A Review on Fairness in Machine Learning

DANA PESSACH and EREZ SHMUELI, Department of Industrial Engineering, Tel-Aviv University

An increasing number of decisions regarding the daily lives of human beings are being controlled by artificial intelligence and machine learning (ML) algorithms in spheres ranging from healthcare, transportation, and education to college admissions, recruitment, provision of loans, and many more realms. Since they now touch on many aspects of our lives, it is crucial to develop ML algorithms that are not only accurate but also objective and fair. Recent studies have shown that algorithmic decision making may be inherently prone to unfairness, even when there is no intention for it. This article presents an overview of the main concepts of identifying, measuring, and improving algorithmic fairness when using ML algorithms, focusing primarily on classification tasks. The article begins by discussing the causes of algorithmic bias and unfairness and the common definitions and measures for fairness. Fairness-enhancing mechanisms are then reviewed and divided into pre-process, in-process, and post-process mechanisms. A comprehensive comparison of the mechanisms is then conducted, toward a better understanding of which mechanisms should be used in different scenarios. The article ends by reviewing several emerging research sub-fields of algorithmic fairness, beyond classification.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Information systems** → **Information systems applications;**

Additional Key Words and Phrases: Algorithmic bias, algorithmic fairness, fairness-aware machine learning, fairness in machine learning

ACM Reference format:
Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (February 2022), 44 pages.
https://doi.org/10.1145/3494672

1 INTRODUCTION

Today, an increasing number of decisions are being controlled by **artificial intelligence (AI)** and **machine learning (ML)** algorithms, with increased implementation of automated decision-making systems in business and government applications. The motivation for an automated learning model is clear—we expect algorithms to perform better than human beings for several reasons. First, algorithms may integrate much more data than a human may grasp and take many more considerations into account. Second, algorithms can perform complex computations much faster than human beings. Third, human decisions are subjective, and they often include biases.

This work was partially supported by the Koret foundation grant for Smart Cities and Digital Living 2030.
Authors' address: D. Pessach (corresponding author) and E. Shmueli, Department of Industrial Engineering, Tel-Aviv University, P.O. Box 39040, 6997801, Tel-Aviv, Israel; emails: danapessach@gmail.com, shmueli@tau.ac.il.

0360-0300/2022/02-ART51 $15.00
https://doi.org/10.1145/3494672

Hence, it is a common belief that using an automated algorithm makes decisions more objective or fair. However, this is unfortunately not the case since ML algorithms are not always as objective as we would expect. The idea that ML algorithms are free from biases is wrong since the assumption that the data injected into the models are unbiased is wrong. More specifically, a prediction model may actually be inherently biased since it learns and preserves historical biases [125].

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. Indeed, in recent years, the concern for algorithm fairness has made headlines. One of the most common examples was in the field of criminal justice, where recent revelations have shown that an algorithm used by the U.S. criminal justice system had falsely predicted future criminality among African-Americans at twice the rate as it predicted for white people [6, 47]. In another case of a hiring application, it was recently exposed that Amazon discovered that their ML hiring system was discriminating against female candidates, particularly for software development and technical positions. One suspected reason for this is that most recorded historical data were for male software developers [54]. In a different scenario in advertising, it was shown that Google's ad-targeting algorithm had proposed higher-paying executive jobs more for men than for women [56, 187].

These lines of evidence and concerns about algorithmic fairness have led to growing interest in the literature on defining, evaluating, and improving fairness in ML algorithms (e.g., see, [20, 48, 79, 97]). It is important to note, however, that the task of improving fairness of ML algorithms is not trivial since there exists an inherent trade-off between accuracy and fairness. In other words, as we pursue a higher degree of fairness, we may compromise accuracy (e.g., see [125]).

This article presents a review of fairness in ML. In contrast to other recent surveys in this field [48, 79, 147], our work proposes a comprehensive and up-to-date overview of the field, ranging from definitions and measures of fairness to state-of-the-art fairness-enhancing mechanisms. Our survey also attempts to cover the pros and cons of the various measures and mechanisms, and guide under which setting they should be used. Finally, although the main part of this article deals primarily with classification tasks, a major goal of this survey is to highlight and discuss emerging areas of research, beyond classification, that are expected to grow in the upcoming years. Overall, this survey provides the relevant knowledge to enable new researchers to enter the field, inform current researchers on rapidly evolving sub-fields, and provide practitioners the necessary tools to apply the results. For further details on the procedure performed to search and select the papers reviewed in this survey, the reader is referred to Appendix B.5.

The rest of this article is structured as follows. Section 2 discusses the potential causes of algorithmic unfairness. Section 3 presents definitions and measures of fairness and their trade-offs. Section 4 reviews fairness mechanisms and methods and a comparison of the mechanisms, focusing on the pros and cons of each mechanism. Section 5 presents several emerging research subfields of fairness in ML beyond classification. Section 6 provides concluding remarks and sketches several open challenges for future research. (For completeness, Appendix A outlines commonly used fairness-related datasets.)

2 POTENTIAL CAUSES OF UNFAIRNESS

We start this section by noting that throughout this article, we use the terms *bias*, *discrimination*, and *unfairness* interchangeably with similar meanings, as is commonly done in the algorithmic fairness literature (e.g., see [206]).

The literature has indicated several causes that may lead to unfairness in ML [48, 145]. These causes can broadly be divided into causes that stem from biases in the data and causes that stem from biases in the algorithm. It is easy to understand why biases in the data can lead to

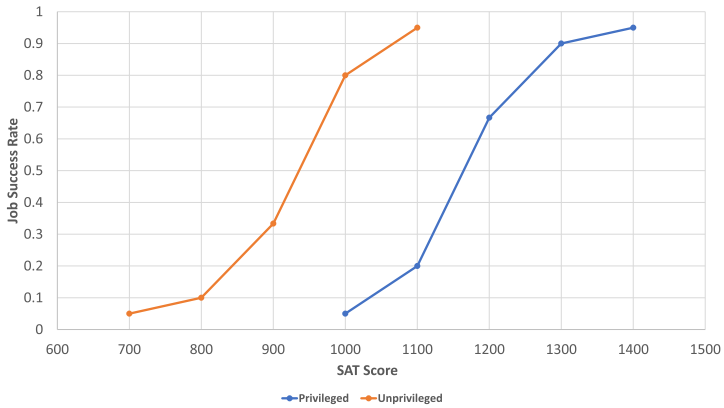

Fig. 1. If the SAT scores were used for hiring, then unprivileged candidates with high potential would be excluded, whereas lower potential candidates from the privileged group would be hired instead.

unfairness—ML models are designed to learn and replicate historical patterns in the data, even if such patterns are biased (e.g., if historically men were hired more frequently than women for technical positions). However, even if the data does not contain any biases, the learned model can still produce unfair results.

To illustrate the latter case, consider the example depicted in Figure 1. The figure illustrates a case of SAT scores for two sub-populations: a privileged one and an unprivileged one. In this example, SAT scores may be used to predict the success (or failure) of candidates in a given job, since the higher the SAT score is, the higher is the probability for success. However, relying solely on the SAT scores while ignoring the group to which the candidates belong may create an undesired bias. More specifically, as can be seen from the figure, unprivileged candidates with SAT scores of approximately 1,100 perform just as well as privileged candidates with SAT scores of 1,400 (e.g., since they may have encountered harder challenges along their way for achieving their scores). Therefore, if SAT scores were used for hiring, such as by just placing a threshold, unprivileged candidates with high potential would be excluded, whereas lower potential candidates from the privileged group would be hired instead.

We hereby summarize the main categories of causes for unfairness identified in the literature:

- Biases already included in the datasets used for learning, which are based on biased device measurements, historically biased human decisions, erroneous reports, or other reasons. ML algorithms are essentially designed to replicate these biases.
- Biases caused by missing data, such as missing values or sample/selection biases, which result in datasets that are not representative of the target population.
- Biases that stem from algorithmic objectives, which aim at minimizing overall aggregated prediction errors and therefore benefit majority groups over minorities.[1]
- Biases caused by "proxy" attributes for sensitive attributes. Sensitive attributes differentiate privileged and unprivileged groups, such as race, gender, and age, and are typically not legitimate for use in decision making. Proxy attributes are non-sensitive attributes that can be exploited to derive sensitive attributes. In the case that the dataset contains proxy attributes,

[1]Note that unreliable algorithms (e.g., algorithms with weak generalization ability) can also lead to unfairness (see more on *bias-variance trade-off* in the work of Belkin et al. [18], *structural risk minimization* in the work of Kim [123] and Vapnik [205], and *inductive bias* in the work of Mitchell [151]).

the ML algorithm can implicitly make decisions based on the sensitive attributes under the cover of using presumably legitimate attributes [15].

Interested readers are referred to the work of Mehrabi et al. [147], which lists a variety of scenarios in which the main causes for biases and unfairness mentioned here are reflected, and discusses several methods for mitigation.

3 FAIRNESS DEFINITIONS AND MEASURES

This section presents some general legal notions for discrimination followed by a survey of the most common measures for algorithmic fairness, as well as the inevitable trade-offs between them.

3.1 Definitions of Discrimination in Legal Domains

The legal domain has introduced two main definitions of discrimination: (i) *disparate treatment* [15, 224]: intentionally treating an individual differently based on his/her membership in a protected class (*direct discrimination*); (ii) *disparate impact* [15, 175]: negatively affecting members of a protected class more than others even if by a seemingly neutral policy (*indirect discrimination*).

Put in our context, it is important to note that algorithms trained with data that do not include sensitive attributes (i.e., attributes that explicitly identify the protected and unprotected groups) are unlikely to produce *disparate treatment* but may still induce unintentional discrimination in the form of *disparate impact* [125].

3.2 Measures of Algorithmic Bias

This section presents the most prominent measures of algorithmic fairness in ML classification tasks. We refer the readers to Appendix B.1 and specifically to Table 4 for a review of additional, less popular measures used in the literature. We additionally refer the readers to Appendix B.2 for an additional summary regarding insights on trade-offs and measures impossibility results discussed in the literature. Moreover, we refer the reader to Appendix B.3 for information on emerging notions of fairness:

(1) *Disparate impact* [74]: This measure was designed to mathematically represent the legal notion of *disparate impact*. It requires a high ratio between the positive prediction rates of both groups. This ensures that the proportion of the positive predictions is similar across groups. For example, if a positive prediction represents acceptance for a job, the condition requires the proportion of accepted applicants to be similar across groups. Formally, this measure is computed as follows:

$$\frac{P[\hat{Y} = 1|S \neq 1]}{P[\hat{Y} = 1|S = 1]} \geq 1 - \varepsilon, \tag{1}$$

where S represents the protected attribute (e.g., race or gender), $S = 1$ is the privileged group, and $S \neq 1$ is the unprivileged group. $\hat{Y} = 1$ means that the prediction is positive. Let us note that if $\hat{Y} = 1$ represents acceptance (e.g., for a job), then the condition requires the acceptance rates to be similar across groups. A higher value of this measure represents more similar rates across groups and therefore more fairness. Note that this notion relates to the "80 percent rule" in disparate impact law [74], which requires that the acceptance rate for any race, sex, or ethnic group be at least 80% of the rate for the group with the highest rate.

(2) *Demographic parity:* This measure is similar to *disparate impact*, but the difference is taken instead of the ratio [36, 60]. This measure is also commonly referred to as *statistical parity*.

Formally, this measure is computed as follows:

$$\left|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]\right| \leq \varepsilon. \tag{2}$$

A lower value of this measure indicates more similar acceptance rates and therefore better fairness. *Demographic parity* (and *disparate impact*) ensure that the positive prediction is assigned to the two groups at a similar rate.

One disadvantage of these two measures is that a fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different. Moreover, to satisfy *demographic parity*, two similar individuals may be treated differently since they belong to two different groups—such treatment is prohibited by law in some cases (note that this notion also corresponds to the practice of *affirmative action* [80]).

(3) *Equalized odds*: This measure was designed by Hardt et al. [94] to overcome the disadvantages of measures such as *disparate impact* and *demographic parity*. The measure computes the difference between the **false-positive rates (FPRs)**, and the difference between the **true-positive rates (TPRs)** of the two groups. Formally, this measure is computed as follows:

$$\left|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]\right| \leq \varepsilon, \tag{3}$$

$$\left|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]\right| \leq \varepsilon, \tag{4}$$

where the upper formula requires the absolute difference in the FPR of the two groups to be bounded by ε, and the lower formula requires the absolute difference in the TPR of the two groups to be bounded ε. Smaller differences between groups indicate better fairness. In contrast to *demographic parity* and *disparate impact* measures, a fully accurate classifier will necessarily satisfy the two *equalized odds* constraints. Nevertheless, since *equalized odds* relies on the actual ground truth (i.e., Y), it assumes that the base rates of the two groups are representative and were not obtained in a biased manner.

One use case that demonstrates the effectiveness of this measure investigated the COMPAS [103] algorithm used in the U.S. criminal justice system. For predicting recidivism, although its accuracy was similar for both groups (African-Americans and Caucasians), it was discovered that the *odds* were different. It was discovered that the system had falsely predicted future criminality (FPR) among African-Americans at twice the rate predicted for white people [6]; importantly, the algorithm also induced the opposite error, significantly underestimating future crimes among Caucasians **(false-negative rate (FNR)**).

(4) *Equal opportunity*: This requires TPRs to be similar across groups (meaning the probability of an individual with a positive outcome to have a positive prediction) [94]. This measure is similar to equalized odds but focuses on the TPRs only. This measure is mathematically formulated as follows:

$$\left|P[\hat{Y} = 1|S \neq 1, Y = 1] - P[\hat{Y} = 1|S = 1, Y = 1]\right| \leq \varepsilon. \tag{5}$$

Let us note that 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 [168]. Moreover, according to Corbett-Davies and Goel [50], this measure may be problematic when base rates differ between groups.

Thus far, we have mapped the most common *group* notions of fairness, which require parity of some statistical measure across groups. The literature has additionally indicated *individual* notions of fairness. It is alternatively possible to match other measures such as accuracy, error rates, or

calibration values between groups (see Appendix B.1 and specifically Table 4). *Group* definitions of fairness, such as *demographic parity, disparate impact, equalized odds*, and *equalized opportunity*, consider fairness with respect to the whole group, as opposed to *individual* notions of fairness.

(5) *Individual fairness*: This requires that similar individuals will be treated similarly. Similarity may be defined with respect to a particular task [60, 107]. Individual fairness may be described as follows:

$$|P(\hat{Y}^{(i)} = y|X^{(i)}, S^{(i)}) - P(\hat{Y}^{(j)} = y|X^{(j)}, S^{(j)})| \leq \varepsilon;\ if\ d(i,j) \approx 0, \tag{6}$$

where i and j denote two individuals, $S^{(\cdot)}$ refers to the individuals' sensitive attributes, and $X^{(\cdot)}$ refers to their associated features. $d(i,j)$ is a distance metric between individuals that can be defined depending on the domain such that similarity is measured according to an intended task. This measure considers other individual attributes for defining fairness rather than just the sensitive attributes. However, note that to define similarity between individuals, a similarity metric needs to be defined, which is not trivial. This measure, in addition to assuming a similarity metric, also requires some assumptions regarding the relationship between features and labels (e.g., see [48]).

3.3 Trade-Offs

Determining the right measure to be used must take into account the proper legal, ethical, and social context. As demonstrated earlier, different measures exhibit different advantages and disadvantages. Next, we highlight the main trade-offs that exist between different notions of fairness, and the inherent trade-off between fairness and accuracy.

Fairness Measures Trade-Offs. Interestingly, several recent studies have shown that it is not possible to satisfy multiple notions of fairness simultaneously [20, 47, 50, 51, 77, 125, 168]. For example, when base rates differ between groups, it is not possible to have a classifier that equalizes both calibration and odds (except for trivial cases such as a classifier that assigns all examples to a single class). Additionally, there is also evidence for incompatibility between equalized accuracy and equalized odds, as in the COMPAS criminal justice use case [6, 20].

We will hereby demonstrate some intuition for trade-offs between measures, and we additionally refer the reader to Appendices B.1 and B.2 for further details on the trade-offs between fairness measures and measures impossibility results. As mentioned, one possible incompatibility can occur between the measures of equalized odds and demographic parity in cases when base rates are different (i.e., different proportions of actual positive outcomes). The intuition can be demonstrated, for example, by considering the case of pursuing demographic parity by requiring that the proportion of inmates selected for parole would be the same for men and women. Note that following the preceding requirement may harm the fairness notion of equalized odds, since women who pose a lower safety risk may be incarcerated, solely for the purpose of achieving the same proportions of releases for men and women [20]. In other words, since base rates are different among men versus women (different proportions of actual recidivism), when following demographic parity, **false-positive (FP)** errors will be increased for one group (errors against women), whereas **false-negative (FN)** errors will be increased for the other group (errors in favor of men), hence harming equalized odds. However, in an extreme case, when the model results in no errors at all on both groups (i.e., FPs and FPRs among groups are equal, since they are all zero)—due to different base rates, the proportion of selected parolees of each group will be different—hence, demographic parity will not be satisfied.

The preceding intuition can also be explained by the three fairness criteria highlighted in the work of Barocas et al. [14]: *independence*, *separation*, and *sufficiency*. Achieving independence refers

to obtaining model predictions that are not dependent on the individual's group membership. The measure of demographic parity is a measure that focuses on this independence criterion. However, measures that rely on independence do not take into account the actual outcomes. The separation criterion is therefore extending independence to be conditional on the actual outcome. The measure of equalized odds is a measure that focuses on the separation criterion. As discussed earlier, if the outcome is indeed dependent on the group membership (i.e., different base rates), then independence and separation cannot be both satisfied [14]. The measure of calibration follows the criterion of sufficiency, which requires that for each predicted score, the outcome is independent of the group membership. Barocas et al. [14] show that these three criteria cannot be satisfied simultaneously, unless in degenerate cases.

Pleiss et al. [168] recommends that in light of the inherent incompatibility between equalized calibration and equalized odds, practical implications require choosing only one of these goals according to the specific application's requirements. We recommend that any selected measure of algorithmic fairness be considered in the appropriate legal, social, and ethical contexts.

Individual and group fairness can sometimes be incompatible as well. For example, Dwork et al. [60] highlight the case of the trade-off between demographic parity and individual fairness and show that they cannot be satisfied simultaneously unless in trivial degenerate solutions. They present an example case to show the intuition, in which imposing demographic parity between two groups S_0 and S_1 can be problematic when members of S_0 are less likely to be "qualified." This will cause unfair preferential treatment. They propose to approach this problem by proposing a linear programming optimization model that enhances a more "fair affirmative action." The model requires demographic parity while satisfying as much individual fairness as possible (recall that individual fairness requires that similar individuals be treated similarly), as well as a utility requirement. They additionally discuss the importance of defining the distances between individuals in a manner that strong individuals of both groups are considered similar, instead of similarity that is derived mostly from the characteristics of the membership in the same group.

Fairness-Accuracy Trade-Off. The literature extensively discusses the inherent trade-off between accuracy and fairness—as we pursue a higher degree of fairness, we may compromise accuracy (e.g., see [125]). A theoretical analysis of the trade-off between fairness and accuracy was studied in the work of Corbett-Davies et al. [51] and Lipton et al. [135]. Since then, many papers have empirically supported the existence of this trade-off (e.g., [16, 79, 149]). Generally, the aspiration of a fairness-aware algorithm is to achieve a model that allows for higher fairness without significantly compromising the accuracy or other alternative notions of utility.

Intuitively, if we consider this as a constrained optimization problem, then a classifier that includes additional fairness constraints will have comparable or worse results in terms of accuracy than a classifier that aims solely at maximizing accuracy. To demonstrate this idea, consider the example presented in Figure 1 in Section 2. As explained earlier, a simple classifier that aims at maximizing accuracy will set a threshold value around an SAT score of 1,200. In this case, the error rate for unprivileged candidates who are not hired but succeed is higher than those in the privileged population. Additionally, the error rate for privileged candidates who are hired but fail is higher than those in the unprivileged population (i.e., relatively high accuracy with a cost of some unfairness, in terms of equalized odds). In contrast, a naive classifier that aims at maximizing fairness can set a threshold value around an SAT score of 0, achieving similar error rates for both groups, but resulting in very low accuracy.

In other words, it is possible to tune the threshold value in a manner that more candidates from the unprivileged population are selected and to add constraints to the problem that will improve

Table 1. Measures and Definitions for Algorithmic Fairness

Measure	Paper	Description	Type	Uses Actual Outcome	Uses Sensitive Attribute	Type of Actual Outcome	Type of Sensitive Attribute	Equivalent Notions
Disparate Impact	[74]	High ratio between positive prediction rates of both groups	Group	✗	✓	–	Binary	For $\varepsilon = 0.2$ relates to the "80 percent rule" in disparate impact law [74]
Demographic Parity	[36], [60]	Similar positive prediction rates between groups	Group	✗	✓	–	Binary	• Statistical parity [60] • Group fairness [60] • Equal acceptance rates [47, 206, 225] • Discrimination score [36]
Equal Opportunity	[94]	Requires that TPRs are similar across groups	Group	✓	✓	Binary	Binary	• Equal TPR • Mathematically equal TPRs will induce equal FNRs (see [51, 206]) • FN error rate balance [47, 206]
Equalized Odds	[94]	Requires that FPRs (1-TNR) and TPRs (1-FNR) are similar across groups	Group	✓	✓	Binary	Binary	• Disparate mistreatment [214] • Error rate balance [47] • Conditional procedure accuracy equality [20]
Fairness through Awareness	[60]	Requires that similar individuals will have similar classifications; similarity can be defined with respect to a specific task	Individual	✗	✗	–	–	Individual fairness

fairness. Intuitively, because a classifier that does not take fairness into account is the one that maximizes accuracy, placing additional constraints on the problem, such as fairness constraints, will result in a decrease in accuracy.

It is worth noting that although a trade-off is generally the expected relationship between fairness and accuracy, this is not the case in all scenarios. For example, as mentioned, under the measure of equalized odds, a fully accurate classifier is considered perfectly fair. In addition, several recent papers have discussed potentially unique scenarios where increasing fairness can also increase accuracy (e.g., see [165, 208]).

Table 1 provides a summary of the measures presented in this section. For further reading about algorithmic fairness measures, we refer the reader to the work of Corbett-Davis et al. [50], Kleinberg et al. [125], and Verma and Rubin [206].

4 FAIRNESS-ENHANCING MECHANISMS

Numerous recent papers have proposed mechanisms to enhance fairness in ML algorithms. These mechanisms are typically categorized into three types: pre-process, in-process, and post-process. The following three sections review studies in each one of these categories. The fourth section is devoted to comparing the three mechanism types and providing guidelines on when each type should be used.

4.1 Pre-Process Mechanisms

Mechanisms in this category involve changing the training data before feeding it into an ML algorithm. Preliminary mechanisms, such as the ones proposed by Kamiran and Calders [111] and Luong et al. [140] proposed changing the labels of some instances or reweighing them before training to make the classification fairer. Typically, the labels that are changed are related to samples that are closer to the decision boundary since these are the ones that are most likely to be discriminated. More recent mechanisms suggest modifying feature representations so that a subsequent classifier will be fairer [38, 74, 139, 180, 218].

For example, Feldman et al. [74] suggest modifying the features in the dataset so that the distributions for both privileged and unprivileged groups become similar, therefore making it more difficult for the algorithm to differentiate between the two groups. A tuning parameter λ was provided for controlling the trade-off between fairness and accuracy ($\lambda = 0$ indicates no fairness considerations, whereas $\lambda = 1$ maximizes fairness). Chierichetti et al. [46] and Backurs et al. [9] use the same notion of fair representation learning and apply it for fair clustering, and Samadi et al. [180] apply it for fair dimensionality reduction (PCA). For more fair representation learning using *adversarial learning*, see Section 5.2.

Note that this approach to achieving fairness is somewhat related to the field of data *compression* [197, 218]. It is also very closely related to *privacy* research since both fairness and privacy can be enhanced by removing or obfuscating the sensitive information, with the adversary goal of minimal data distortion [65, 118].

4.2 In-Process Mechanisms

These mechanisms involve modifying the ML algorithms to account for fairness during the training time [4, 16, 17, 36, 84, 114, 209, 214, 215].

For example, Kamishima et al. [114] suggest adding a regularization term to the objective function that penalizes the mutual information between the sensitive feature and the classifier predictions. A tuning parameter η was provided to modulate the trade-off between fairness and accuracy.

Zafar et al. [214, 215] and Woodworth et al. [209] suggest adding constraints to the classification model that require satisfying a proxy for *equalized odds* [209, 214] or *disparate impact* [215]. Woodworth et al. [209] also show that there exist difficult computational challenges in learning a fair classifier based on *equalized odds*.

Bechavod and Ligett [16, 17] suggest incorporating penalty terms into the objective function that enforce matching proxies of FPR and FNR. Kamiran et al. [112] suggest adjusting a decision tree split criterion to maximize information gain between the split attribute and the class label while minimizing information gain with respect to the sensitive attribute. Zemel et al. [218] combine fair representation learning with an in-process model by applying a multi-objective loss function based on logistic regression, and Louizos et al. [139] apply this notion using a variational autoencoder.

Quadrianto and Sharmanska [169] suggest using the notion of *privileged learning*[2] for improving performance in cases where the sensitive information is available at training time but not at testing time. They add constraints and regularization components to the privileged learning **support vector machine (SVM)** model proposed by Vapnik and Izmailov [204]. They combine the sensitive attributes as privileged information that is known only at training time, and they additionally use a **maximum mean discrepancy (MMD)** criterion [89] to encourage the distributions to be similar across privileged and unprivileged groups.

Berk et al. [19] propose a convex in-process fairness mechanism for regression tasks and use three regularization terms that include variations of individual fairness, group fairness, and a combined hybrid fairness penalty term. Agarwal et al. [5] propose an in-process minimax optimization formulation for enhancing fairness in regression tasks based on the suggested design of Agarwal et al. [4] for classification tasks. They use two fairness metrics adjusted for regression tasks. One is an adjusted *demographic parity* measure, which requires the predictor to be independent of the sensitive attribute as measured by the **cumulative distribution function (CDF)** of the protected group compared to the CDF of the general population [5] using the *Kolmogorov-Smirnov statistic*

[2]Privileged learning is designed to improve performance by using additional information, denoted as the "privileged information," which is present only in the training stage and not in the testing stage [204].

[133]. The second measure is the bounded group loss (BGL), which requires that the prediction error of all groups remain below a predefined level [5].

4.3 Post-Process Mechanisms

These mechanisms perform post-processing of the output scores of the classifier to make decisions fairer [51, 62, 94, 149]. For example, Hardt et al. [94] propose a technique for flipping some decisions of a classifier to enhance equalized odds or equalized opportunity. Corbett-Davies et al. [51] and Menon and Williamson [149] similarly suggest selecting separate thresholds for each group separately, in a manner that maximizes accuracy and minimizes demographic parity. Dwork et al. [62] propose 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 (to read more about transfer learning, see the work of Pan and Yang [161]).

In this section, we so far briefly described the various mechanisms for enhancing fairness. In Section 4.4, we expand the discussion to further provide insights for comparison between mechanisms, and for how to select which mechanisms are suitable for use in different scenarios, based on their respective advantages and disadvantages.

Moreover, Table 2 presents a summary of the details of pre-process, in-process, and post-process mechanisms discussed in this section. The table additionally reviews the measures used by the mechanisms in each of the papers. It should be noted that our analysis differentiates between the measures used for optimization purposes versus the measures used for evaluation purposes. Note that these two could be different due to reasons such as computational complexity. Moreover, the table lists the datasets that were utilized by the authors of the cited papers for examining the performance of their proposed mechanisms. More details with respect to the various datasets that are mentioned in the table can be found in Appendix A.

Unfortunately, we cannot fully cover all mechanisms in-depth; however, we believe it is worthwhile to demonstrate several methods in more detail, to alleviate an entry-level understanding of some of the most important principles of mechanisms. In Appendix B.4, we describe three mechanisms in more depth, one of each mechanism type.

Recent papers have proposed new approaches for how to define fairness and have consequently applied new adequate enhancement mechanisms. Grgić-Hlača et al. [91] and Green and Hu [88] discuss the notion of *procedural fairness*; Zafar et al. [216], Kim et al. [124], and Ustun et al. [199] propose using preference-based fairness measures that leverage notions from the fields of economics and game theory; Speicher et al. [190] and Lohia et al. [138] propose a combined approach to both individual and group fairness; and Lahoti et al. [129] and Jung et al. [108] introduce *pairwise fairness*. We refer the reader to Appendix B.3 for more information.

Note that the methods discussed in this section were designed for the task of classification. Fairness mechanisms for other learning tasks are discussed in Section 5.

4.4 Which Mechanism to Use?

The different mechanism types present respective advantages and disadvantages. Pre-process mechanisms can be advantageous since they can be used with any classification algorithm. However, they may harm the explainability of the results. Moreover, since they are not tailored for a specific classification algorithm, there is high uncertainty with regard to the level of accuracy obtained at the end of the process.

Similar to pre-process mechanisms, post-process mechanisms may be used with any classification algorithm. However, due to the relatively late stage in the learning process in which they are applied, post-process mechanisms typically obtain inferior results [209]. In a post-process mechanism, it may be easier to fully remove bias types such as *disparate impact*; however, this is not

Table 2. Pre-Process, In-Process, and Post-Process Mechanisms for Algorithmic Fairness

Paper	Mechanism Type	Base Algorithm	Optimization Measure	Evaluation Measure	Method Name	Datasets	General Description
[74]	Pre-process	Any	Earth moving distance	Disparate impact	Removing Disparate Impact	• Adult • German	Pre-process data to decrease earth moving distance between distributions of both groups. Note that it does not require any ground truth data.
[111]	Pre-process	Any score based	Acceptance probabilities, distance from boundary	Demographic parity	• Massaging • Reweighing • Sampling • Suppression	• German • Adult • Communities • Dutch	*Massaging*: Changing some labels before training (selecting points with lower acceptance rate, trying to replace a small number of samples as opposed to the work of Luong et al. [140]). *Reweighing*: Assigning different weights to samples. *Sampling*: Uniform sampling or preferential sampling, using the idea that the closest the samples are to the boundary, the more likely they are to be discriminated. *Suppression*: Baseline approach to remove correlated features (performed worse).
[140]	Pre-process	Any	Conditional statistical parity	Conditional statistical parity	Discrimination Prevention with KNN	• Adult • Communities • German	*Changing labels before training* such that individuals that are similar except their sensitive attribute will not have different labels, and more individuals from the unprivileged group will get positive labels.
[38]	Pre-process	Any	• Disparate impact • Individual fairness	Disparate impact	Optimized Pre-Processing for Discrimination Prevention	• ProPublica • Adult	Transforming the data to a new mapping that enhances both group fairness and individual fairness while maintaining utility. An optimization formulation that incorporates these three requirements.
[199]	Pre-process	Any	Group preference-based fairness (see Appendix B.3)	Group preference-based fairness (see Appendix B.3)	Decoupled Classifiers with Preference Guarantees	• ProPublica • Adult • Apnea [200] • Cancer [196]	Recursive feature selection procedure for building decoupled classifiers.
[129]	Pre-process	Any	Pairwise fairness (see Appendix B.3)	Pairwise fairness (see Appendix B.3)	Pairwise Fair Representation	• ProPublica • Communities	Pre-process mechanism to create fair representations by leveraging external knowledge on human judgments on pairs of individuals that should be treated similarly. Use fairness graph constraints combined with data-driven similarities.
[114]	In-process	Logistic regression	Mutual information between prediction and sensitive attribute	Normalized prejudice index	Prejudice Remover Regularizer	Adult (test only)	Solving an optimization problem with a penalty that minimizes mutual information between prediction and sensitive attribute.
[215]	In-process	Decision boundary based	Covariance (between sensitive attributes and distance to the decision boundary)	Disparate impact	Fairness Constraints	ProPublica	Optimization using additional constraints to control the covariance.
[214]	In-process	Decision boundary based	Proxy for equalized odds	Equalized odds	Removing Disparate Mistreatment	ProPublica	Minimizing a proxy for equalized odds with constraints.
[17]	In-process	Decision boundary based	Proxy for equalized odds	Equalized odds	Penalizing Unfairness	• ProPublica • Adult • Loans • Admissions	Minimizing a proxy for equalized odds in a penalty to the objective function.
[84]	In-process	SVM	Disparate impact, Equal opportunity	Disparate impact	Dataset Constraints	Adult	Constrained non-convex optimization using SVM with ramp loss, to enhance disparate impact.

(Continued)

Table 2. Continued

Paper	Mechanism Type	Base Algorithm	Optimization Measure	Evaluation Measure	Method Name	Datasets	General Description
[91]	In-process	Logistic regression	Procedural fairness (see Appendix B.3)	Procedural fairness (see Appendix B.3)	Feature Selection for Procedurally Fair Learning	• ProPublica • NYPD SQF [158]	Perform fair feature selection to balance procedural fairness (that considers fair process rather than fair outcome) and accuracy, using constrained submodular optimization. Combines human judgments to assess procedural fairness.
[216]	In-process	Convex boundary-based classifiers	Group preference-based fairness (see Appendix B.3)	Group preference-based fairness (see Appendix B.3)	Achieving preferred treatment and preferred impact	• ProPublica • Adult • NYPD SQF [158]	In-process mechanisms that solves an optimization problem that balances accuracy and tractable proxies for preferred treatment and preferred impact (see Appendix B.3).
[124]	In-process	Classifiers with convex objectives	Individual preference-based fairness (see Appendix B.3)	Individual preference-based fairness (see Appendix B.3)	Optimization subject to Preference Informed Individual Fairness (PIIF)	–	Train a linear classifier for each group by solving an optimization problem that balances PIIF and accuracy. PIIF is a relaxation of both individual fairness and envy-freeness.
[108]	In-process	Linear classifier	Pairwise fairness (see Appendix B.3)	Pairwise fairness (see Appendix B.3)	Algorithmic Framework for Fairness Elicitation	ProPublica	In-process mechanism that leverages external information on equally deserving individuals. The optimization procedure trades off classification accuracy with frequency of decisions that violates human subjective fairness judgments.
[169]	In-process	SVM	MMD	• Equalized odds • Overall accuracy equality	Recycling Privileged Learning and Distribution Matching	• ProPublica • Adult	Use SVM with privileged learning [204]. Combine the sensitive attributes as privileged information that is known only at training time. Use the MMD criterion to encourage the distributions to be similar across privileged and unprivileged groups [89].
[4]	In-process	Any	• Demographic parity • Equalized odds	• Demographic parity • Equalized odds	Reductions Approach	• ProPublica • Adult • Dutch • Admissions	A minimax optimization formulation. that aims at maximizing the predictor's capability to predict the outcome while minimizing the capability to predict the sensitive feature. Solved using *saddle point* methods.
[94]	Post-process	Any score based	Equalized odds	Equalized odds	Equality of Opportunity in Supervised Learning	FICO scores [94]	Selecting different decision thresholds for different groups (privileged and unprivileged), enhancing equalized odds or equal opportunity. zation—
[51]	Post-process	Any score based	• Demographic parity • Conditional statistical parity • Predictive parity	• Demographic parity • Conditional statistical parity • Predictive parity	Cost of Fairness	ProPublica	Selecting separate thresholds for each group that maximizes accuracy while minimizing demographic parity.
[138]	Post-process	Any	Combination of individual and group fairness (see Appendix B.3)	Combination of individual and group fairness (see Appendix B.3)	Individual+Group Debiasing (IGD)	• ProPublica • German • Adult	Post-process mechanism to balance accuracy with both individual and group fairness. Begin by detecting samples that are prone to individual bias and consider them for prediction change for enhancing disparate impact. Note that this method does not require ground truth class labels for the validation set.
[149]	Post-process	Any score based	• Demographic parity • Equal opportunity	• Demographic parity • Equal opportunity	Plugin Approach	-	Selecting different thresholds for each of the groups. Using Lagrange relaxation.

(Continued)

Table 2. Continued

Paper	Mechanism Type	Base Algorithm	Optimization Measure	Evaluation Measure	Method Name	Datasets	General Description
[112]	In-process, Post-process	In-process—decision tree; Post-process—any algorithm	Information gain	Demographic parity	• Discrimination Aware Tree Construction (new split criterion) • Relabeling (post-process)	• Adult • Communities • Dutch census	*New split criterion* (in-process)—while maximizing information gain between selected split attribute to class label, also minimizing information gain to sensitive attribute. *Leaf relabeling* (post-process)—change class labels of some leaves trading as little accuracy as possible, using KNAPSACK optimization formulation [8].
[36]	In-process, Post-process	Naive Bayes	Acceptance probabilities	Demographic parity	• Modifying Naive Bayes • Two Naive Bayes • Expectation Maximization (EM)	Adult (test only)	*Modifying naive Bayes*—modify probabilities so that the unprivileged group would get more positive predictions, enhancing demographic parity. *Two naive Bayes*—using two separate models to predict each of the groups, and balance demographic parity of both groups in an iterative process, by making slight changes to the observed probabilities. *EM*—add a latent variable, which represents actual labels, to the Bayesian model and optimize using EM.
[218]	Pre-process + In-process	Logistic regression	Demographic parity	Demographic parity	Learning Fair Representations	• Adult • German • Heritage health	Combining fair representation learning with an in-process model by a multi-objective loss function that requires representations to retain information of the original vectors, induce fair statistical parity, and maintain as high accuracy as possible.
[139]	Pre-process + In-process	Any	MMD	• Demographic parity • Mean difference	Variational Fair Autoencoder	• Adult • German • Heritage health	Fair representation learning using a variational autoencoder. Additional penalty for controlling distribution matching between groups using MMD criterion [89].
[209]	In-process + Post-process	Convex linear	Equalized odds	Equalized odds	Learning Non-Discriminatory Predictors	–	A two-step framework that combines in-process and post-process mechanisms to enhance equalized odds. A heuristic relaxation to solve the resulting hard non-convex problem.
[62]	In-process + Post-process	Any score-based	• Demographic parity • Equalized odds	• Demographic parity • Equalized odds	Decoupled Classifiers	ImageNet [57]	Decoupling technique to learn a different classifier for each group and then use transfer learning technique to learn from out-of-group samples.

always the desired measure, and it could be considered as discriminatory since it deliberately damages accuracy for some individuals to compensate others (this is also related to the controversies in the legal and economical field of *affirmative action*; see [80]). Specifically, post-process mechanisms may treat differently two individuals who are similar across all features except for the group to which they belong. This approach requires the decision maker at the end of the loop to possess the information of the group to which individuals belong (this information may be unavailable due to legal or privacy reasons).

In-process mechanisms are beneficial since they can explicitly impose the required trade-off between accuracy and fairness in the objective function [209]. However, such mechanisms are tightly coupled with the machine algorithm itself.

Hence, we see that the selection of the method depends on the availability of the ground truth, the availability of the sensitive attributes at test time, and on the desired definition of fairness, which can also vary from one application to another.

Several preliminary attempts were made to understand which methods are best for use. The study in Hamilton [93] was a first effort in comparing several fairness mechanisms previously

proposed in the literature [36, 74, 114, 215]. The analysis focuses on binary classification with binary sensitive attributes. The authors have demonstrated that the performances of the methods vary across datasets, and there was no conclusively dominating method.

Another study by Roth [172] has shown as a preliminary benchmark that in several cases, in-process mechanisms perform better than pre-process mechanisms, and for other cases, they do not, leading to the conclusion that there is a need for much more extensive experiments.

A recent empirical study by Friedler et al. [79] has provided a benchmark analysis of several fairness-aware methods and compared the fairness-accuracy trade-offs obtained by these methods. The authors have tested the performances of these methods across different measures of fairness and across different datasets. They have concluded that there was no single method that outperformed the others in all cases and that the results depend on the fairness measure, on the dataset, and on changes in the train-test splits.

More research is required for developing robust fairness mechanisms and metrics or, alternatively, for finding the adequate mechanism and metric for each scenario. For instance, the conclusions reached when considering missing data might be very different from those reached when all information is available [109, 145]. Martínez-Plumed et al. [145] have tested imputation strategies to deal with the fairness of partially missing examples in the dataset. They observe, for instance, that unprivileged individuals are more reluctant to reveal some information, assuming it can be used against them. In turn, this self-reporting missing information itself might cause unintentional amplification of unfairness, due to the nature of the relationship between missing values and unprivileged groups. They have empirically shown that the imputation of rows with missing values may be fairer than the deletion of these rows. However, they note that imputation methods can introduce bias as well, so they recommend examining, for each particular task, if indeed imputation improves fairness compared to the deletion of missing values (depending on the metric, mechanism, and dataset). Pessach and Shmueli [165] find that when there is an evident selection bias in the data, meaning that there is an extreme under-representation of unprivileged groups, pre-process mechanisms can outperform in-process mechanisms.

5 EMERGING RESEARCH ON ALGORITHMIC FAIRNESS

In this section, we review selected emerging sub-fields of algorithmic fairness.

5.1 Fair Sequential Learning

Most existing research on algorithmic fairness considers batch classification, where the complete data are available in advance. However, many learning settings have a dynamic nature in which data are collected over time. In these settings, in contrast to batch learning, the system includes feedback loops so that the decision at each step may influence future states and decisions. This holds also with regard to fairness decisions, as they should now be considered at each step, where short-term fairness decisions may affect long-term fairness results, a setting typically referred to as *sequential learning*. In this context, there is a need to balance *exploitation* of existing knowledge (e.g., hiring an already known population) and *exploration* of sub-optimal solutions to gather more data (e.g., hiring populations of different backgrounds that differ from current employees).

Several studies have investigated fairness in sequential learning [95, 105, 107, 115, 201]. For example, Jabbari et al. [105] study fairness in reinforcement learning and model the environment as a *Markov decision process*. In their model, fairness is defined such that one action will never be preferred over another if its long-term discounted reward is lower (e.g., enriching the overall applicant pool in the long run might require sacrifices in the short run such as productivity losses). Heidari and Krause [95] define fairness as time-dependent individual fairness and require that algorithmic decisions be *consistent* over time. They propose a post-process mechanism that imposes

these time-dependent constraints such that two individuals who arrive during the same period and are similar in their feature dimension must be assigned similar labels.

One challenge that arises in this domain is defining the length of the time windows to which fairness refers. On the one hand, short time windows allow for increased control on the process; however, on the other hand, they endure higher complexity and a smaller space of feasible solutions. Therefore, in many cases, longer time windows may be preferable.

Another challenge exists in choosing the right balance between exploration and exploitation. It is worth noting that an exploration process that aims at increasing fairness in the long run may be unethical on its own (e.g., taking medical action [48] or hiring sub-populations for which the results may be sub-optimal).

In a similar line of work, researchers have investigated scenarios where feedback loops have the potential to cause amplification of bias. In these scenarios, the decisions based on the ML models then affect the future collected data. The risk of feedback loops is that they can introduce self-fulfilling predictions, where acting on a prediction can change the outcomes. For example, sending more police officers to an area that was predicted to be at high risk for crime will inevitably cause the arrest of more individuals in this area, then the prediction model will eventually further increase the risk prediction for the area [72].

Note that fair sequential learning is somewhat different from fair selection in a multi-stage process. In the latter, more information is gained about individuals during each stage—for example, when screening candidates first by their resumes, then by their test scores, and finally by interviews. This field is sometimes referred to as *fair pipelines* [29, 61, 71, 99, 142]. In these studies, fairness is revised to be considered in each stage, not only in the final stage.

5.2 Fair Adversarial Learning

Generative adversarial networks (GANs) [86] are commonly utilized, among other purposes, for the generation of simulated representative samples of a dataset. In this field of research, input data can be of various domains such as images or tabular data. In **computer vision (CV)**, for instance, adversarial learning is used for tasks such as image generation (e.g., [12]) or modification (e.g., [7]), along with other CV tasks.

Today, fair adversarial learning is attracting increasing attention with respect to both fair classification and the generation of fair representations. In one distressing incident, a face-modifying application was exposed as racist when the app's "image filter" that was intended to change face images to more "attractive" made skin lighter [167].

GANs are generally constructed as a feedback loop model, starting from a generator G that generates "fake" simulated samples and a discriminator D (the "adversary") that determines whether the generated samples are real or fake and returns the decisions as feedback to the generator G to improve its model. Improving G means enhancing its ability to generate samples that are increasingly similar to real samples in a manner that "fools" the discriminator D, thus minimizing its ability to differentiate between real and fake samples. Typically, both G and D are multi-layer neural networks.

To use GANs for fair learning, previous studies have developed different approaches. Models that are based on GANs are often constructed as minimax optimization problems that aim at maximizing the predictor's capability to accurately predict the outcomes while minimizing the adversary's capability to predict the sensitive feature.

In one approach, it was suggested to use the feedback structure to check whether a trained classifier is fair or not and then update the model accordingly [40, 207, 219].

A different approach encourages the use of GANs to additionally learn fair representations or embedding from the training data so that it is more difficult for a subsequent classifier to

distinguish which samples belong to a privileged group and which belong to an unprivileged group [23, 66, 141].

Another approach [3, 211] is using GANs for generating fair *synthetic* data from the initial input data and then using them to train any classifier. Xu et al. [211], for example, use a GAN framework with one generator and two discriminators. One discriminator is trained to distinguish whether a generated sample is real or fake (as in conventional GANs), and the second identifies whether the sample belongs to the privileged or unprivileged group.

Some of these methods make efforts to also preserve semantic information of the data while learning fair representations [170, 182]. This is essential since the interpretability and transparency of how fairness is approached in algorithmic decision making are crucial to enhancing the understanding of decisions and trust in algorithms, and it is a paramount challenge that future research should face.

Some other papers have studied fair adversarial learning. Beutel et al. [23] investigate the effect of input sample selection on the resulting fairness in adversarial fair learning models and show that a small balanced dataset can be effective for achieving fair representations. Bose and Hamilton [28] extend fair adversarial learning to improve fairness in *graph embedding*. Beutel et al. [22] have raised some concerns about fair adversarial learning that should be further investigated. They argue that these models may sometimes be unstable and therefore may present some risks when using them in a production environment.

Note that other closely related problems are aimed at finding equilibrium (minimax) points in the field of *game theory*, and therefore game-theoretic schemes may also be used for solutions [4, 5, 40, 76].

From the literature on adversarial learning, we can also note that learning fairly to predict the outcome of an unprivileged group can be thought of as learning to predict in a different domain [66, 141], and therefore notions from the field of *domain adaptation* can be adopted to enhance the study of the field of *algorithmic fairness* and vice versa.

5.3 Fair Word Embedding

Word embedding models construct representations of words and map them to vectors (also commonly referred to as *word2vec* models). Training of word embeddings is performed using raw textual data with an extremely large number of text documents and is based on the assumption that words that occur in the same contexts tend to have similar meanings. These models are primarily designed such that the embedding vectors will indicate something about the meanings and relationships between words (i.e., words with similar meanings have vectors that are close in the vector space). As such, they are broadly used in many natural language processing applications, such as in search engines, machine translation, resume filtering, job recommendation systems, online reviews, and more.

However, previous studies have shown that there are inherent biases in word embeddings (e.g., [26, 32, 37, 223]). These studies showed that word embedding models have exhibited social biases and gender stereotypes. For instance, it has been shown that the embedding of "computer programmer" is closer to "male" than it is to "female." The implications of this are disturbing since these biases may affect people's lives and cause discrimination in social applications such as in job recruitment or school admissions. Another example is Microsoft's AI chat bot, named *Tay*, which learned abusive language from Twitter data after only the first day of release (the bot was eventually shelved; see [101]).

A common definition of gender bias in word embedding is the measure of cosine similarity from a selected word to the words "he" and "she" (or any other two pronouns that indicate gender, e.g., "him"/"her" or "Mr."/"Mrs."). The difference between these two similarities may indicate the extent

of bias in the embedding model. For example, consider the sentence "The *CEO* raised the salary of the receptionist because *he* is generous." In this sentence, "he" refers to "CEO," and further similar references in many additional sentences and texts may create a large contextual connection between these two words or other occupational nouns [223]. Hence, even if the algorithms themselves are not biased, historical biases and norms may be embedded into the algorithm results.

To mitigate these types of biases, several studies have developed methods for debiasing the results. For example, Bolukbasi et al. [26] suggest a post-process mechanism for removing gender bias, referred to as *hard-debiasing*. Their method first identifies a *gender dimension*, which is determined by a set of words that indicate gender definitions (e.g., "he"/"she"). Second, it inherently defines *neutral words* (e.g., occupations) and then negates the projection of all of these neutral words with respect to the gender direction (so that the bias of neutral words is now zero by definition) by re-embedding the word w:

$$\vec{w} := (\vec{w} - \vec{w_B}) / \left\| \vec{w} - \vec{w_B} \right\|, \tag{7}$$

where $\vec{w}$ is the embedding of the selected word and $\vec{w_B}$ is the projection of w with respect to the gender direction.

The same paper also suggests an additional *soft-debiasing* mechanism that reduces bias while still maintaining some similarity to the original embedding, thus providing a parameter that controls the trade-off between debiasing and the need to preserve information.

Zhao et al. [223] suggest an in-process mechanism, referred to as *Gender-Neutral Global Vectors* (*GN-GloVe*), to reduce bias. Their method trains word embedding using GloVe [164] with an altered loss function that constructs an embedding such that the protected attribute (e.g., gender) is represented in a certain dimension (a sub-vector of the embedding), which can then be ignored for debiasing. This is done by encouraging pairs of words with gender indication (e.g., "mother" and "father") to have larger distance in their gender dimension and gender-neutral words to be orthogonal to the gender direction.

Brunet et al. [32] propose a pre-process mechanism for reducing bias by perturbing or removing documents during the training stage that are traced as the origin for the word embedding bias.

One challenge of these fairness mechanisms is the need for lists of words that indicate the sensitive attribute dimension and words that should be considered as neutral. Bolukbasi et al. [26], for example, tackle this challenge by using an initial set of gender-definitional words and train an SVM to expand the list.

Let us note that the challenges of fair word embedding also affect many other downstream applications that use these word representations, for example, in *coreference resolution* [173, 220, 222], in *sentence encoding* [146], in *machine translation* [75, 203], in *language models* [27], and in *semantic role labeling* [221].

Curiously, a recent study has argued that a major concern is that some of the proposed methods for removing biases in word embeddings are actually not able to remove biases but rather just "hide" them [85]. They show, by a clustering illustration, that gender biases are still reflected in the embedding even after applying these methods. They additionally show that by using an SVM classifier, most of the gender information can be recovered from the embedding.

Hence, it seems that existing methods as well as definitions for fair embeddings might be insufficient, and these challenges require more extensive research.

5.4 Fair Visual Description

The study of fairness in CV has recently gained extensive interest since CV models have been shown to produce disturbingly biased results with respect to several tasks. For example, Buolamwini and Gebru [33] have found that facial analysis models were negatively affected toward

discriminating results by the under-representation of female dark-skinned faces in datasets. Kay et al. [117] show that image searches of occupations in Google's engine resulted in gender-biased results. Google's labeling application has recklessly identified black Americans as "gorillas" [160, 188]. Furthermore, an app that classified the attractiveness of individuals from photos turned out to be discriminative against dark skin [144].

There are several previous papers in the domain of fair image classification [33, 62, 170, 182]. However, the task becomes much more complex when a fair description of images is required, such as in *multi-label classification* tasks [117, 193, 202, 221], in *face attribute detection* [116, 177], or in the task of *image captioning* [96]. The last task is even more complicated than the others because of the unstructured character of the problem.

Mitigating bias in these types of problems is challenging for several reasons. First, the multi-modal nature of the task requires handling fairness at both levels of natural language processing models and CV models. As mentioned in a previous section, the challenges of fair word embedding also affect many other downstream applications that use these word representations (e.g., see *image captioning* in the work of Hendricks et al. [96]). Second, the labels of the data usually depend on annotators and are not always accurate [193, 202]. For example, Stock and Cisse [193] show that human annotators fail to capture parts of the visual concepts in images, and Van Miltenburg [202] show that the crowdsourced descriptions of the images in the *Flickr30K dataset* are often based on stereotypes and prejudices of the annotators; another challenge is that the datasets are inherently unbalanced since to have balance, all possible co-occurrences must be balanced [96, 202]

Zhao et al. [221] consider multi-label role classification, show that some tasks display severe gender bias, and suggest a method to re-balance the resulting predictions by solving a constrained optimization problem using Lagrange relaxation. Additional constraints require the model predictions to have a similar distribution as the training set in terms of co-occurrences of gender indication and the predicted target value (this is applied to the entire corpus level since the entire corpus is needed to assess the frequency of occurrences).

Hendricks et al. [96] consider the problem of fair image captioning. They propose a method to reduce gender bias by directly using a person's appearance information in the image. The method is designed to be more cautious when there is no gender information in the image. Moreover, it is constructed to function even when the distributions of genders differ between training and test sets.

5.5 Fair Recommender Systems

Recommender systems are prevalent in many automated and online systems and are designed to analyze users' data to provide them with personalized propositions that correspond to each user's tastes and interests. An inherent concept in recommendations is that the best items for one user may be different from those for another. Some examples of recommended items are movies, news articles, products, jobs, and loans, among others. These systems have the potential to facilitate activities for both providers and consumers; however, they have also been found to exhibit fairness issues [34, 35, 67, 68]. For instance, it was shown that Google's ad-targeting algorithm had proposed higher-paying executive jobs more commonly for men than for women [56, 187].

Most recommender systems employ user and item similarities and are therefore prone to result in homogeneous selections that may not provide sufficient opportunities for minority populations.

A recent paper, by Burke [34], notes that extending the notion of fairness from general classification tasks to recommender systems should take personalization into account. Several studies are investigating the user's perspective [41, 65, 67, 212], where fairness is considered as recommending items equitably to different user groups (referred to as *C-fairness*), such as recommending high-paying jobs to both men and women. Other studies refer to the provider's perspective

[68, 132], where items from different providers should be recommended equally (referred to as *P-fairness*), for instance, when trying to avoid market monopolistic domination. Burke [34] and Burke et al. [35] note that many recommender system applications involve multiple stakeholders and may therefore give rise to fairness issues for more than one group of participants simultaneously, as well as achieving fairness at a regulatory level or the level of the entire system (referred to as *multisided fairness*).

It is interesting to note that *P-fairness* is somewhat related to the *categorical diversity* in recommender systems, requiring that recommendation lists are diverse [127, 136]. For example, consider a hiring recommender system—we may observe all male candidates as items of one provider and all female candidates as items of another provider. We may then ensure the equal recommendation of men and women to each of the positions. In diversity-enhancing methods, some common measures of equality may be considered, such as the Gini coefficient that is also used in economic contexts [127]. Moreover, note that an *individual* definition of *P-fairness*, rather than *group-fairness*, may be somewhat similar to the definition of *coverage* in recommender systems, requiring that each *item* be recommended fairly [34].

Sürer et al. [195] propose enhancing multi-stakeholder fairness using a constraint-based integer programming optimization model. The problem is computationally difficult, and hence a following relaxation heuristic is proposed to solve it. Edizel et al. [65] suggest a post-processing mechanism that alters a fraction of the entries in the recommendation matrix so that it would be more difficult to predict the sensitive attributes from the matrix while preserving the high utility of the matrix. The ε*-fairness* of a recommendation matrix with respect to a certain sensitive attribute is defined as the error level when predicting the sensitive attribute using the recommendation matrix. The price of achieving ε*-fairness* is measured by the distance between the two matrices: the original one obtained from the prediction algorithm and the altered one obtained after the post-processing mechanism is performed.

Yao and Huang [212] and Kamishima et al. [113] suggest an in-process mechanism by including additional regularization factors in the objective functions. Kamishima et al. [113] introduce the notion of *recommendation independence* to fairness-aware recommender systems. This notion requires statistical independence between recommendation results and the sensitive attribute. This means that the sensitive information will not affect the results. Yao and Huang [212] define several notions of fairness in recommender systems. *Value unfairness* measures the inconsistency in the signed estimation error across user groups. *Absolute unfairness* is similar to the previous measure but with absolute estimation. *Underestimation unfairness* represents the extent to which predictions underestimate the true ratings. *Overestimation unfairness* is similar to the previous measure but with overestimation. Celis et al. [39] and Bredereck et al. [30] use similar notions for fairness in multiwinner voting. Note that many of these problems have a multi-objective nature, which requires to balance multiple trade-offs (e.g., see [148]).

Most recommender systems implement ranking procedures to present the best items. Therefore, studies have investigated fair ranking notions for recommender systems. For example, group-based fairness notions in ranking [21, 69, 217], individual-based fairness notions in ranking [25], or both group and individual fairness simultaneously [87]. Several studies explore the notion of *fair exposure* of items in ranking based on their merit [25, 87, 189] (e.g., fair ranking of job candidates). Others aim to ensure sufficient representation of items from different groups in the top-k rankings (e.g., [41, 217]). Beutel et al. [21] propose several measures based on *pairwise fairness*. According to their notion, a ranking function is considered to obey pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item is the same across groups (other works that follow the notion of pairwise fairness are, e.g., those of Kuhlman et al. [126] and

Narasimhan et al. [156]). Biega et al. [25] introduce the notion of *amortized individual fairness*, which measures accumulated relevance across a series of rankings dynamically over time.

One challenge in fairness-aware recommender systems is the possibility of discriminated groups based on more than one sensitive attribute [65]. It is required to first devise a definition for fairness in such cases to address the problem. In multi-stakeholder fairness optimization, the problems are computationally difficult [195], so there is a need for more computational enhancements to work with large datasets. Future work may also focus on fairness in systems where there are network structures that define relationships between providers and between users (e.g., [28]). Another possible research direction may be the incorporation of sequential notions of fairness into recommender systems through the introduction of additional time-dependent constraints [163, 194].

Note that the domain of recommender systems is also closely related to other common multi-stakeholder environments [1, 2], like *resource allocation* problems such as police distribution to districts [70], the allocation of aid in disaster response, where fairness is also a major concern, or other allocation problems such as fair allocation of indivisible goods [162].

5.6 Fair Causal Learning

Observational data collected from real-world systems can mostly provide associations and correlations rather than causal structure understandings. In contrast, causal learning relies on additional knowledge structured as a model of causes and effects.

One limitation of measures that are based solely on observable data is that they do not consider the mechanisms by which the data is generated and therefore may have wrong interpretations [143]. Moreover, as mentioned in Section 3, there is a challenge of incompatibility of fairness notions. Another limitation of observational measures is that they can be highly affected by missing data, as mentioned in Section 4.4.

Causal approaches may assist in enhancing fairness in several manners. For instance, by understanding causes and effects in the data, the model may assist in tackling the challenges of fairness definitions by analyzing which types of discrimination should be allowed and which should not [128, 174]. Another approach to improve fairness using causal reasoning is to provide an understanding of how to perform imputation of missing values or how to repair a dataset that contains sample or selection bias [13, 191]. Moreover, understanding a causal model may help with other ethical issues, such as defining liability and responsibility by understanding the sources of biases. This may increase the transparency and explainability of the fair models, which is also crucial for trust.

Kusner et al. [128] have suggested a measure of causal fairness called *counterfactual fairness*. It measures the extent to which it is possible to build two identical predictions $\hat{Y}$—one trained on the privileged group and the second trained on the unprivileged group—using any combination of variables in the system that are not caused by the sensitive attribute. The intuition is that, in a fair model, the prediction would not change if only the sensitive attribute (and its affected variables) is changed. More precisely, a causal graph satisfies *counterfactual fairness* when the predicted label is not dependent on any descendant of the sensitive attribute.

Several studies have proposed alternative notions to *counterfactual fairness*, which relax the rigid restriction on any descendant to less strict limitations. For example, the graph does not suffer from *proxy discrimination* [120] if the predicted label is not dependent on any *proxy* of the sensitive attributes (a *proxy* feature is a feature that can be exploited to derive the sensitive feature). Moreover, the graph does not suffer from *unresolved discrimination* if the predicted label is not dependent on any *resolving* variable (a *resolving* variable is influenced by the sensitive feature but is accepted by practitioners as non-discriminatory). Chiappa [43] introduces a notion of *path-specific*

counterfactual fairness. This notion assists in identifying whether an indirect path between a sensitive attribute and the outcome is unfair.

Commonly, causal methods for fairness mitigation are aimed at pre-processing the training data [44, 81, 83, 122, 128, 155, 174, 179]. This is reasonable considering, for example, that including proxies of the sensitive attributes in the learning model can cause unfairness. Hence, using causal structure knowledge to remove or modify proxy variables can assist in mitigating unfairness (e.g., see [83, 178]). Moreover, in a recent work, Salimi et al. [178, 179] suggest to use causal knowledge on dependencies between sensitive and non-sensitive attributes to add or remove samples from the training dataset in a manner that improves fairness.

Additional Intuition and Examples. The need for causal models for identifying and mitigating unfairness is based on the intuition that unfairness is reflected in situations where individuals experience different outcomes due to factors outside of their control (factors like gender or race) [137]. Causal models are therefore useful for investigating which of the factors cannot be controlled by individuals and use the resulted understandings to identify and deal with unfairness.

One of the common examples that is used to illustrate the importance of causality to identify and mitigate unfairness is the gender bias in Berkeley admission in 1973 [24, 137]. This example demonstrates how causal models can be useful for the removal of the preceding uncontrolled factors.

In the discussed case, the admission rates for studies at Berkeley were about 44.3% of the men who applied and about 34.6% of the women who applied. This raises a question as to whether the difference in admission rates is evidence of an existing bias toward women. An additional question is if there is bias indeed, where does it stem from—whether it is a discriminatory activity of the educational institution, or alternatively the cause is from another origin, that is not evident without investigating the causal structure of the problem more deeply. After examining the case more deeply, it can be seen that the decisions made in each department separately had the same admission rates of men and women [24]. However, it can be further noticed that a larger proportion of women had applied for the most selective departments, resulting in the observed difference in admission rates.

In other words, understanding the structure of the root causes of the problem can assist in identifying unfairness and interpreting where it can be originated from. In the Berkley example, the origin appears to be the socialization of women over the years that motivated their choice for the field of studies [24, 137]. So, in this example, there is a causal structure that must be taken into account rather than just the correlation between the sensitive attribute and the outcome.

Another example in which causal models are necessary is the situation where a selection bias may exist. A selection bias is a situation in which a group of individuals may experience underrepresentation in the selected population—for example, in a case of historically hiring more men than women to technological positions [54].

In the case of a selection bias, relying solely on the observable data, with no additional causal information, is limited, as the dataset represents only the selected population, without any information on the groups who were not selected. Hence, the interpretation could be wrong without additional knowledge external to the database describing causes and effects. Such information can be achieved using knowledge of a causal graph or by a controlled experiment making use of interventions.

It is important to note that causal fairness models can indeed help us overcome many of the challenges encountered with respect to fair prediction tasks; however, in practice, it is difficult to obtain the correct causal model. Moreover, removing all correlated features found through a causal model may significantly compromise accuracy.

We presented here key notions of causal fairness. Interested readers are referred to the work of Loftus et al. [137] for further information about categorization of causality-based notions, Makhlouf et al. [143] for causal measures, and Barocas et al. [14] and Chiappa and Isaac [45] for the relationships between causality and fairness.

5.7 Fair Private Learning

Dwork et al. [60] introduced a discussion about the relationship between privacy and fairness. We note that algorithmic fairness research is very closely related to *privacy* research since both fairness and privacy can be enhanced by obfuscating sensitive information, with the adversary goal of minimal data distortion [65, 118]. Moreover, a violation of privacy (e.g., as captured by the term *inferential privacy* [53, 64, 82]) can lead to unfairness due to the ability of an adversary to infer sensitive information about an individual and use it in a discriminating manner.

There were several recent studies that incorporated both privacy and fairness considerations to ML algorithms. Most of these studies have investigated the case of a centralized setting, in which all the dataset is held by a single party [10, 52, 100, 106, 153, 210]. The goal of those studies is to train an ML model over the centralized dataset, making sure that the released model and its future outputs are fair, as well as private, in the sense that one cannot infer from them (meaningful) information about individual data records of the dataset. These papers have used *differential privacy* [63] as their approach, in which the underlying idea is to modify the values of attributes randomly so that privacy will be maintained, whereas the results of the desired analysis will still be close to the original ones. An algorithm is considered differentially private if the addition or removal of a single record from the dataset does not significantly affect the output of the algorithm. Typically, differential privacy is obtained by incorporating a Laplacian noise to the results of the data analysis [58].

A few other studies have investigated a distributed setting that includes a "main" party that wishes to train a fair ML model over the data it holds, and a third party to which the sensitive attributes are outsourced [73, 98, 121]. The motivation behind this setting is that although the sensitive attributes should not be exposed to the main party, they should still be used in the training process of the model (in a privacy-preserving manner) to ensure that the resulting model is fair. To obtain this goal, these studies used either random projections [73, 98] or **secure multi-party computation (SMC)** techniques [121]. Pessach et al. [166] proposed a collaborative private pre-process fairness-enhancing mechanism based on SMC that does not require a third party.

It is worth noting that the term *fair* is also used in the cryptographic literature in a different sense. For example, an SMC protocol [134] is considered fair, in that context, if it ensures that either all parties receive their designated outputs or none of them does [58]. The notion of fairness discussed in this article is different from the SMC fairness discussed previously.

6 DISCUSSION AND CONCLUSION

In this article, we presented a comprehensive and up-to-date overview of the algorithmic fairness research field. We started by describing the main causes of unfairness, followed by common definitions and measures of fairness, and the inevitable trade-offs between them. We then presented fairness-enhancing mechanisms, focusing on their pros and cons, aiming at better understanding which mechanisms should be used in different scenarios. Last, we listed several emerging research sub-fields of algorithmic fairness including fair sequential learning, fair adversarial learning, fair word embedding, fair visual description, fair recommender systems, fair causal learning, and fair private learning. Overall, this survey provides the relevant knowledge to enable new

researchers to enter the field and inform current researchers on rapidly evolving sub-fields. Interested readers can expand their in-depth understanding with the assistance of the work of Barocas et al. [14].

This study was largely based on the Google Scholar search engine (see Appendix B.5), which has some limitations (a limit of 1,000 visible search results for each query, lack of stability, incomplete coverage, etc.). However, we note that a study from 2014 [119] estimated that Google Scholar covers more than 87% of English scholarly documents available on the web. Moreover, a recent comparative research [92] assessed that Google Scholar's size might have been underestimated previously and is currently the most comprehensive academic search engine.

In addition to the already studied topics and the emerging ones, we identify several open challenges that make fairness in ML still hard to achieve and should be further investigated in future research. One major challenge stems from biases inherent in the dataset. Such biases may arise, for example, when the labeling process was performed in an already unfair manner, or if there are under-represented populations in the dataset, or in the case of systematic lack of data and in particular labels. Representative datasets are difficult to achieve, and therefore it is crucial to devise methods to overcome these issues.

Another challenge is the proliferation of definitions and measures, fairness-related datasets, and fairness-enhancing mechanisms. It is not clear how newly proposed mechanisms should be evaluated, and particularly which measures should be considered, which datasets should be used, and which mechanisms should be used for comparison. A closely related challenge is the difficulty in determining the balance between fairness and accuracy. In other words, what are the costs that should be assigned to each of these measures for evaluation purposes? Future efforts should be invested in generating a benchmarking framework that will allow a more unified and standard evaluation process for fairness mechanisms.

The interpretability and transparency of how fairness is addressed by ML algorithms impose another important challenge. Such transparency is crucial to increase the understanding and trust of users in these algorithms and in many domains is even required by law. This need is further supported by several recent studies that have addressed the question of how devised mathematical notions of fairness are perceived by users [90, 192]. It turns out that users tend to prefer the simpler notion of demographic parity, probably due to the difficulty of grasping more complex definitions of fairness.

A promising solution to some of the preceding challenges can be found in causality-based approaches. For instance, a better understanding of the origin of unfairness can improve explainability and guide in choosing suitable measures and mechanisms. Additionally, causality can provide tools to assess fairness when the dataset includes missing data or a selection bias (Section 5.6). We note that causality-based approaches, although promising, still have challenges to overcome (e.g., correctly identifying the origin of unfairness can be challenging or even infeasible).

Moreover, to deal with the trade-offs between multiple objectives, we recommend that achieving fairness should be treated as a multi-objective task, weighting the different objectives according to the proper legal, social, and ethical contexts. Furthermore, we suggest monitoring the long-term effect of the changes entailed by the chosen fairness mechanism while paying attention to feedback loops that can unintentionally introduce self-fulfilling predictions (see the policing example in Section 5.1).

To conclude, since the use of algorithms is expanding to all aspects of our lives, demanding that automated decisions be more ethical and fair is inevitable. We should aspire to not only develop fairer algorithms but also to design procedures to reduce biases in the data. Such procedures may rely, for example, on integrating both humans and algorithms in the decision pipeline. However, thus far, it seems that biased algorithms are easier to fix than biased humans or procedures [154].

APPENDIX

A FAIRNESS-RELATED DATASETS

In this appendix, we review the most commonly used datasets in the literature of algorithmic fairness. Most of these datasets are publicly available, and we further indicate this for each of these datasets in their description.

ProPublica Risk Assessment Dataset. The ProPublica dataset includes data from the COMPAS risk assessment system (see [6, 103, 130]).

This dataset was previously extensively used for fairness analysis in the field of criminal justice risk [20]. The dataset includes 6,167 individuals, and the features in the dataset include number of previous felonies, charge degree, age, race, and gender. The target variable indicates whether an inmate recidivated (was arrested again) within 2 years after release from prison.

As for the sensitive variable, this dataset was previously used with two variations: the first when race was considered as the sensitive attribute and the second when gender was considered as the sensitive attribute [17, 38, 71, 79, 145]. The dataset is publicly available from Larson et al. [131].

Adult Income Dataset. The Adult dataset is a publicly available dataset in the UCI repository [59, 183] based on 1994 U.S. census data. The goal of this dataset is to successfully predict whether an individual earns more or less than $50,000 per year based on features such as occupation, marital status, and education. The sensitive attributes in this dataset include age [139], gender [218], and race [79, 145, 215].

This dataset is used with several different pre-processing procedures. For example, the dataset of Zafar et al. [215] includes 45,222 individuals after pre-processing (48,842 before pre-processing).

German Credit Dataset. The German dataset is a publicly available dataset in the UCI repository [59, 186] that includes information of individuals from a German bank in 1994.

The goal of this dataset is to predict whether an individual should receive a good or bad credit risk score based on features such as employment, housing, savings, and age. The sensitive attributes in this dataset include gender [79, 139] and age [110, 218]. This dataset is significantly smaller, with only 1,000 individuals with 20 attributes.

Ricci Promotion Dataset. The Ricci dataset includes the results of an exam administered to 118 individuals to determine which of them would receive a promotion. The dataset originated from a case that was brought to the U.S. Supreme Court [150, 176]. The goal of this dataset is to successfully predict whether an individual receives a promotion based on features that were tested in the exam, as well as the current position of each individual. The sensitive attribute in this dataset is race. The dataset is publicly available from Friedler et al. [78].

Mexican Poverty Dataset. The Mexican poverty dataset includes poverty estimation for determining whether to match households with social programs. The data originated from a survey of 70,305 households in 2016 [102]. The target feature is poverty level, and there are 183 features. This dataset was studied, for example, in the work of Bakker et al. [11] and Noriega-Campero [157]. The authors studied two sensitive features: young and old families, as well as urban and rural areas.

Diabetes Dataset. The Diabetes dataset includes hospital data for the task of predicting whether a patient will be readmitted. It is publicly available in the UCI repository [59, 185]. The data contain

approximately 100,000 instances and 235 attributes. This dataset was studied, for example, in the work of Edwards and Storkey [66], where race was the sensitive feature.

Heritage Health Dataset. The Heritage health dataset originated from a competition conducted by the United States as a competition to improve healthcare through early prediction. It includes data of 147,473 patients with 139 features. The goal of this dataset is to predict whether an individual will spend any days in the hospital during the next year [31]. This dataset was studied, for example, in the work of Zemel et al. [218], Louizos et al. [139], and Tramer et al. [198], where age was the sensitive feature. The dataset is publicly available from Heritage Health Prize Contest Data [55].

College Admissions Dataset. The College Admissions dataset was collected by the UCLA law school [181]. It includes data from more than 20,000 records of law school students who took the bar exam. The goal of this dataset is to predict whether a student will pass the exam based on factors such as LSAT score, undergraduate GPA, and family income.

This dataset was used, for example, by Berk et al. [19], where gender was studied as the sensitive feature, and Bechavod and Ligett [17], where race was studied as the sensitive feature.

Bank Marketing Dataset. The Bank Marketing dataset is a publicly available dataset in the UCI repository [59, 152, 184], and it includes 41,188 individuals with 20 attributes. The task is to predict whether the client has subscribed to a term deposit service based on features such as marital status and age. It was previously investigated by Zafar et al. [215], where age was studied as the sensitive attribute.

Loans Default Dataset. The Loans Default dataset includes 30,000 instances and 24 attributes of credit card users. It is publicly available in the UCI repository [59, 159, 213]. The goal is to predict whether a customer will default on payments. The features include age, gender, marital status, past payments, credit limit, and education.

This dataset was used, for example, by Bechavod and Ligett [17] and Yeh and Lien [213], where gender was studied as the sensitive feature.

Dutch Census Dataset. The Dutch Census dataset includes 189,725 instances and 13 attributes of individuals. It is publicly available in the IPUMS repository [42]. Kamiran and Calders [111] and Agarwal et al. [4] use this dataset with only the 60,420 individuals who are not underaged. Their goal is to predict whether an individual holds a highly prestigious occupation by using features such as gender, age, household details, location, citizenship, birth country, education, economic status, and marital status. The sensitive feature utilized is gender.

Communities and Crimes Dataset. The Communities and Crimes dataset includes 1,994 instances and 128 attributes of communities in the United States. It is publicly available in the UCI repository [49, 59, 171]. The goal is to predict the number of violent crimes per 100,000 individuals based on features such as percentage of population by age, by marital status, by number of children, by race, and more. Kamiran and Calders [111] add a new sensitive attribute that represents whether the percentage of the African-American population in the community is greater than 0.06.

Table 3 presents a summary of the benchmark datasets for algorithmic fairness that were discussed in this section. We note that the reviewed datasets may have some limitations that were not analyzed in this appendix, such as missing values and duplicate instances.

Table 3. Common Benchmark Datasets for Algorithmic Fairness

Dataset Name	Domain	# Records	Sensitive Attributes	Target Attributes	Publicly Available
ProPublica [6, 130, 131]	Criminal risk assessment	6,167	Race; Gender	Whether an inmate has recidivated (was arrested again) in less than 2 years after release from prison	✓
Adult [59, 183]	Income	48,842	Age; Gender	Whether an individual earns more or less than $50,000 per year	✓
German [59, 186]	Credit	1,000	Gender; Age	Whether an individual should receive a good or bad credit risk score	✓
Ricci [78, 150, 176]	Promotion	118	Race	Whether an individual receives a promotion	✓
Mexican poverty [102]	Poverty	183	Young and old families; Urban and rural areas	Poverty level of households	✗
Diabetes [59, 185]	Health	100,000	Race	Whether a patient will be readmitted	✓
Heritage health [31, 55]	Health	147,473	Age	Whether an individual will spend any days in the hospital in the next year	✓
College Admissions [181]	College Admissions	20,000	Gender; Race	Whether a law student will pass the bar exam	✗
Bank Marketing [59, 152, 184]	Marketing	41,188	Age	Whether the client subscribed to a term deposit service	✓
Loans Default [59, 159, 213]	Loans	30,000	Gender	Whether a customer will default on payments	✓
Dutch Census [42]	Census	189,725	Gender	Whether an individual holds a highly prestigious occupation	✓
Communities and Crimes [49, 59, 171]	Crime	1,994	Percentage of African-American population	For each community, the number of violent crimes per 100,000 individuals	✓

B SUPPLEMENTARY MATERIALS

B.1 Additional Measures for Algorithmic Fairness

Section 3 discussed the most prominent definitions and measures of algorithmic fairness. This appendix reviews additional measures used in the literature:

(1) *Overall accuracy equality*: This requires similar accuracy across groups [20]. This measure is mathematically formulated as follows:

$$\left|P[Y = \hat{Y}|S = 1] - P[Y = \hat{Y}|S \neq 1]\right| \leq \varepsilon, \tag{8}$$

where S represents the sensitive attribute (e.g., race and gender), $S = 1$ is the privileged group, and $S \neq 1$ is the unprivileged group. $\hat{Y} = Y$ means that the prediction was correct. A lower value indicates better fairness. It should be noted that this measure does not guarantee *equalized odds* or fair decisions (see [130]).

(2) *Predictive parity*: This requires that the **positive predictive values (PPVs)** are similar across groups (meaning the probability of an individual with a positive prediction actually experiencing a positive outcome) [47]. This measure is mathematically formulated as follows:

$$\left|P[Y = 1|S = 1, \hat{Y} = 1] - P[Y = 1|S \neq 1, \hat{Y} = 1]\right| \leq \varepsilon. \tag{9}$$

Note that a lower value indicates better fairness. This measure uses the ground truth of the outcome, assuming that the outcome was achieved fairly. However, it has been shown to be incompatible with *equalized odds* and *equal opportunity* when prevalence differs across groups [47, 50, 51].

(3) *Equal calibration*: This requires that, for any predicted probability value, both groups will have similar PPVs (PPV represents the probability of an individual with a positive prediction actually experiencing a positive outcome) [47, 125]. Note that this measure is similar to *predictive parity* when the score value is binary (but does not guarantee predictive parity when the score is not binary) [47]. This measure is mathematically formulated as follows:

$$|P[Y = 1|S = 1, V = v] - P[Y = 1|S \neq 1, V = v]| \leq \varepsilon, \tag{10}$$

where V is the predicted probability value. Note that in some studies, the definition of calibration requires that the PPV also be equal to V [50, 125]. A lower value indicates better fairness. Although in some cases equal calibration may be the desired measure, it has been shown that it is incompatible with *equalized odds* [168] and is insufficient to ensure accuracy or equitable decisions [50]. Moreover, it conflicts with *balance for the positive class* and *balance for the negative class* [47, 51].

(4) *Conditional statistical parity*: Controlling for a limited set of "legitimate" features, an equal proportion of individuals is selected from each group [51]. This measure is mathematically formulated as follows:

$$\left|P[\hat{Y} = 1|S = 1, L = l] - P[\hat{Y} = 1|S \neq 1, L = l]\right| \leq \varepsilon. \tag{11}$$

L is a set of legitimate factors. A lower value indicates better fairness. Note that using this measure requires defining which features are legitimate, which is not a trivial task. It is not practical to find features that are entirely independent of the sensitive attributes.

(5) *Predictive equality*: This requires FPRs (meaning the probability of an individual with a negative outcome to have a positive prediction) to be similar across groups [51]. This measure is mathematically formulated as follows:

$$\left|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]\right| \leq \varepsilon. \tag{12}$$

This measure requires the ground truth of the outcome, assuming that the outcome was achieved fairly. A lower value indicates better fairness. However, it considers only one type of error (as opposed to *equalized odds*, e.g., which requires the equality of both FPRs and FNRs). As mentioned, following equality in terms of only one type of error will increase the disparity in terms of the other error [168].

(6) *Conditional use accuracy equality*: This requires PPVs and **negative predictive values (NPVs)** to be similar across groups [20]. NPV represents the probability of an individual with a negative prediction actually experiencing a negative outcome. PPV represents the

probability of an individual with a positive prediction actually experiencing a positive outcome. This measure is mathematically formulated as follows:

$$\begin{gathered}\left|P[Y=1|S=1,\hat{Y}=1]-P[Y=1|S\neq 1,\hat{Y}=1]\right|\leq\varepsilon\\ \wedge\\ \left|P[Y=0|S=1,\hat{Y}=0]-P[Y=0|S\neq 1,\hat{Y}=0]\right|\leq\varepsilon.\end{gathered}\tag{13}$$

This measure requires the ground truth of the outcome, assuming that the outcome was achieved fairly. A lower value indicates better fairness. This measure considers more than one type of error. According to Berk et al. [20], following this measure does not guarantee *equalized odds*.

(7) *Treatment equality:* This requires an equal ratio of FNs and FPs [20]. The FN cases are all of the cases that were predicted to be in the negative class when the actual outcome belongs to the positive class. The FP cases are all of the cases that were predicted to be in the positive class when the actual outcome belongs to the negative class. This measure is mathematically formulated as follows:

$$|FN_{S=1}/FP_{S=1}-FN_{S\neq 1}/FP_{S\neq 1}|\leq\varepsilon.\tag{14}$$

This measure requires the ground truth of the outcome, assuming that the outcome was achieved fairly. A lower value indicates better fairness. Note that according to Berk et al. [20], following this measure may harm the *conditional use accuracy equality*.

(8) *Balance for the positive class*: This requires an equal mean of predicted probabilities for individuals that experience a positive outcome [125]. This measure is mathematically formulated as follows:

$$|E[V|Y=1,S=1]-E[V|Y=1,S\neq 1]|\leq\varepsilon,\tag{15}$$

where V is the predicted probability value. A lower value indicates better fairness. This measure was proven to be incompatible with *equal calibration* [125].

(9) *Balance for the negative class*: This requires an equal mean of predicted probabilities for individuals that experience a negative outcome [125]. This measure is mathematically formulated as follows:

$$|E[V|Y=0,S=1]-E[V|Y=0,S\neq 1]|\leq\varepsilon,\tag{16}$$

where V is the predicted probability value. A lower value indicates better fairness. This measure was proven to be incompatible with *equal calibration* [125].

(10) *Fairness through unawareness*: This requires that no sensitive attributes are explicitly used in the algorithm. The predicted outcomes are the same for candidates with the same attributes [128]. This measure is mathematically formulated as follows:

$$X_i=X_j\rightarrow\hat{Y}_i=\hat{Y}_j,\tag{17}$$

where i and j denote two individuals, and X are the attributes describing an individual except for the sensitive attributes. This measure requires predictions to be the same for candidates with the same attributes. However, note that even when not considering the sensitive attributes, the model could still be biased through "proxies" or other causes such as sample or selection bias. Moreover, explicitly considering sensitive attributes is sometimes required, and excluding information can lead to discriminatory decisions [50].

(11) *Mutual information*: This measures the mutual dependence between the sensitive feature and the predicted outcome [113]. This measure is mathematically formulated as follows:

$$\sum \left(P(\hat{y}, s) log(\frac{P(\hat{y}, s)}{P(\hat{y})P(s)}) \right) \leq \varepsilon. \tag{18}$$

Note that this measure does not consider the actual outcomes. The lower the measure is, the lower the dependence between the sensitive attribute and the predictions; thus, lower values represent better fairness. One advantage of this measure is that it can consider binary, categorical, or numerical predictions.

(12) *Mean difference*: This measures the difference between the means of the predictions across groups [225]. For example, Louizos et al. [139] use a variation of this measure that computes the difference between the means of the predicted probabilities across groups.

$$\left| E[\hat{Y}|S = 1] - E[\hat{Y}|S \neq 1] \right| \leq \varepsilon \tag{19}$$

Note that this measure does not consider the actual outcomes and that lower values indicate better fairness. One advantage of this measure is that it can consider binary, categorical, or numerical predictions.

Table 4 presents a summary of the measures described in this appendix.

B.2 Measure Categories

This appendix summarizes insights regarding measures subgroups, trade-offs, and impossibility results discussed in the literature (Figure 2).

B.3 Emerging Notions of Fairness Measures for Classification

Recent works have proposed new approaches for defining fairness notions including *procedural fairness*, preference-based fairness, combined notions for individual and group fairness, and *pairwise fairness*.

Procedural Fairness. Grgić-Hlača et al. [91] and Green and Hu [88] discuss the notion of *procedural fairness*. They review the differences between *distributive fairness* that considers decision outcomes (e.g., disparate impact) and *procedural fairness* that considers the process of decision making. Assessing procedural fairness depends on societal context and subjective human perceptions on the fairness and reliability of the decision-making process. This background knowledge is commonly not captured in the available datasets. For example, Grgić-Hlača et al. [91] collect human judgments through surveys and introduce three fairness measures that take into account the perceived fairness of the features that are used for decision making. Their proposed measures combine people's perceived judgments regarding the fairness of features together with the impact of these features on the fairness of the outcome. They additionally propose in-process fairness mechanisms based on logistic regression that are aimed at selecting fair features for enhancing procedural fairness while balancing the trade-off with accuracy. Their mechanisms are mathematical optimization formulations that are solved by applying a *constrained submodular optimization* [104]. They observe that enhancing procedural fairness requires the removal of highly informative features, and hence might cause a greater reduction in accuracy compared to mechanisms that enhance distributive fairness. Green and Hu [88] argue that procedural fairness and distributive fairness should be addressed together for a more holistic view of fairness in ML, and neither is sufficient on its own. They illustrate their argument, for example, by discussing the limitations of

Table 4. Additional Measures and Definitions for Algorithmic Fairness

Measure	Paper	Description	Type	Uses Actual Outcome	Uses Sensitive Attribute	Type of Actual Outcome	Type of Sensitive Attribute	Equivalent Notions
Overall Accuracy Equality	[20]	Requires similar accuracy across groups	Group	✓	✓	Binary	Binary	
Predictive Parity	[47]	Requires that PPVs are similar across groups	Group	✓	✓	Binary	Binary	• Equal PPV [47] • Equal precision (e.g., see [50]) • Mathematically equal PPVs will induce equal false discovery rates (see [206]).
Equal Calibration	[47], [125]	Requires that for any predicted probability value, both groups will have similar PPV	Group	✓	✓	Binary	Binary	Similar to predictive parity when the score value is binary [47]
Conditional Statistical Parity	[51]	Controlling for a limited set of "legitimate" features, an equal proportion of individuals is selected from each group	Group	✗	✓	-	Binary	Similar to the notion of *fairness through unawareness* (see [51])
Predictive Equality	[51]	Requires that FPRs are similar across groups	Group	✓	✓	Binary	Binary	• FP error rate balance [47, 206] • Equal FPRs [51] • Mathematically equal FPRs will induce equal true negative rates (see [206])
Conditional Use Accuracy Equality	[20]	Requires that PPVs and NPVs are similar across groups	Group	✓	✓	Binary	Binary	
Treatment Equality	[20]	Requires equal ratio of FNs and FPs	Group	✓	✓	Binary	Binary	
Balance for the Positive Class	[125]	Requires equal mean predicted probabilities for individuals who experience a positive outcome	Group	✓	✓	Binary	Binary	
Balance for the Negative Class	[125]	Requires equal mean predicted probabilities for individuals who experience a negative outcome	Group	✓	✓	Binary	Binary	
Fairness through Unawareness	[128]	Requires that no sensitive attributes are explicitly used in the algorithm	Individual	✗	✗	-	-	• Fairness through blindness [51] • Anti-classification [50]
Mutual Information	[113]	Measures mutual dependence between the sensitive feature and the predicted outcome	Group	✗	✓	-	Binary, Numerical, Categorical	Related to prejudice index [113]
Mean Difference	[225]	Measures the difference between the means of the targets across groups	Group	✗	✓	-	Binary, Numerical	• Discrimination probability [139] • Similar to *demographic parity* when the target is binary

individual fairness. Recall that individual fairness (Section 3) requires a definition of a similarity metric to assess the satisfaction of the requirement that similar individuals receive similar treatment. Green and Hu [88] note that this definition of similarity should be based on the social context and judgments of humans, not solely on mathematical measurements.

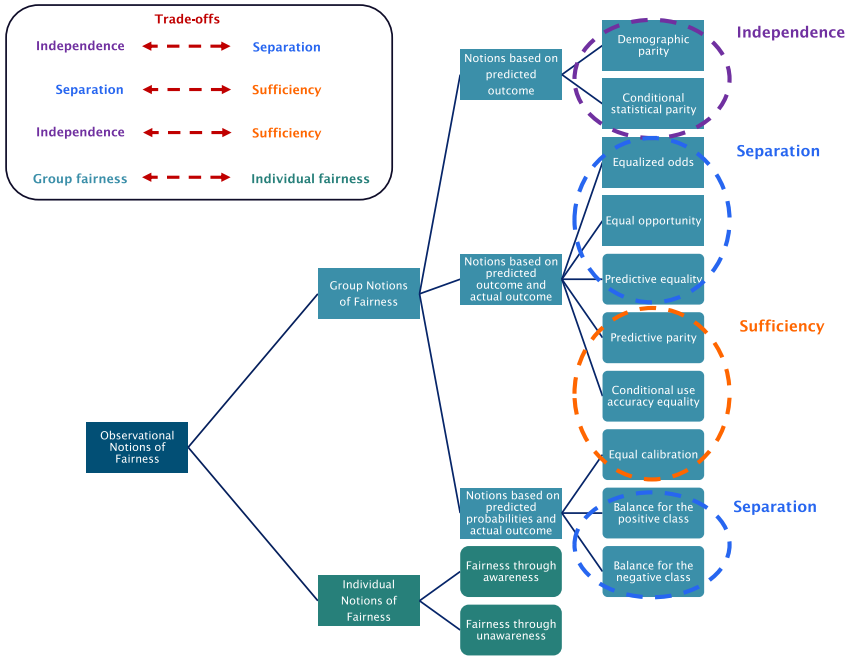

Fig. 2. The figure presents categorization of measures, as well as their classification to the fairness criteria of independence, separation, and sufficiency, as indicated by Barocas et al. [14]. In their analysis, they discuss how these fairness criteria have trade-offs and cannot be satisfied simultaneously. The red dashed lines in the legend represent trade-offs between fairness notions.

Preference-Based Fairness. Zafar et al. [216], Kim et al. [124], and Ustun et al. [199] propose using preference-based fairness measures that leverage notions from the fields of economics and game theory. For example, Zafar et al. [216] propose two preference-based notions of fairness in the context of the preferences of the entire sub-group (e.g., men, women): *preferred treatment* and *preferred impact*. Preferred treatment ensures envy-freeness in a manner that no group would prefer the outcome from the classifier of another group, and therefore no group of users would feel that they would be collectively better off by switching their group membership. Preferred impact requires that the classifier will achieve the best benefit for each of the groups, while ensuring at least the benefit of impact parity classifier (equal fraction of individuals of each group who receive beneficial decision outcomes), but allowing disparities in benefits received by different groups (this is justified when demanding impact parity worsen the results for one group without improving another group). Note that these two notions refer to the preferences of an entire group rather than individuals in the group. They introduce in-process mechanisms that solve an optimization problem that balances accuracy and tractable proxies for the proposed preference-based notions. Their method is limited to convex boundary-based classifiers. They experiment with classifiers based on logistic regression and show that their preference-based fairness can allow for higher accuracy than existing parity-based fairness notions; however, they acknowledge that parity-based notions are sometimes more desirable. Similarly, Ustun et al. [199] focus on a preference-based group fairness notion. They argue that deviation from group fairness notions is acceptable only if it is based on the best interest of each group. They propose a pre-process mechanism that leverages the sensitive attributes and a recursive feature selection procedure for building decoupled

classifiers that aim to maximize accuracy while satisfying constraints for preference guarantees, such as envy-freeness. Their preferences guarantees are based on requiring that the majority of individuals in each group prefer their assigned classifier (in terms of low generalization error) to other groups' classifiers and to a model that is trained without sensitive attributes. Kim et al. [124] focus on a preference-based individual fairness notion. They propose a notion of fairness referred to as **Preference Informed Individual Fairness (PIIF)**. The new notion suggests that allowing for deviations from individual fairness is beneficial when individuals have diverse preferences over the possible outcomes. The PIIF notion is a relaxation of both individual fairness and envy-freeness (envy-freeness requires that no individual prefers another individual's outcome over their own). They introduce an in-process mechanism that performs optimization to classifiers with convex objectives, subject to PIIF.

Combining Individual and Group Fairness. Speicher et al. [190] and Lohia et al. [138] propose a combined approach to both individual and group fairness. Speicher et al. [190] note that demanding between-group fairness can harm within-group fairness, and therefore propose a new measure for combining both individual and group fairness. Their proposed measure is based on a Generalized Entropy Index (GEI), and it leverages notions from economics theory to assess the inequality of benefits of algorithm results. The measure considers differences in prediction outcomes of individuals and the average prediction accuracy and considers both between-group and within-group components. Lohia et al. [138] proposed a post-processing mechanism that aims to balance accuracy with both individual and group fairness. They start by detecting samples that are prone to individual bias, by finding those that experience a different prediction outcome when the sensitive attribute changes (leaving all other features constant). These individuals are then considered for alternation of their prediction outcome while considering a group-fairness measure (disparate impact). Note that this method does not require ground truth class labels for the validation set.

Pairwise Fairness. Lahoti et al. [129] and Jung et al. [108] introduce *pairwise fairness*. Lahoti et al. [129] discuss the challenge in defining the similarity metric for individual fairness assessment and propose a pre-process mechanism that tackles this challenge. They propose to use external knowledge on human judgments for pairs of individuals who should be treated similarly and maintain this information in a fairness graph structure, which represents a set of pairwise constraints. This information is then used for a mathematical optimization model that is aimed at creating pairwise fair representations (PFR) of the data combined with the graph embedding. The mechanism requires pairwise judgments only for the training phase, and they are not needed for the test data. They recognize that although this information may be valuable, it is also subjective and noisy, and they perform sensitivity analysis to assess the deviations. They show that their mechanism has the potential to improve both individual fairness and group fairness with a low loss in utility. Jung et al. [108] similarly propose to leverage external human judgments regarding pairwise information on equally deserving individuals, which should be treated similarly for a given task. They propose an in-process optimization mechanism that maximizes accuracy subject to the pairwise fairness constraints, which are derived from the external judgments information.

B.4 A Demonstration of Selected Fairness Enhancing Mechanisms

In this appendix, we describe three mechanisms in more detail, one of each mechanism type. Through the description of specific details, we provide a further understanding of the most important principles that similar mechanisms are also based on.

A Pre-Process Fairness Mechanism. First, we describe the approach suggested by Feldman et al. [74]. This method suggests to pre-process the data to decrease the earth mover's distance between feature distributions of both groups (privileged and unprivileged). More precisely, the procedure will move the distributions closer to a distribution referred to as "the median distribution." The intuition is that the classifier should not make decisions that depend on the group through proxy attributes.

The repair mechanism processes each feature individually, ranks the values of each feature, and determines the number of bins based on the smaller population of the two groups. Then, the larger population is divided into the same number of bins and each item is moved toward the median of the associated distribution. Equation (20) defines a geometric repair $\tilde{F}_s^{-1}(\alpha)$:

$$\tilde{F}_s^{-1}(\alpha) = (1-\lambda)\cdot F_s^{-1}(\alpha) + \lambda \cdot F^{-1}_{MedianDistribution}(\alpha), \tag{20}$$

where $F_s(x)$ is the CDF for the population of $S = s$ (meaning $P(X \geq x|S = s)$) and $F_s^{-1}(\alpha)$ is the value of x, a non-sensitive attribute, so that $P(X \geq x|S = s) = \alpha$. The λ parameter is used to repair distributions to be closer to or farther from each other and therefore can serve as a tuning parameter to balance the trade-off between fairness and accuracy.

Note that this approach does not require nor considers labeled data at all. This characteristic can act as an advantage if there is little available information about the labels of one of the groups. However, in many scenarios, it might be beneficial to learn fair representations that additionally take into account the actual outcome (e.g., see [218]). As other pre-process mechanisms, this approach is not tailored to a specific ML algorithm and therefore can be used with any classification algorithm.

Several other pre-process mechanisms follow similar principles of modifying the feature representations so that the distributions for both privileged and unprivileged groups become similar, therefore making it more difficult for the algorithm to differentiate between the two groups (e.g., [38, 74, 139, 180, 218]). Except for the way they operate, these mechanisms sometimes differ in the measures used for evaluation (see Table 2).

An In-Process Fairness Mechanism. We now describe the approach suggested by Kamishima et al. [114]. Their method is a generalization of logistic regression that takes fairness into account by adding a regularization term to the objective function. The added regularization term penalizes mutual information between the sensitive attribute and the model's predictions.

Their method suggests to balance the requirement of prediction accuracy by demanding higher log likelihood, and fairness by requiring lower mutual information between the sensitive attribute and the predictions, as follows:

$$Minimize(-LL^{preds,actual} + \eta \cdot MI^{preds,sens} + \theta \cdot L_2), \tag{21}$$

where $LL^{preds,actual}$ is the log likelihood, calculated by taking into account the predictions and actual outcomes (Equation (22)), $MI^{preds,sens}$ is the mutual information between the sensitive attribute and the predictions (Equation (23)), L_2 is the norm of the weights (Equation (24)), and η and θ are regularization parameters. The η parameter can serve as a tuning parameter to balance the trade-off between fairness and accuracy.

In particular, the calculation of $LL^{preds,actual}$ is given by

$$LL^{preds,actual} = \sum_i \left[y_i log\left(\sigma\left(x_i^T w\right)\right) + (1-y_i)\, log\left(1-\sigma\left(x_i^T w\right)\right)\right], \tag{22}$$

where y_i are the actual outcomes of the privileged group, $\sigma(\cdot)$ is a sigmoid function, and w are the weight vectors for x—the non-sensitive feature vectors.

The calculation of $MI^{preds,sens}$ is given by

$$MI^{preds,sens} = \sum \left(P(\hat{y}, s) log(\frac{P(\hat{y}, s)}{P(\hat{y})P(s)}) \right), \tag{23}$$

where $\hat{y}$ are the predictions, s are the sensitive attribute values, $P(\hat{y}, s)$ is the joint probability, and $P(\hat{y})$ and $P(s)$ are the marginal probabilities.

The calculation of L_2 is given by

$$L_2 = \|w\|_2^2 . \tag{24}$$

The optimization problem is solved by the computational steps and *conjugate gradient decent* method as described by Kamishima et al. [114].

There are other in-process mechanisms that are based on similar concepts of penalizing prediction results that "leak" information about the sensitive attribute. One of the most common approaches is through adding regularization terms to the objective function of the optimization problem. The intuition is that if the sensitive attribute cannot be learned from the prediction output, it means that the prediction is not biased.

These mechanisms differ from each other in the measures that are used for the optimization problem, the measures used for evaluation, and the classification algorithm that the mechanism is based on. As a result, the mechanisms also differ in the techniques used for solving them, depending on linearity and computational complexity of the optimization problem (e.g., see the works of Bechavod and Ligett [16, 17], which use a regularization penalty term that aims to minimize a proxy for equalized odds).

We refer the reader to Table 2 for further details on other in-process methods. The table additionally reviews the different measures used by each mechanism and differentiates between the measures used for optimization purposes versus the measures used for evaluation purposes.

A Post-Process Fairness Mechanism. Hardt el al. [94] propose a technique to adjust any classifier in a post-process mechanism. Their approach is based on first training a classifier without any fairness constraints, then deriving an equalized odds predictor in a second step.

After obtaining the output of a classifier, they suggest to solve an optimization problem for selecting separate thresholds for each group separately, in a manner that minimizes the loss or cost of wrong classifications (i.e., maximizing prediction accuracy) and satisfies constraints for equalized odds or equalized opportunity.

Their method assumes available information about the output of the classifier, the actual outcome, and the membership of individuals in the privileged or unprivileged group. The information about the features and the trained model itself are not required. The optimization problem is formalized as follows.

Formulation 1. *Hardt et al. [94] optimization problem:*

$$\min \mathbb{E}\ell(\tilde{Y}, Y). \tag{25}$$

Subject to the constraints:

$$\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}), \tag{26}$$

where ℓ is the loss function that indicates the cost of predictions, and $\tilde{Y}$ is the derived predictor that minimizes the expected loss and satisfies the set of two constraints, which requires equalized odds (see Section 3). $\gamma_s(\tilde{Y})$ are the FPRs and the TPRs for each of the group $S = s$.

The authors show that if the prediction output of the classifier is binary (i.e., $\hat{Y} \in \{0, 1\}$), then the optimization problem is linear. Otherwise, if the output is a real-valued score R, a binary classifier

can be obtained by thresholding the score. They further show that finding optimal thresholds for minimizing loss while satisfying equalized odds is not a linear problem but can be efficiently optimized numerically using ternary search. They additionally propose to use a randomization technique for cases in which intersection points are difficult to find.

As shown in the work of Hardt et al. [94], this method can also be interpreted as finding the optimal point that minimizes the loss on the pointwise minimum of the two ROC curves (one curve for each of the two groups).

Other mechanisms (e.g., [51, 62, 149]) are based on similar principles based on post hoc corrections of classification thresholds. These mechanisms differ from each other in the measures and the formulation of the optimization problem, as well as in the techniques used for solving them.

B.5 Paper Selection Procedure

The selection process started by using Google Scholar with several main queries: algorithmic bias, algorithmic fairness, fairness-aware ML, fairness in ML, fairness measures, fairness mechanisms, algorithmic bias mitigation, and fairness improvement. Queries were searched mainly between February and April 2021 and between November 2019 and January 2020. We included papers from both journals and conferences and focused on papers that were published in recent years. This initial set of papers was then expanded by adding papers that cited or were cited by papers in this set. We then analyzed these papers and categorized them into several groups to suit the intended structure of the review as follows: definitions and measures, datasets, mechanisms, trade-offs, and topics beyond classification.

For the review of emerging fairness topics beyond classification tasks, we used an extended set of queries: fair sequential learning, fair adversarial learning, fair word embedding, fair visual description, fair recommender systems, multi-sided fairness, fair causal learning, and fair private learning. It should be noted that we did not aim at covering these emerging topics to their full depth, and where possible we referred the readers to other relevant surveys.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2019. Beyond personalization: Research directions in multistakeholder recommendation. *arXiv preprint arXiv:1905.01986* (2019).

[2] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158.

[3] Adel Abusitta, Esma Aïmeur, and Omar Abdel Wahab. 2019. Generative adversarial networks for mitigating biases in machine learning systems. *arXiv preprint arXiv:1905.09972* (2019).

[4] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the International Conference on Machine Learning*. 60–69.

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

[6] Julia Angwin. 2016. Machine Bias: ProPublica. Retrieved June 5, 2019 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. 2017. Face aging with conditional generative adversarial networks. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP'17)*. IEEE, Los Alamitos, CA, 2089–2093.

[8] Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, Viggo Kann, Alberto Marchetti-Spaccamela, and Marco Protasi. 2012. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer Science & Business Media.

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

[10] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*. 15479–15488.

[11] Michiel A. Bakker, Humberto Riverón Valdés, Duy Patrick Tu, Krishna P. Gummadi, Kush R. Varshney, Adrian Weller, and Alex Pentland. 2020. Fair enough: Improving fairness in budget-constrained decision making using confidence thresholds. In *Proceedings of the Workshop on Artificial Intelligence Safety, Co-Located with the 34th AAAI Conference on Artificial Intelligence (SafeAI@AAAI'20)*.

[12] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: Fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*. 2745–2754.

[13] Elias Bareinboim and Judea Pearl. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113 (2016), 7345–7352.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and machine learning. fairmlbook.org.

[15] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[16] Yahav Bechavod and Katrina Ligett. 2017. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044* (2017).

[17] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116 (2019), 15849–15854.

[19] 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).

[20] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 6 (2018), 0049124118782533.

[21] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2212–2220.

[22] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. *arXiv preprint arXiv:1901.04562* (2019).

[23] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[24] Peter J. Bickel, Eugene A. Hammel, and J. William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187 (1975), 398–404.

[25] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. 405–414.

[26] 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.

[27] Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 7–15.

[28] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *Proceedings of the International Conference on Machine Learning*. 715–724.

[29] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391* (2017).

[30] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. 2018. Multiwinner elections with diversity constraints. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[31] Phil Brierley, David Vogel, and Randy Axelrod. 2011. Heritage Provider Network Health Prize Round 1 Milestone Prize: How we did iTeam "Market Makers." Retrieved on 16 Dec., 2021 from https://foreverdata.org/1015/content/milestone1-2.pdf.

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

[33] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 77–91.

[34] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[35] Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. 2017. Balanced neighborhoods for fairness-aware collaborative recommendation. In *Proceedings of the ACM FATRec Workshop*.

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

[37] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR abs/1608.07187* (2016).
[38] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.
[39] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. 2018. Multiwinner voting with fairness constraints. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 144–151.
[40] L. Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443* (2019).
[41] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with fairness constraints. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming (ICALP'18)*.
[42] Minnesota Population Center. 2015. Integrated Public Use Microdata Series, International: Version 6.4 [The Dutch Virtual Census of 2001]. Retrieved November 10, 2019 from https://doi.org/10.18128/D020.V6.4
[43] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.
[44] Silvia Chiappa and William S. Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.
[45] Silvia Chiappa and William S. Isaac. 2019. A causal Bayesian networks viewpoint on fairness. *arXiv preprint arXiv:1907.06430* (2019).
[46] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*. 5029–5037.
[47] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5 (2017), 153–163.
[48] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).
[49] UCI Machine Learning Repository: Communities and Crime Data Set. 2009. Retrieved October 7, 2021 from http://archive.ics.uci.edu/ml/datasets/communities+and+crime.
[50] 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).
[51] 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*. ACM, New York, NY, 797–806.
[52] 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.
[53] Tore Dalenius. 1977. Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 15 (1977), 429–444.
[54] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved June 6, 2019 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
[55] Heritage Health Prize Contest Data. 2013. Collection 1015. Retrieved January 29, 2021 from https://foreverdata.org/1015/index.html.
[56] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015 (2015), 92–112.
[57] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 248–255.
[58] Josep Domingo-Ferrer and Alberto Blanco-Justicia. 2020. Privacy-preserving technologies. In *The Ethics of Cybersecurity*. Springer, Cham, Switzerland, 279–297.
[59] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. Retrieved October 7, 2021 from http://archive.ics.uci.edu/ml.
[60] 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*. ACM, New York, NY, 214–226.
[61] Cynthia Dwork and Christina Ilvento. 2018. Fairness under composition. In *Proceedings of the 10th Innovations in Theoretical Computer Science Conference (ITCS'19)*.
[62] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 119–133.

[63] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference*. 265–284.
[64] Cynthia Dwork and Moni Naor. 2010. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality* 2 (2010), 1–12.
[65] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2019. FaiRecSys: Mitigating algorithmic bias in recommender systems. *International Journal of Data Science and Analytics* 9 (2019), 197–213.
[66] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).
[67] Michael D. Ekstrand and Maria Soledad Pera. 2017. The demographics of cool. In *Poster Proceedings at ACM RecSys*. ACM, New York, NY. .
[68] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, 242–250.
[69] Shady Elbassuoni, Sihem Amer-Yahia, Ahmad Ghizzawi, and Christine Atie. 2019. Exploring fairness of ranking in online job marketplaces. In *Proceedings of the 22nd International Conference on Extending Database Technology (EDBT'19)*.
[70] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. 2019. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 170–179.
[71] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna Gummadi, and Patrick Loiseau. 2019. The price of local fairness in multistage selection. *arXiv preprint arXiv:1906.06613* (2019).
[72] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
[73] James Fantin. 2020. A Distributed Fair Random Forest. University of Wyoming. Retrieved on 16 Dec., 2021 https://mountainscholar.org/bitstream/handle/20.500.11919/7072/STUW_HT_COSC_2020_Fantin_James?sequence=1.
[74] 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*. ACM, New York, NY, 259–268.
[75] Joel Escudé Font and Marta R. Costa-Jussa. 2019. Equalizing gender biases in neural machine translation with word embeddings techniques. *arXiv preprint arXiv:1901.03116* (2019).
[76] Yoav Freund and Robert E. Schapire. 1996. Game theory, on-line prediction and boosting. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*. 325–332.
[77] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).
[78] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2018. Comparing fairness-aware machine learning techniques. Retrieved October 7, 2021 from https://github.com/algofairness/fairness-comparison/tree/master/fairness/data.
[79] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 329–338.
[80] Robert Fullinwider. 2018. Affirmative action. In *The Stanford Encyclopedia of Philosophy* (summer 2018 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.
[81] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. 498–510.
[82] Arpita Ghosh and Robert Kleinberg. 2016. Inferential privacy guarantees for differentially private mechanisms. *arXiv preprint arXiv:1603.01508* (2016).
[83] Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 269–278.
[84] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P. Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.
[85] 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. 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)*. 609–614.
[86] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[87] Sruthi Gorantla, Amit Deshpande, and Anand Louis. 2020. Ranking for individual and group fairness simultaneously. *arXiv preprint arXiv:2010.06986* (2020).
[88] Ben Green and Lily Hu. 2018. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Proceedings of Machine Learning: The Debates Workshop at the 35th International Conference on Machine Learning (ICML'18)*.
[89] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13 (2012), 723–773.
[90] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.
[91] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[92] Michael Gusenbauer. 2019. Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118, 1 (Jan. 2019), 177–214. https://doi.org/10.1007/s11192-018-2958-5
[93] Evan Hamilton. 2017. *Benchmarking Four Approaches to Fairness-Aware Machine Learning*. Ph.D. Dissertation. Haverford College.
[94] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.
[95] Hoda Heidari and Andreas Krause. 2018. Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. 2248–2254.
[96] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*. 793–811.
[97] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 600.
[98] Hui Hu, Yijun Liu, Zhen Wang, and Chao Lan. 2019. A distributed fair machine learning framework with private demographic data protection. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM'19)*. IEEE, Los Alamitos, CA, 1102–1107.
[99] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*. 1389–1398.
[100] Chong Huang, Xiao Chen, Peter Kairouz, Lalitha Sankar, and Ram Rajagopal. 2018. Generative adversarial models for learning private and fair representations. Retrieved on 16 Dec., 2021 https://openreview.net/pdf?id=H1xAH2RqK7.
[101] Elle Hunt. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Retrieved October 19, 2019 from https://goo.gl/mE8p3J.
[102] Pablo Ibarrarán, Nadin Medellín, Ferdinando Regalia, Marco Stampini, Sandro Parodi, Luis Tejerina, Pedro Cueva, et al. 2017. *How Conditional Cash Tansfers Work*. IDB Books.
[103] Northpointe Inc. 2012. Practitioners Guide to COMPAS. Retrieved November 18, 2021 from http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf.
[104] Rishabh K. Iyer and Jeff A. Bilmes. 2013. Submodular optimization with submodular cover and submodular knapsack constraints. In *Proceedings of Neural Information Processing Systems*.
[105] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70*. 1617–1626.
[106] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. 2019. Differentially private fair learning. In *Proceedings of the International Conference on Machine Learning*. 3000–3008.
[107] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. 325–333.
[108] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660* (2019).
[109] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2019. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285* (2019).
[110] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *Proceedings of the 2009 2nd International Conference on Computer, Control, and Communication*. IEEE, Los Alamitos, CA, 1–6.
[111] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2012), 1–33.

[112] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Proceedings of the 2010 IEEE International Conference on Data Mining*. IEEE, Los Alamitos, CA, 869–874.
[113] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Enhancement of the neutrality in recommendation. In *Proceedings of the Workshop on Human Decision Making in Conjunction with the 6th ACM Conference on Recommender Systems (Decisions@RecSys'12)*. 8–14.
[114] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 35–50.
[115] Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. 2018. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems*. 2227–2236.
[116] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913* (2019).
[117] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, 3819–3828.
[118] Ehsan Kazemi, Morteza Zadimoghaddam, and Amin Karbasi. 2018. Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. In *Proceedings of the International Conference on Machine Learning*. 2549–2558.
[119] Madian Khabsa and C. Lee Giles. 2014. The number of scholarly documents on the public web. *PloS One* 9, 5 (2014), e93949.
[120] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.
[121] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, Vol. 80. 2635–2644.
[122] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 277–287.
[123] DaeEun Kim. 2006. Minimizing structural risk on decision tree classification. In *Multi-Objective Machine Learning*. Springer, 241–260.
[124] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. 2019. Preference-informed fairness. *arXiv preprint arXiv:1904.01793* (2019).
[125] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS'17)*.
[126] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *Proceedings of the World Wide Web Conference*. 2936–2942.
[127] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems—A survey. *Knowledge-Based Systems* 123 (2017), 154–162.
[128] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 4069–4079.
[129] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439* (2019).
[130] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. Retrieved August 28, 2019 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
[131] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. ProPublica Compas Analysis—Data and Analysis for 'Machine Bias.'Retrieved October 7, 2021 from https://github.com/propublica/compas-analysis.
[132] Eric L. Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. 2014. Fairness-aware loan recommendation for microfinance services. In *Proceedings of the 2014 International Conference on Social Computing*. ACM, New York, NY, 3.
[133] Erich L. Lehmann and Joseph P. Romano. 2006. *Testing Statistical Hypotheses*. Springer Science & Business Media.
[134] Yehuda Lindell and Benny Pinkas. 2000. Privacy preserving data mining. In *Proceedings of the Annual International Cryptology Conference*. 36–54.
[135] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. 2017. Does mitigating ML's disparate impact require disparate treatment? *Stat* 1050 (2017), 19.

[136] Weiwen Liu and Robin Burke. 2018. Personalizing fairness-aware re-ranking. *arXiv preprint arXiv:1809.02921* (2018).
[137] Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).
[138] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2019. Bias mitigation post-processing for individual and group fairness. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'19)*. IEEE, Los Alamitos, CA, 2847–2851.
[139] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2016. The variational fair autoencoder. In *Proceedings of the International Conference on Learning Representations (ICLR'16)*.
[140] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 502–510.
[141] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* (2018).
[142] David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict responsibly: Increasing fairness by learning to defer. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*.
[143] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553* (2020).
[144] Rowland Manthorpe. 2017. Beauty.AI's 'robot beauty contest' is back—And this time it promises not to be racist. *Wired*. Retrieved November 12, 2019 from https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai.
[145] Fernando Martínez-Plumed, Cèsar Ferri, David Nieves, and José Hernández-Orallo. 2019. Fairness and missing values. *arXiv preprint arXiv:1905.12728* (2019).
[146] Chandler May, Alex Wang, Shikha Bordia, Samuel Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*. 622–628.
[147] 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).
[148] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2243–2251.
[149] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 107–118.
[150] Weiwen Miao. 2010. Did the results of promotion exams have a disparate impact on minorities? Using statistical evidence in *Ricci v. DeStefano*. *Journal of Statistics Education* 18, 3 (2010), 1–26.
[151] Tom M. Mitchell. 1980. *The Need for Biases in Learning Generalizations*. Technical Report CBM-TR-117. Department of Computer Science, Laboratory for Computer Science Research, Rutgers University.
[152] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31.
[153] Hussein Mozannar, Mesrob I. Ohannessian, and Nathan Srebro. 2020. Fair learning with private demographic data. *arXiv preprint arXiv:2002.11651* (2020).
[154] Sendhil Mullainathan. 2019. Biased Algorithms Are Easier to Fix Than Biased People. Retrieved December 26, 2019 from https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html?smid=nytcore-ios-share.
[155] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
[156] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.
[157] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, 77–83.
[158] NYPD. 2017. Stop, Question and Frisk (SQF) Data. Retrieved February 14, 2021 from https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page.
[159] UCI Machine Learning Repository. 2016. Default of Credit Card Clients Data Set. Retrieved October 7, 2021 from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.
[160] Pete Pachal. 2015. Google Photos identified two black people as 'gorillas.' *Mashable*. Retrieved November 9, 2019 from https://mashable.com/2015/07/01/google-photos-black-people-gorillas/.
[161] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (2009), 1345–1359.

[162] Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the Web Conference*. 1194–1204.

[163] Gourab K. Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna Gummadi. 2020. Incremental fairness in two-sided market platforms: On smoothly updating recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 181–188.

[164] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

[165] Dana Pessach and Erez Shmueli. 2021. Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings. *Expert Systems with Applications* 185 (2021), 115667.

[166] Dana Pessach, Tamir Tassa, and Erez Shmueli. 2021. Fairness-driven private collaborative machine learning. *arXiv preprint arXiv:2109.14376* (2021).

[167] Lizzie Plaugic. 2017. FaceApp's creator apologizes for the app's skin-lightening 'hot' filter. *The Verge*. Retrieved November 12, 2019 from https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology.

[168] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[169] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems*. 677–688.

[170] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8227–8236.

[171] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141 (2002), 660–678.

[172] Derek Roth. 2018. *A Comparison of Fairness-Aware Machine Learning Algorithms*. Ph.D. Dissertation. Haverford College.

[173] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 8–14.

[174] Chris Russell, Matt J. Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.

[175] George Rutherglen. 1987. Disparate impact under title VII: An objective theory of discrimination. *Virginia Law Review* 73 (1987), 1297.

[176] George Rutherglen. 2009. Ricci v DeStefano: Affirmative action and the lessons of adversity. *Supreme Court Review* 2009 (2009), 83–114.

[177] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. 2017. InclusiveFaceNet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193* (2017).

[178] Babak Salimi, Bill Howe, and Dan Suciu. 2019. Data management for causal algorithmic fairness. *IEEE Data Engineering Bulletin* 2019 (2019), 24.

[179] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD'19)*.

[180] Samira Samadi, Uthaipon Tantipongpipat, Jamie H. Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*. 10976–10987.

[181] Richard H. Sander. 2004. A systemic analysis of affirmative action in American law schools. *Stanford Law Review* 57 (2004), 367.

[182] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush Raj Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development* 63, 4–5 (2019), Article 3, 9 pages.

[183] UCI Machine Learning Repository. 1996. Adult Data Set. Retrieved October 7, 2021 from https://archive.ics.uci.edu/ml/datasets/adult.

[184] UCI Machine Learning Repository. 2012. Bank Marketing Data Set. Retrieved October 7, 2021 from https://archive.ics.uci.edu/ml/datasets/bank+marketing.

[185] UCI Machine Learning Repository. n.d. Diabetes Data Set. Retrieved October 7, 2021 from https://archive.ics.uci.edu/ml/datasets/diabetes.

[186] UCI Machine Learning Repository. 1994. Statlog (German Credit Data) Data Set. Retrieved October 7, 2021 from https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

[187] Tom Simonite. 2015. Probing the Dark Side of Google's Ad-Targeting System. Retrieved July 31, 2019 from https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/.

[188] Tom Simonite. 2018. When It Comes to Gorillas, Google Photos Remains Blind. Retrieved September 17, 2019 from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.
[189] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056* (2019).
[190] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2239–2248.
[191] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. 499–506.
[192] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. *arXiv preprint arXiv:1902.04783* (2019).
[193] Pierre Stock and Moustapha Cisse. 2017. ConvNets and ImageNet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv preprint arXiv:1711.11443* (2017).
[194] Tom Sühr, Asia J. Biega, Meike Zehlike, Krishna P. Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 3082–3092.
[195] Özge Sürer, Robin Burke, and Edward C. Malthouse. 2018. Multistakeholder recommendation with provider constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, New York, NY, 54–62.
[196] National Lung Screening Trial Research Team. 2011. The National Lung Screening Trial: Overview and study design. *Radiology* 258 (2011), 243–253.
[197] Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communications, Control, and Computing*. 368–377.
[198] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *Proceedings of the 2017 IEEE European Symposium on Security and Privacy (EuroS&P'17)*. IEEE, Los Alamitos, CA, 401–416.
[199] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the International Conference on Machine Learning*. 6373–6382.
[200] Berk Ustun, M. Brandon Westover, Cynthia Rudin, and Matt T. Bianchi. 2016. Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine* 12 (2016), 161–168.
[201] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. 2018. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*. 1769–1778.
[202] Emiel Van Miltenburg. 2016. Stereotyping and bias in the Flickr30k dataset. In *Proceedings of the Workshop on Multimodal Corpora (MMC'16)*. 1–4.
[203] Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2019. Getting gender right in neural machine translation. *arXiv preprint arXiv:1909.05088* (2019).
[204] Vladimir Vapnik and Rauf Izmailov. 2015. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research* 16 (2015), 2023–2049.
[205] Vladimir N. Vapnik. 2000. *Controlling the Generalization Ability of Learning Processes*. Springer New York, New York, NY, 93–122. https://doi.org/10.1007/978-1-4757-3264-1_5
[206] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (FairWare'18)*. IEEE, Los Alamitos, CA, 1–7.
[207] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).
[208] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: A trade-off revisited. In *Advances in Neural Information Processing Systems 32 (NeurIPS'19)*.
[209] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Proceedings of the Conference on Learning Theory*. 1920–1953.
[210] Depeng Xu, Shuhan Yuan, and Xintao Wu. 2019. Achieving differential privacy and fairness in logistic regression. In *Companion Proceedings of the 2019 World Wide Web Conference*. 594–599.
[211] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware generative adversarial networks. In *Proceedings of the 2018 IEEE International Conference on Big Data (Big Data'18)*. IEEE, Los Alamitos, CA, 570–575.
[212] Sirui Yao and Bert Huang. 2017. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838* (2017).
[213] I-Cheng Yeh and Che-Hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications* 36 (2009), 2473–2480.

[214] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[215] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. 962–970.

[216] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From parity to preference-based notions of fairness in classification. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'17)*.

[217] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*. 1569–1578.

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

[219] 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*. ACM, New York, NY, 335–340.

[220] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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)*. 629–634.

[221] 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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

[222] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 15–20.

[223] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4847–4853.

[224] Michael J. Zimmer. 1995. Emerging uniform structure of disparate treatment discrimination litigation. *Georgia Law Review* 30 (1995), 563.

[225] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31 (2017), 1060–1089.

Received January 2020; revised October 2021; accepted October 2021