

# Fairness-Aware Machine Learning

## An Extensive Overview

Jannik Dunkelau<sup>(✉)</sup> and Michael Leuschel

Heinrich-Heine-Universität Düsseldorf  
Universitätsstraße 1 · 40225 Düsseldorf  
jannik.dunkelau@hhu.de michael.leuschel@hhu.de

**Abstract.** We provide an overview of the state-of-the-art in fairness-aware machine learning and examine a wide variety of research articles in the area. We survey different fairness notions, algorithms for pre-, in-, and post-processing of the data and models, and provide an overview of available frameworks.

**Keywords:** Fairness · Algorithmic Fairness · Machine Learning · Survey

### Table of Contents

1	Introduction	2
1.1	Legal Motivation	3
1.2	Types and Causes of Discrimination	4
1.3	Causes for Machine Bias	5
2	Related Work	7
3	Necessary Nomenclature	9
3.1	Machine Learning Nomenclature	9
3.2	Fairness Nomenclature	10
4	Mathematical Notation	12
5	Notions of Fairness	13
5.1	Unawareness	14
5.2	Group fairness	14
5.3	Predictive Parity	15
5.4	Calibration	19
5.5	Individual Fairness	20
5.6	Preference-Based Fairness	21
5.7	Causality	21
5.8	Note on the Selection of Fairness Metrics	23
6	Pre-Processing: Discrimination-free Training Data	24
6.1	Relabelling	24
6.2	Resampling	25
6.3	Fair Representations	26
6.4	Further Algorithms	27
7	In-Processing: Discrimination-aware Learners	27

7.1	Adjusted Learning Algorithms	28
7.2	Adapted Loss Function	29
7.3	Adversarial Approaches	31
7.4	Optimisation Subject to Fairness Constraints	32
7.5	Compositional Approaches	34
7.6	Further Algorithms	35
8	Post-Processing: Correcting Biased Classifiers	35
8.1	Output correction	35
8.2	Input correction	38
8.3	Classifier correction	39
8.4	Further Algorithms	40
9	Fairness Toolkits	40
9.1	Datasets	41
9.2	Frameworks	41
10	Further Discussion and Final Remarks	44
10.1	Accountability and Transparency	44
10.2	Ethics and the Use of Protected Attributes	45
10.3	Critiques and Perspectives on Current Practices	47

## 1 Introduction

In today’s world, artificial intelligence (AI) increasingly surrounds us in our day-to-day lives. This is especially true for machine learning algorithms, which learn their behaviours by recognising patterns in existent data and apply it to new instances to make correct precisions quickly. This is desirable as it reduces the factor of human error and speeds up various processes, taking less than a second for a decision which would take a human worker multiple minutes.

For instance, a company can reliably speed up its hiring process by algorithmically filtering through hundreds of applications, leaving a more manageable amount for human review. The recidivism risk-scores for criminals can also be computationally determined, reducing human error in this regard, leading to a more reliant scoring system altogether. Another example might be the admission of students into universities, favouring those who have a higher chance of graduating instead of dropping out. However, besides not using any sensitive attribute like race, sex, age, or religion as input the algorithms might still learn how to discriminate against them. This gives way for new legal implications, as well as ethical problems.

The fairness-aware machine learning community only began to develop in the last ten years, with the first publication to the best of our knowledge leading back to Pedreschi et al. in 2008 [151]. Since then there is a steady grow of interest, giving way to a multitude of different fairness notions, as well as algorithms for preventing machine-learned bias in the first place.

In this survey paper, we will compile the current stand of research regarding fairness-aware machine learning. This includes definitions of different fairness notions and algorithms, as well as discussion of problems and sourced of machine

discrimination. By bundling the information from different sources, this paper serves as a rich entry-point for researchers new to the area, as well as an extensive fairness bibliography, spanning also legal references and examples of employed machine learning systems.

The remainder of this paper is structured as follows: the rest of this section motivates fairness-aware algorithms on legal grounds, discusses various causes of unfairness, and the resulting implications of discrimination in Sections 1.1 to 1.3 respectively. Section 2 goes over related work, i.e. other survey papers considering different parts of the whole research area. Section 3 establishes a common ground for nomenclature used throughout the paper with Section 4 introducing the necessary mathematical notation. Sections 5 to 8 list various definitions of algorithmic fairness, as well as pre-, in-, and post-processing algorithms found in an extensive literature review. In Section 9, different frameworks, toolkits, as well as common databases used in literature are presented. The paper concludes with some final remarks in Section 10.

## 1.1 Legal Motivation

Unequal opportunities are known of in employment rates [17, 121] or the mortgage market [17, 124]. As a countermeasure, various legislations are in place to ensure the non-discrimination of minority groups. In the U.K., the Equality Act [188] is in place since October 2010, consolidating the previous Sex Discrimination Act [185], the Race Relations Act [186], and the Disability Discrimination Act [187]. In U.S.A., these legislations are regulated by the Civil Rights Act of 1968 [190], the Equal Pay Act of 1963 [189], and the Equal Credit Opportunity Act of 1974 [191]. The European Union passed the Racial Equality Directive 2000 [66], the Employment Equality Framework Directive 2000 [67], and the Equal Treatment Directive 2006 [68]. A further EU proposed directive for implementing the Equal Treatment Directive was proposed and partially agreed on in 2009 [64], yet is still pending.<sup>1</sup>

Intuitively, employing AI systems driving the decision process in such cases would give the benefit of being objective and hence free of any discrimination. Unfortunately, this is not the case. Google Ads show less high-paying job offers for females [52] while searching for black-identifying names results in ads suggestive of an arrest [173]. Amazon provided same-day-delivery offers to certain neighbourhoods, chosen by an algorithm which ultimately reinforced racial bias and never offered same-day-delivery for neighbourhoods consisting mainly of minority groups [105]. Commercial, image-based gender classifiers by Microsoft, IBM, and Face++ all increasingly mispredict once the input individual is dark-skinned or female, with an error rate for dark-skinned females (20.8%–34.7%) significantly worse than those for light-skinned males (0.0%–0.3%) [36].

Further examples include gender bias in word embeddings [32], discrimination in assigning credit scores [48], bias in image search results [116], racial bias

---

<sup>1</sup> As of October 2019.

in recidivism prediction [7], or prejudice in New York City’s stop-and-frisk policy [81, 82].

Note that these systems did not explicitly discriminate, e.g. the Amazon system did not explicitly have access to race in its decision making process. Although ignoring sensitive attributes like race, sex, age, or religion intuitively should be sufficient for a fair classification, it was shown that such systems make use of indirect discrimination [15, 40, 151, 162]. The discriminatory attribute is deduced from seemingly unrelated data, which is also known as disparate impact [15]. Further, the algorithms are trained on historical data, which can contain previous discrimination which is hence learned by the model [15, 40]

The subject of employing machine learning algorithms in such cases henceforth introduces new legislative problems, as well as ethical problems. The Leadership Conference on Civil and Human Rights published a set of five principles to respect the value of equal opportunity and justice [125], following up with a report on the current stand of social justice and technology [159]. A White House Report to the Obama administration points toward the discrimination potential of such techniques [177, 178] with a consecutive report calling for *equal opportunity by design* [179]. The European Union released a governance framework for algorithmic transparency and accountability [65], giving policies for accountable and transparent algorithms, as well as an in-depth overview on the subject.

**Success Stories.** Besides all the negative examples of discriminatory systems mentioned above, there do exist examples in which fairness-aware machine learning was successfully employed.

In a study by Danner et al. [51] a risk assessment tool was used at selected pretrial services agencies. These selected agencies were able to release more defendants which were less risky on average compared to those agencies who did not use the tool.

Ahlman and Kurtz [6] put a predictor by Berk et al. [24, 26] into production for randomised control trials of prisoners on probation. By querying the predictor, the burden on parolees could efficiently be reduced while not increasing the re-offence rates significantly.

Gender and race discrimination could further be overcome in face recognition [164] as well as image captioning [99].

## 1.2 Types and Causes of Discrimination

The fairness literature mainly concentrates on two types of discrimination [60, 70, 76, 202], namely *disparate treatment* and *disparate impact*. Feldman et al. [70] and Dwork et al. [60] delivered good surveys on different aspects of each type, which we will briefly summarise in this section.

**Disparate Treatment.** Disparate treatment describes the practise of treating individuals differently based on their sensitive attributes like race, gender, age or religion. It is intuitively the most direct form of discrimination.

An approach to hide disparate treatment is reverse tokenism [60]. Tokenism describes the act of admitting few individuals of minority groups (the tokens) which serve as display for equal treatment. Reverse tokenism is the practice of rejecting applicants of minority groups by having an actually qualified member of the majority group which was also rejected. Thus claims of discrimination against the minority group can be refuted by referencing to a more qualified individual (the token) which was treated equally.

A form of disparate treatment is known as taste-based discrimination [19], which describes an economic theory of labour-market discrimination. Here, an employer actively discriminates against a minority group, either due to his personal tastes or due to his subordinates which would avoid any interaction with a co-worker from a minority group.

**Disparate Impact.** We already outlined the idea of disparate impact as indirect discrimination in Section 1.1. The difference to disparate treatment is that the sensitive attributes are not used in the decision making process but minorities still obtain different outcomes (the impact) from the majority group. This is possible, e.g., due to correlations between the sensitive attributes and other attributes which are not protected by any law.

One practice which corresponds to disparate impact is redlining [104]. Redlining refers initially to declining mortgage loans for residents of specific neighbourhoods, either directly or indirectly [27,62,98]. While one’s neighbourhood is not a legally protected attribute, it is usually correlated with race or social status and thus serves as a proxy for discrimination. Alternatively, reverse redlining describes the practice to actively target minority groups, e.g. for charging them higher interest rates [35].

Predictive policing is the practice of statistically predicting crimes to identify targets of police intervention [133]. However, this can lead to disparate impact as well. Due to a historical bias of police forces focussing attention on certain neighbourhoods [82], more crimes are recorded there simply by increased police presence. As the crime rate there is perceived as statistically higher, more attention is put on those neighbourhoods by predictive policing, again leading to an increased track record [133]. This relates to the notion of the self-fulfilling prophecy [60]. Here, decisions are intentionally based upon building a bad track record for a minority group, i.e. by admitting explicitly unqualified individuals. Due to corresponding poor performance potential prejudice can be amplified and serves as historical justification for future discrimination. This is also known as statistical discrimination [8], which is the second big theory on the cause of labour-market discrimination (contrasting taste-based discrimination).

### 1.3 Causes for Machine Bias

As the examples in Section 1.1 have proved, AI systems are well capable of discrimination despite being possibly perceived as impartial and objective [15, 40, 177]. In this section, we will give a brief overview of how machine bias can be

introduced to such systems. For a more thorough assembly and discussion, we point the reader to the excellent works of Calders and Žliobaitė [40] as well as Barocas and Selbst [15], upon which this section is based.

**Data Collection** The training data can contain an inherent bias which in turn leads to a discriminatory model.<sup>2</sup>

One kind of unfair training data is due to incorrect distribution of the ground truths. These ground truths can either be objective or subjective [40]. Examples for objective ground truths include whether a credit was repaid, or whether a criminal did reoffend, i.e. outcomes which can objectively be determined without any influence of personal opinion. Subjective ground truths on the other hand are dependent on the individual creator and assignment might differ depending on the assigner. Examples include whether an applicant was hired or a student was admitted to university. Note that only subjective ground truths can be incorrect [40]. Objective ground truths however still do not imply a discrimination-free dataset, as shown by the predictive policing practice presented in Section 1.2.

The collection of data over historical records is always at risk of carrying biases. Besides subjectively assigned ground truths the dataset itself can vary in quality. On one hand, the data collected by companies might contain several mistakes [195]. The collection process can easily be biased in the first place, as shown by Turner and Skidmore [184]. On the other hand, minority groups can simply not be well represented in the data due to varying access to digitalisation or otherwise different exposure to the employed data collection technique [126]. This is correlated with the notion of negative legacy [113] where prior discrimination resulted in less gathered data samples of a minority group as respective individuals were automatically denied due to disparate treatment. It is known that machine learning performance on data with sample-bias may lead to inaccurate results [201]. This all can cause statistical biases which the machine learning model eventually reinforces.

**Selecting the Features.** Related to *how* the data is collected is the question of *what* data is collected. This is known as feature selection.

One already discussed problem is disparate impact or redlining, where the outcome of a prediction model is determined over a proxy variable [182], i.e. a variable which is directly dependent on the sensitive attribute. However, incomplete information also introduces problems to the system. That is, not all information necessary for a correct prediction is taken into account as feature and remains unobserved, either due to oversight, lack of domain knowledge, privacy reasons, or simply that the information is difficult to observe to begin with [40]. This lack of correctly selected features might only affect specific subgroups of the population, leading to less accurate predictions only for their individuals while the majority does not experience this issue. However, in certain cases the acquisition of more precise features can come at a rather expensive costs for only a marginal increase in prediction accuracy [129].

<sup>2</sup> This corresponds to the computer science proverb ‘garbage in, garbage out’.

**Wrong Assumptions.** Having outlined possible reasons for discriminatory datasets, such as sample-bias or missing features, there is a set of assumptions usually made regarding these datasets which however are not necessarily true [40].

The first assumption is that the dataset is representative of the whole population. Considering the data collection problems above, specifically the sample-bias, this is easily refuted. A related assumption is that the data is representative of the future samples the resulting classifier will be used upon. Hoadley [101] emphasises that a dataset is only ever a capture of the current population. Due to population drift, future observation might differ (significantly) than those over which the current dataset was conducted [96]. For instance, economic changes could greatly influence consumer behaviour, thus rendering corresponding old datasets inadequate.

**Masking.** Another problem to consider is the question of *who* provided the data. Having a set of unintentional ways to induce bias into a dataset also comes with the same set of intentional practices to achieve quite this.

Intentionally applying the above practices for masking the embedding of prejudice into the dataset leads consequently to biased machine learning predictors [15]. This corresponds to the self-fulfilling prophecy discussed above [60].

As Barocas and Selbst [15] point out, data mining can help decision makers in detecting reliable proxy variables “to distinguish and disadvantage members of protected classes” intentionally. However, it is already difficult to prove most cases of, e.g., employment discrimination, hence “employers motivated by conscious prejudice would have little to gain by pursuing these complex and costly mechanisms”. Barocas and Selbst conclude that unintentional discrimination is the more pressing concern.

## 2 Related Work

This article is by no means the first attempt to conduct a survey over the fair machine learning literature. Other authors already tackled different parts of the field before us and did a great job in their ambitions and serve as a great entry point into the field. In this section we will present their works and place it in contrast to ours.

Romei and Ruggieri [160] conducted a broad multidisciplinary survey on discrimination analysis. Surveying core concepts, problems, methods, and more from the perspectives of legal, statistical, economical and computational grounds, the authors gathered a great resource for different fields and delivered a broad, multidisciplinary bibliography. Focus are the applied approaches for discrimination analysis, which the authors divided into four categories: observational, quasi-experimental, experimental, and knowledge discovery studies. In contrast to our survey, we did not attempt to create a multidisciplinary resource, but rather a resource specifically for fair machine-learning researchers, although our overview of fairness notions in Section 5 makes an attempt to be comprehensible

by a more general audience. Romei and Ruggieri focused mainly on methods to analyse discrimination in a datamining sense, whereas we focus mainly on algorithms for bias prevention.

Gajane and Pechenizkiy [76] aimed to formalise different fairness notions. Dividing fairness into six categories, they give a formal definition from the fairness literature and further unify these with corresponding notions of the social sciences literature. Thereby they provide excellent discussion about the respective social limitations in applicability of the notions onto given problem domains. Our overview of fairness notions in Section 5 is partially influenced by their work, although we provide a greater list of different notions.

Žliobaitė [212] conducted a survey on the measurement of indirect discrimination. He considered 20 different measures, divided into four categories: statistical tests, absolute measures, conditional measures, and structural measures. Hereby he considered tests which were not previously used for discrimination measurements in the fairness literature. After giving a review and analysis of the measures in question the author concludes in recommendations for researchers and practitioners, encouraging the use of normalised difference and discourage the use of ratio based measures due to challenges in their interpretation. This contrasts our work as we do not provide a general guidance framework but an extensive overview of the field itself.

Verma and Rubin [194] conducted also a survey over different fairness notions. Their goal was to contrast the surveyed measurements in their rationales and to show how the same problem setting can be considered either as fair or unfair depending on the chosen notion. For this, the authors trained a linear regression classifier on the German Credit dataset and evaluated whether the classifier satisfied each individual fairness criterion. In total, their survey was conducted over 20 different notions of fairness. Our Section 5 was greatly influenced by their work. However, instead of training a classifier and stating which fairnesses were achieved, we present separate examples to each notion aiming for increased comprehensibility by visualisation. We further expand their list to 26 notions.

Friedler et al. [75] presented a comparative study conducted over a set of discrimination-aware algorithms (see Sections 6 to 8) in which they analysed achieved performance on five real-world datasets. The performance is evaluated on default accuracy measures of the machine learning literature, as well as on eight notions of fairness. The authors took four algorithms of the fairness literature into account (two naïve bayes, disparate impact remover, prejudice remove regularizer, and avoiding disparate mistreatment, as well as common machine learning algorithms from literature as baseline approaches. Note that they only considered pre- and in-processing approaches. The work of Friedler et al. contrasts ours as we do not aim to give a performance evaluation of existing algorithms, but rather to describe their algorithms and set them into context with related ones. This allows us to capture a broader range of algorithms as we do not bestow ourselves with the burden of unified implementation and training procedures.



The book ‘Fairness in machine learning’ by Barocas et al. [14] appears to be a promising addition to the fairness literature, aiming for much the same goals as this article.<sup>3</sup> Unfortunately, it is still a work in progress with most of the essential chapters remaining yet to be released. As the authors explicitly solicit for feedback “[i]n the spirit of open review”, we think it is important for the community to actively increase visibility on their project. Eventually the book could contribute in the same meaningful manner to the fairness community, as the Deep Learning book [84] did to the deep learning community.

### 3 Necessary Nomenclature

Before Section 5 introduces various fairness notions, some common ground regarding proper nomenclature might be of need. In the following, Sections 3.1 and 3.2 will define the terminology of machine learning and of fairness problems respectively.

#### 3.1 Machine Learning Nomenclature

Generally speaking, a machine learning algorithm processes a set of input data, the *training set*, and aims to distinguish patterns inside those data. Given new data, it can detect those found patterns in the *samples* and map them onto corresponding predictions. Those predictions might be a classification, where input is assigned an assumed outcome class, or regression, where a continuous value is assigned.

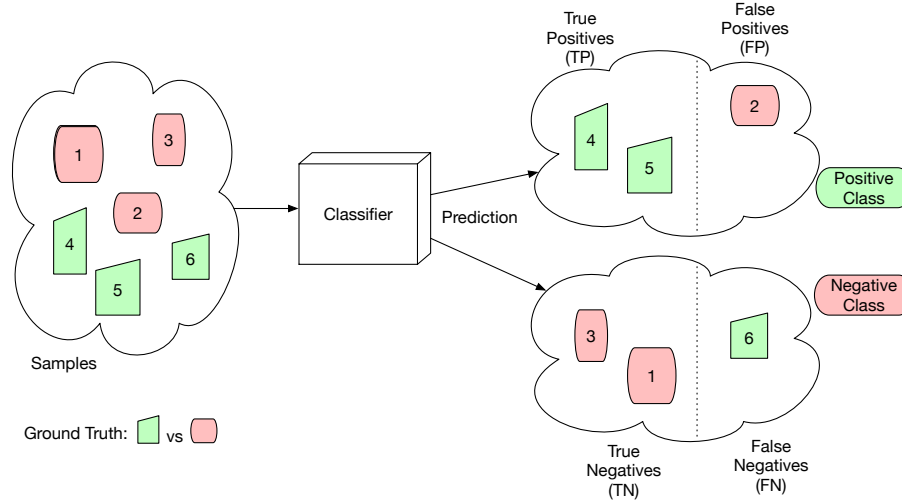
*Training Data and Samples.* Training Data is a collection of data samples which are used to train the machine learning algorithm for its eventual prediction task. Regarding fairness and discrimination such a data set usually refers to a population of humans, with each individual’s personal attributes defining a data sample.

*Classifier.* A machine learning predictor which assigns a class to each data sample is known as a classifier. Classification happens over a finite set of distinct classes. In the scope of fairness, this article is mainly concerned with binary classification, e.g. predicting between only two distinct classes. For instance, assume a binary classifier that is used to decide about the credit worth of an individual. The two classes in use would be ‘approve credit’ or ‘deny credit’.

*Positive and Negative Class.* As binary classification is used for automated decision making, the output classes correspond to one positive class (the yes-decision, ‘approve credit’) and to one negative one (the no-decision, ‘deny credit’).

---

<sup>3</sup> This judgement is based on a comparison of their announced outline of chapters with our section outline.



**Fig. 1.** Overview of Classification

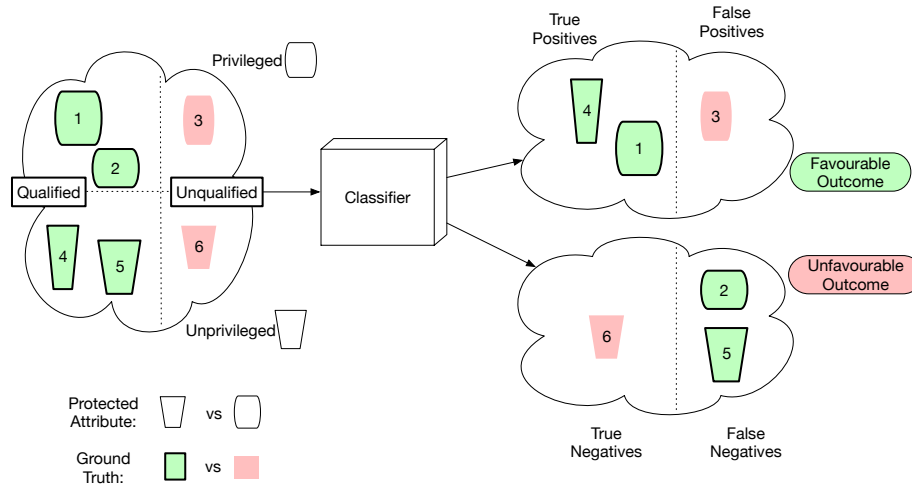
*Prediction and Ground Truth.* This machine learning nomenclature differentiates between what an individual belongs to and what the classifier predicts. Assuming a classifier is used to determine whether an individual is given a credit or not. The individual could be creditworthy (ground truth) but the classifier could still reject the credit application, as it classifies the individual to not be creditworthy (prediction). In the learning algorithms considered later on, for each individual in the training data the corresponding ground truth is known.

*True and False Positives.* A true positive (TP) is a sample which was correctly classified to belong to the positive class, i.e. the ground truth corresponds to the positive class as well. If the ground truth would actually have been the negative class, then it is called a false positive (FP). For instance, a non-creditworthy individual which still has their credit approved would be a false positive.

*True and False Negatives.* Analogous to true and false positives, a true negative (TN) is a sample which was correctly classified to belong to the negative class according to its ground truth. A false negative (FN) corresponds to a negatively classified sample which actually should have been positive respectively. For instance, a creditworthy individual which still has their credit approved would be a false negative.

### 3.2 Fairness Nomenclature

A fair classifier should guarantee that the *predictions* assigning the *favourable outcome* to individuals of the *privileged and unprivileged groups* do not discriminate over the *protected attribute*.



**Fig. 2.** Fairness and Classification

*Protected Attribute.* A property of an individual that must not influence the decision process of the machine learning algorithm is called a protected attribute. Typical examples include sex, race, religion, age, or caste. In a decision scenario, there can be multiple protected attributes.

*Privileged and Unprivileged Group.* Given a binary protected attribute like sex (male, female)<sup>4</sup>, the individuals over which decisions are made are divided into two demographic groups: sex = male and sex = female.

Assuming discrimination against one group (i.e. females), the other group experiences a favourable treatment. The latter group is referenced to as the privileged group, whereas the group experiencing discrimination is known as the unprivileged group.

*Favourable and Unfavourable Outcome.* In a binary classification scenario, the positive class corresponds to the favourable outcome the individuals wish to achieve, whereas the negative class corresponds to an unfavourable outcome respectively.

*Qualified and Unqualified Individuals.* Regardless of the protected attributes of an individual, the individual might be qualified for the favourable outcome or not. This corresponds to the ground truth whether an individual should be classified as belonging to the positive class or not. For instance, a creditworthy individual which still has their credit approved would be a false negative. For instance, an individual which is approved for a credit (favourable outcome) might actually be non-creditworthy (hence, unqualified).

<sup>4</sup> This example does not account for non-binary genders, but only the determined sex at birth.

## 4 Mathematical Notation

As usual, we denote by  $P(A | B) = P(A \cap B)/P(B)$  the conditional probability of the event  $A$  happening given that  $B$  occurs.

In the following, assume a finite dataset of  $n$  individuals  $\mathcal{D}$  in which each individual is defined as a triple  $(X, Y, Z)$ :

- $X$  are all attributes used for predictions regarding the data sample.
- $Y$  is the corresponding ground-truth of the sample.
- $Z$  is a binary protected attribute,  $Z \in \{0, 1\}$ , which might be included in  $X$  and hence used by the predictor.

The privileged group will be denoted with  $Z = 1$ , whereas  $Z = 0$  corresponds to the unprivileged group. The favourable and unfavourable outcomes correspond to  $Y = 1$  and  $Y = 0$  accordingly.

For a sample in Figure 2, we would have:

- $Y=1$  for green samples with a solid outline,
- $Y=0$  for red samples with no outline,
- $Z=1$  for rounded rectangles,
- $Z=0$  for trapezoids,
- the attributes  $X$  are not visible in the figure.

A (possibly unfair) classifier is a mapping  $h : X \rightarrow [0, 1]$ , yielding a score  $S = h(X)$  which corresponds to the predicted probability of an individual to belong to the positive class. For a given threshold  $\sigma$  the individual is predicted to belong to the positive class  $Y = 1$  iff  $h(X) > \sigma$ .

The final prediction based on the threshold is denoted as  $\hat{Y}$  with

$$\hat{Y} = 1 \Leftrightarrow h(X) > \sigma.$$

In the rest of the article we assume binary classifiers and always talk about a given fixed classifier.

Hence,

$$P(\hat{Y} = 1 | Z = 1)$$

represents the probability that the favourable outcome will be predicted for individuals from the privileged group, whereas

$$P(Y = 0 | Z = 0, \hat{Y} = 1)$$

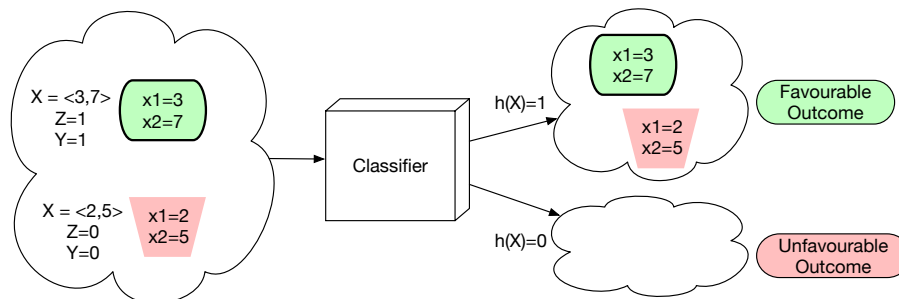
represents the probability that a positively classified individual from the unprivileged group is actually unqualified. Positive examples for these two cases are illustrated in Figure 3.

To keep notation short, let

$$P_i(E) := P(E | Z = i) \quad i \in \{0, 1\}$$

and more generally, let

$$P_i(E | C) := P(E | C, Z = i) \quad i \in \{0, 1\}$$



**Fig. 3.** Two example classifications

define the probability of event  $E$  conditional to  $C$  given an individual's protected attribute to be  $Z = i$ . For instance,

$$P_0(\hat{Y} = 0 | Y = 1) = P(\hat{Y} = 0 | Y = 1, Z = 0)$$

is the probability of an individual in the unprivileged group to be mispredicted for the unfavourable outcome, despite being qualified.

In the following Sections 5 to 8, we present notions and algorithms found in the fairness literature. In an attempt to provide a unified, consistent notation, the notation and variable naming might slightly differ from the original papers.





## 5 Notions of Fairness

For achieving a fair predictor, a metric on how to measure fairness is needed first. Depending on the use case, however, what is to be perceived as fair differs. This leads to multiple different notions of fairness, some of which were already compiled separately by Gajane and Pechenizkiy [76] Verma and Rubin [194], as well as Friedler et al. [75].

In line with Verma and Rubin [194], we will list the various fairness notions together with their formal definitions. Besides those notions already compiled by Verma and Rubin, the list of notions is expanded where applicable. Further, our summary provides visual, minimal examples for the given, parity-based notions. The example visualisations are defined in Table 1.

In line with Gajane and Pechenizkiy [76], we split the different notions into seven categories:

1. unawareness (Section 5.1),
2. group fairness (Section 5.2),
3. predictive parity (Section 5.3),
4. calibration (Section 5.4),
5. individual fairness (Section 5.5),
6. preference-based fairness (Section 5.6), and
7. causality (Section 5.7).

	Image Qualified?	Privileged?
	✓	✓
	✗	✓
	✓	✗
	✗	✗

**Table 1.** Summary of example illustrations.

Further, we will provide some discussion present in the literature regarding the choice of fairness metrics in Section 5.8

### 5.1 Unawareness

Fairness through unawareness [87, 122] is fulfilled when  $Z \notin X$ , that is when the protected attributes are not used by the predictor.

This fairness notion avoids *disparate treatment* [15, 202] as described in Section 1.2 Formally, a binary classifier avoids disparate treatment [202] if:

$$P(\hat{Y} = y \mid X = x) = P(\hat{Y} = y \mid X = x, Z = z), \quad (1)$$

i.e. the knowledge of  $Z$  does not alter the outcome.

As already discussed, removing the protected attribute from the decision process is insufficient for training a non-discriminating predictor, as other features can serve as proxies for the protected attribute [15, 40, 151, 202].

### 5.2 Group fairness

Group fairness [60] (a.k.a. statistical parity [49, 60, 208], demographic parity [75, 122], equal acceptance rate [210], mean difference [212], benchmarking [168], affirmative action [141]) requires the probability for an individual to be assigned the favourable outcome to be equal across the privileged and unprivileged groups [76]:

$$P_1(\hat{Y} = 1) = P_0(\hat{Y} = 1). \quad (2)$$

In practice, Eq. (2) (as well as all parity-based fairness notions) can be relaxed by defining the notion of  $\epsilon$ -fairness: let  $\epsilon > 0$ , we say a parity based fairness notion is  $\epsilon$ -fair iff

$$|P_1(\hat{Y} = 1) - P_0(\hat{Y} = 1)| < \epsilon. \quad (3)$$

By considering the ratio between groups instead of their absolute difference, group fairness relates to the 80% rule of U.S. employment law [30]. This rule

favourable outcome	unfavourable outcome	fair?

**Fig. 4.** Example on Group Fairness.

states that the employment rate between the privileged and unprivileged group must not differ more than 20%.

A problem of group fairness is that it can be easily reached by a randomised classifier on the unprivileged group. Given a trained predictor, one can measure the acceptance rate of the privileged group  $P_1(\hat{Y} = 1)$ , then simply assign the favourable outcome to unprivileged individuals at random with respect to the measured acceptance rate [60].

**Conditional Statistical Parity.** Conditional statistical parity [49, 60, 112] expands upon group fairness by taking a set of legitimate factors  $L \subset X$  into account, over which a decision needs to be equal regardless of the protected attribute:

$$P_1(\hat{Y} = 1 \mid L = l) = P_0(\hat{Y} = 1 \mid L = l). \quad (4)$$

**Normalised Difference.** The normalised difference [210, 212] is the normalised difference of acceptance rates (group fairness). Let

$$d_{\max} = \min \left( \frac{P(\hat{Y} = 1)}{P(Z = 1)}, \frac{P(\hat{Y} = 0)}{P(Z = 0)} \right).$$

The normalised difference is defined as

$$\delta = \frac{P_1(\hat{Y} = 1) - P_0(\hat{Y} = 1)}{d_{\max}} \quad (5)$$

with  $\delta = 0$  indicating complete fairness and  $\delta = 1$  indicating maximum discrimination.

### 5.3 Predictive Parity

Group fairness is evaluated only on the prediction outcome alone. This can be expanded by also taking into account the ground truth of the samples. Predictive parity [46] (a.k.a. outcome test [168]) requires equal *precision* for all demographic groups. Precision hereby is the positive predictive value (PPV),





favourable outcome	unfavourable outcome	fair?
		✓
		✗

Fig. 5. Example on Normalised Difference.

that is the probability of a positive classified sample to be a true positive:  $PPV = \frac{TP}{TP+FP}$  [172, 194]. This results in the following parity to hold

$$P_1(Y = 1 | \hat{Y} = 1) = P_0(Y = 1 | \hat{Y} = 1). \quad (6)$$

In short, individuals for which the favourable outcome was predicted need to have equal probability to actually belong to the positive class in both groups.





favourable outcome	unfavourable outcome	fair?
		✓
		✗

Fig. 6. Example on Predictive Parity.

**Predictive Equality.** Instead of comparing the PPV, it is also possible to compare the false positive rate (FPR), i.e. the rate of actual negative samples to be assigned the favourable outcome:  $FPR = \frac{FP}{FP+TN}$ . This approach is known as predictive equality [49] (a.k.a. false positive error rate balance [46]).

In short, the probability for unqualified individuals to be classified for the favourable outcome needs to be equal in both groups. This formalises to

$$P_1(\hat{Y} = 1 | Y = 0) = P_0(\hat{Y} = 1 | Y = 0). \quad (7)$$

**Equality of Opportunity.** The notion of equality of opportunity [97, 122] (a.k.a. false negative error rate balance [46]) complements predictive equality. Instead of expecting equal FPRs, equal false negative rates (FNR) are required:  $FNR = \frac{FN}{FN+TP}$ . This means, the probability for qualified individuals to be classified for the unfavourable outcome has to be equal in both groups:

$$P_1(\hat{Y} = 0 | Y = 1) = P_0(\hat{Y} = 0 | Y = 1). \quad (8)$$



favourable outcome	unfavourable outcome	fair?
		✓
		✗

Fig. 7. Example on Predictive equality.

favourable outcome	unfavourable outcome	fair?
		✓
		✗

Fig. 8. Example on Equality of Opportunity.

**Equalised Odds.** Requiring both, predictive equality and equality of opportunity to hold leads to the definition of equalised odds [97] (a.k.a. disparate mistreatment [202], conditional procedure accuracy equality [25]). As the equality of FNRs is equivalent to the equality of true positive rates (TPR) [194], equalised odds can be formalised as

$$P_1(\hat{Y} = 1 | Y = i) = P_0(\hat{Y} = 1 | Y = i), \quad i \in 0, 1. \tag{9}$$

Note that in cases where the prevalence of qualified individuals is not equal in all groups, a classifier can only satisfy both, predictive parity and equalised odds once it achieves perfect predictions [46, 120, 202]. Hence, in domains where the prevalences differ between groups or a perfect predictor is impossible, only one fairness notion can be satisfied at any time.

favourable outcome	unfavourable outcome	fair?
		✓
		✗

Fig. 9. Example on Equalised Odds.

**Conditional Use Accuracy Equality.** The requirement for a classifier to not only satisfy predictive parity (i.e. equal PPVs for both groups), but also to have equal negative prediction values (NPV) across groups, defines the conditional use accuracy equality [25]. NPV is defined as  $\frac{TN}{TN+FN}$ , which leads to the formalisation of this fairness notion as

$$P_1(Y = \hat{Y} \mid \hat{Y} = i) = P_0(Y = \hat{Y} \mid \hat{Y} = i), \quad i \in 0, 1. \quad (10)$$

In short, this notion is satisfied if the true positive rates and the true negative rates each are equal for both groups. Note that this does not require the PPV to be equal to the NPV.







favourable outcome	unfavourable outcome	fair?
		
		

Fig. 10. Example on Conditional Use Accuracy Equality.

**Overall Accuracy Equality.** Expecting equal prediction accuracy on both groups leads to overall accuracy equality [25]:

$$P_1(Y = \hat{Y}) = P_0(Y = \hat{Y}). \quad (11)$$

As the name suggests, the accuracy [172] over both groups needs to be equal. In contrast to conditional use accuracy equality, this notion combines the focus on true positives and true negatives.







favourable outcome	unfavourable outcome	fair?
		
		

Fig. 11. Example on Overall Accuracy Equality.

**Treatment Equality.** Treatment equality [25] is satisfied if the ratio of false positives and false negatives is equal among groups:

$$\frac{FN_1}{FP_1} = \frac{FN_0}{FP_0} \quad \text{or equivalently} \quad \frac{FP_1}{FN_1} = \frac{FP_0}{FN_0}. \quad (12)$$

The term ‘treatment’ hereby used to convey that such ratios can be used as policy lever [25]. If the classifier produces more false negatives than false positives for the privileged group, this means more unqualified individuals receive the favourable outcome than the other way around. Given that the unprivileged group has an even ratio, the misclassified privileged individuals receive an unfair advantage.

favourable outcome						unfavourable outcome						fair?	
													✓
													✗

Fig. 12. Example on Treatment Equality.

### 5.4 Calibration

Calibration [46] (a.k.a. matching conditional frequencies [97], test-fairness [194]) is a notion which is accompanied by a score  $S$ , which is the predicted probability for an individual  $X$  to be qualified (i.e. the probability to have the favourable outcome assigned):  $S = P(\hat{Y} = 1 | X)$ .

A classifier is said to be calibrated if

$$P_1(Y = 1 | S = s) = P_0(Y = 1 | S = s), \quad \forall s \in [0, 1]. \quad (13)$$

That is, the probabilities for individuals with the same score to actually be qualified has to be equal for each score value.

**Well-calibration.** Well-calibration [120] (a.k.a. perfect calibration [153]) extends the previous notion by further requiring those probabilities to be equal to  $S$ :

$$P_1(Y = 1 | S = s) = s = P_0(Y = 1 | S = s), \quad \forall s \in [0, 1]. \quad (14)$$

The aim of this notion is to ensure that if a set of individuals has a certain probability of having the favourable outcome assigned, then approximately the same percentage of individuals is indeed qualified [194].

**Balance for negative class.** The balance for negative class notion [120] requires equal average scores between the set of unqualified individuals in both groups. Thus, no group’s unqualified individuals have a statistical advantage over those of the other group to be misclassified for the favourable outcome.

Verma and Rubin [194] formalised this notion as

$$E_1(S | Y = 0) = E_0(S | Y = 0). \quad (15)$$

**Balance for positive class.** Similar to the previous notion, balance for the positive class [120] is satisfied if both groups have an equal average score among their qualified individuals. This ensures that no group is in disadvantage by resulting in more false negatives.

It formalises equal to Eq. (15) [194]:

$$E_1(S | Y = 1) = E_0(S | Y = 1). \quad (16)$$

## 5.5 Individual Fairness

As the statistical measures seen so far in Sections 5.2 to 5.4 are mostly concerned with fairness over demographic groups, the features  $X$  inhibited by each individual are largely ignored [194]. Assume a classifier for which holds i.e.  $P_1(\hat{Y} = 1) = P_0(\hat{Y} = 1) = 0.7$ . However, whereas the privileged group is classified as expected, the unprivileged group is assigned the favourable outcome at random with a  $\frac{7}{10}$  chance. Eq. (2) is still satisfied and hence the classifier is fair under the notion of group fairness, but equally qualified individuals are treated differently depending on group [194]

To counter this, the notion of individual fairness [60, 76, 122, 135] (a.k.a. fairness through awareness [60]) is based on metrics above the individuals themselves, formulating a  $(D, d)$ -Lipschitz property.

Let  $D$  be a distance metric over the room of possible classifications, and let  $d$  be a distance metric over individuals. Then, a classifier is said to fulfil individual fairness if

$$D(h(x_i), h(x_j)) \leq d(x_i, x_j) \quad \forall x_i, x_j \quad (17)$$

where  $x_i, x_j$  denote individuals. That is, the distance of predicted outcomes must not be greater than the distance between the respective individuals in the first place was. In other words: similar individuals need to be similarly classified.

This notion is similar to that of monotone classification [58] in which additional to the training data a function is given for which the predictions need to be monotone.

Dwork et al. [60] have shown, that a predictor satisfying individual fairness also satisfies group fairness with certain bias.

**Causal discrimination.** The notion of causal discrimination [77] requires individuals from different demographic groups with otherwise equal attributes to receive equal outcome:

$$x_i = x_j \wedge z_i \neq z_j \implies h(x_i) = h(x_j), \quad (18)$$

with  $z_k$  being the protected and  $x_k$  being the unprotected attributes of individual  $k$ . Though similar to the notion of conditional statistical parity (cf. Eq. (4)) with  $L = X$ , this notion is not defined over probabilities. Hence, each pair of individuals with equal attributes receives the same outcome, whereas in conditional statistical parity it is sufficient when the same percentage of individuals with equal legitimate attributes receives the favourable outcome.

## 5.6 Preference-Based Fairness

Zafar et al. [202] relax the notions of disparate treatment and disparate impact by proposing two notions of preference-based fairnesses. Those notions, as they state, are rooted on the economic and game theoretic concepts of fair division and envy-freeness [28,143,193]. The difference to group-based fairnesses as introduced above (Sections 5.2 and 5.3) is that the demographic groups do not need to experience the same treatment but merely need to prefer their received treatment over the treatment they would have received in another group.

Given a classifier  $h$  which yields prediction  $\hat{Y}$ , the group benefit  $\mathcal{B}_z$  of the subset of individuals with shared protected attribute  $z$  is defined as [202]

$$\mathcal{B}_z(h) := E(\hat{Y} = 1 \mid Z = z). \quad (19)$$

That is the fraction of favourable outcomes for individuals with protected attribute  $z$ .

**Preferred Treatment.** Let  $h_z$  denote a group-conditional classifier, i.e.  $h = \{h_z\}_{z \in Z}$ . Preferred treatment [202] is satisfied if

$$\mathcal{B}_z(h_z) \geq \mathcal{B}_z(h_{z'}) \quad \forall z, z' \in Z. \quad (20)$$

In other words, each group receives a better outcome on average by their given treatment as opposed to as treated as another group.

Note, that if  $h$  satisfies fairness through unawareness, i.e.  $h_a = h \forall a$ , then  $h$  also satisfies preferred treatment.

**Preferred Impact.** Let  $h'$  be a classifier which avoids disparate impact (i.e. it satisfies group fairness). A classifier  $h$  offers preferred impact [202] over  $h'$  if

$$\mathcal{B}_z(h) \geq \mathcal{B}_z(h') \quad \forall z, z' \in Z. \quad (21)$$

Thus each group receives at least as often the favourable outcome over  $h$  as it would have over  $h'$ , maintaining the core fairness achievable by  $h'$  [202].

## 5.7 Causality

This family of fairness notions assumes a given causal graph. A causal graph is a directed, acyclic graph representation having the features of  $X$  as vertices [118, 122, 142, 149]. Let  $G = (V, E)$  be a causal graph, then two vertices  $v_i, v_j \in V$  have a directed edge  $(v_i, v_j) \in E$  between them if a direct causal relationship exists between them, i.e.  $v_j$  is (potentially) a direct cause of  $v_i$  [142].

**Counterfactual Fairness.** A classifier is counterfactually fair [122] if for all individuals the outcome is equal to the outcome of its counterfactual individual (i.e. the same individual with flipped protected attribute). That is,  $Z$  is not a cause for  $\hat{Y}$  in any instance, i.e. there is no path  $(Z, \dots, \hat{Y})$  in  $G$ .

This can be formalised as

$$P_z(\hat{Y}_z = Y \mid X = x) = P_z(\hat{Y}_{z'} = Y \mid X = x) \quad \forall z, z' \in Z \quad (22)$$

where  $\hat{Y}_z$  is the prediction yielded by the classifier if  $Z = z$ .

Note the relation between counterfactual fairness and individual fairness: counterfactual imply a notion of similarity. Kusner et al. [122] state that individual fairness “can be defined by treating equally two individuals with the same [set of attributes] in a way that is also counterfactually fair.”

**No Unresolved Discrimination.** Unresolved discrimination [118] is avoided if there exists no path  $(Z, v_1, \dots, v_n, \hat{Y})$  in  $G$ , except via a resolving variable  $v_i, i \in \{1, \dots, n\}$ .

A resolving variable is a variable which is dependent on  $Z$ , but in a manner which is understood and accepted as non-discriminatory [118].

As resolved paths  $Z, \dots, \hat{Y}$  are allowed, this notion is a relaxation on counterfactual fairness.

**No Potential Proxy Discrimination.** This notion is dual to that of unresolved discrimination [118]. A proxy [182] is a descendant of  $Z$  which should not affect the prediction. It is meant to be a “clearly defined observable quantity that is significantly correlated [with  $Z$ ]”, as stated by Kilbertus et al. [118].

Potential proxy discrimination [118] is observed when there exists a path  $(Z, v_1, \dots, v_n, \hat{Y})$  in  $G$  which is blocked by a proxy variable  $v_i, i \in \{1, \dots, n\}$ .

**No Proxy Discrimination.** This notion is a refinement of the previous one. Let  $P$  be a proxy for  $Z$ . While potential proxy discrimination can be avoided by simply designing a classifier to be unaware of  $P$ , a classifier  $h(X, P)$  can be carefully tuned to cancel the influence of  $P$  on  $X$  [118]. This is known as intervention on  $P$  [149], denoted as  $do(P = p)$ , replacing  $P$  by putting point mass on value  $p$  [118].

Now, a classifier exhibits no proxy discrimination based on  $P$  if [118]

$$P(\hat{Y} \mid do(P = p)) = P(\hat{Y} \mid do(P = p')) \quad \forall p, p'. \quad (23)$$

Remark that if there exists no such path over  $P$  in  $G$ , fairness through unawareness satisfies no proxy discrimination [118].

**Fair Inference.** The notion of fair inference [142] proposes to select a set of paths in  $G$  which are classified as legitimate paths. Which paths to choose is hereby a domain-specific problem. Along legitimate paths the protected attribute

is treated as having its active value  $Z = z$ , whereas it is treated as baseline value  $Z = z'$  on any non-legitimate path.

The idea stems from splitting the average causal effect of  $Z$  on  $\hat{Y}$  into path-specific effects [150], which can be formulated as nested counterfactuals [167]. As long as  $\hat{Y}$  is a descendant of  $Z$  only by legitimate paths, the outcome is deemed as fair.

### 5.8 Note on the Selection of Fairness Metrics

The notion of total fairness, which corresponds to satisfying all fairness conditions, is unfortunately shown to be impossible. For unequal distribution of groups in the population it was shown that group fairness (Eq. (2)), equalised odds (Eq. (9)), and conditional use accuracy equality (Eq. (10)) are mutually exclusive [25, 46, 74, 212]. Also, group fairness, equalised odds, and calibration (Eq. (13)) as well as group fairness and predictive parity contradict each other [46, 153, 212] given unequal base rates.

This leads to the problem of deciding which fairness measures are desirable. Žliobaitė [212] recommends defaulting to normalised difference while discouraging the use of ratio-based measures due to interpretability issues. He further finds the use of core measures standalone is insufficient as fairness criteria. Core measures are hereby measures which are unconditional over the whole population, group fairness for instance. These are, given unequal distribution of qualification throughout the groups, not applicable to the problem [212] and should be set into a conditional context (i.e. segmenting the population beforehand).

Saleiro et al. give for their Aequitas tool [165] a fairness tree, guiding a user through the decision process of finding a suitable fairness notion to follow.<sup>5</sup> The notions considered are group fairness, disparate impact, predictive parity, predictive equality, false omission rate parity, and equality of opportunity. More information to Aequitas and the fairness tree is listed in Section 9.2.

Another discussion is the long-term impact onto the populations depending on choice of fairness notion. Mouzannar et al. [141] consider affirmative action, which corresponds to either reducing the positive prediction rates of the privileged group or increasing that of the unprivileged group. Specifically, they analysed the conditions under which society equalises for both policies, i.e. achieving equal qualification rates among groups. For instance, consider the case of a company hiring equal rates of qualified individuals among groups, say 20% each. Given that 40% of the privileged and 10% of the unprivileged groups are indeed qualified, this results in hiring 8% and 2% of individuals in those groups respectively. This is fair under predictive parity. However, more privileged individuals were hired than unprivileged, giving them better paying jobs and hence more resources to provide to their children. Hence, the number of qualified individuals in the privileged group can grow stronger in the next generation than in the unprivileged group, increasing the initial gap between them further. Paaßen et al. [147] provide a theoretical approach, which shows that only the enforcement

<sup>5</sup> <http://dsapp.uchicago.edu/aequitas/>

of group fairness actually results in an equilibrium rather than increasing the gap between groups.

Other discussion is concerned with whether the research should rather focus on sub-group fairness [10, 47, 148]. As a predictor which achieves a fairness criterion over the whole population still can show significant differences in classification results on certain subgroups, there always exists a fraction of individuals for whom the predictions can be perceived as unfair [47]. Chouldechova and G'Sell [47] hence, have proposed a model which can automatically find such subgroups on which the fairness measure differs. Such techniques allow for further discrimination analysis beyond the scope of the overall performance.

## 6 Pre-Processing: Discrimination-free Training Data

The method of pre-processing is the approach of removing the bias from the training data such that the predictor does not have to account for discrimination. Instead, during training only fair examples are shown, resulting in a fair classifier. Depending on the pre-processing technique, this has to be applied to each individual in the final production system as well.

In the following, we divided several pre-processing techniques from literature into three categories:

- relabelling (Section 6.1),
- resampling (Section 6.2), and
- fair representations (Section 6.3).

### 6.1 Relabelling

Relabelling approaches aim to alter the ground truth values in the training set such that it satisfies the fairness notion.

**Massaging.** The pre-processing technique of massaging the data [38, 107, 109] takes a number of individuals in the training data and changes their ground truth values. This allows any classifying machine learning algorithm (Bayesian networks, support vector machines, ...) to learn on a fair dataset, aiming to fulfil group fairness (Section 5.2).

For this, a ranker  $R$  is employed which ranks the individuals by their probability to receive the favourable outcome. The more likely the favourable outcome is, the higher the individual will rank.

Let  $\epsilon = P_1(Y = 1) - P_0(Y = 1)$  denote the measured discrimination of the training data (cf. group fairness, Section 5.2). The number  $M$  of required modifications is calculated as [107]

$$M = \epsilon \times \frac{|\mathcal{D}_1| \times |\mathcal{D}_0|}{|\mathcal{D}_1| + |\mathcal{D}_0|}. \quad (24)$$



where

$$\begin{aligned}\mathcal{D}_1 &= \{X \mid Z = 1\} \quad \text{and} \\ \mathcal{D}_0 &= \{X \mid Z = 0\}\end{aligned}$$

denote the sets of privileged and unprivileged individuals respectively.

The massaging happens on the sets  $pr = \{X \in \mathcal{D}_0 \mid Y = 0\}$  and  $dem = \{X \in \mathcal{D}_1 \mid Y = 1\}$  by sorting both sets w.r.t. their ranks:  $pr$  descending and  $dem$  ascending. Labels of the top- $M$  individuals in both sets gets flipped (i.e. massaged), which are respectively the  $M$  individuals closest to the decision border.

## 6.2 Resampling

Resampling methods impact the sampling rate of the training data by either dropping or doubling specific samples or altering their relative impact at training time.

**Reweighting.** The authors of the massaging technique also introduced the method of reweighting [38, 109].

As massaging is rather intrusive as it alters the dataset, this alternative keeps the dataset intact but associates to each individual  $X$  with  $Y = y, Z = z$  a weight  $W_X = W(y, z)$  with

$$W(y, z) = \frac{|\{X \mid Z = z\}| \times |\{X \mid Y = y\}|}{|\mathcal{D}| \times |\{X \mid Z = z \wedge Y = y\}|}. \quad (25)$$

The weighed dataset can now be used for learning a fair classifier. Note that there are only four different weights, depending on whether the individual is privileged or not and qualified or not.

Drawback of this method is the need of a classifier, which is able to incorporate the weights.

**Preferential Sampling.** The authors of reweighting additionally proposed preferential sampling [108, 109] to counter the need of a learner that is able to incorporate the individual's weights. Following the same idea as used in massaging that individuals close to the decision border are more likely to suffer from expected discrimination, again a ranker is employed to determine individuals closest to the border.

The data is divided into four subsets  $\mathcal{D}_0^0, \mathcal{D}_0^1, \mathcal{D}_1^0$ , and  $\mathcal{D}_1^1$ , where

$$\begin{aligned}\mathcal{D}_z &= \{(X, Y, Z) \mid Z = z\}, \\ \mathcal{D}^y &= \{(X, Y, Z) \mid Y = y\}, \text{ and} \\ \mathcal{D}_z^y &= \mathcal{D}_z \cap \mathcal{D}^y = \{(X, Y, Z) \mid Y = y \wedge Z = z\}\end{aligned}$$

for  $y, z \in \{0, 1\}$ . For instance,  $\mathcal{D}_0^1$  is the set of all unprivileged, qualified individuals. For each of these sets, the expected cardinality is calculated by [108]

$$C_z^y = \frac{|\mathcal{D}_z| \times |\mathcal{D}^y|}{|\mathcal{D}|} \quad (26)$$

and the sets are sorted according to their ranks:  $\mathcal{D}_0^1, \mathcal{D}_1^1$  ascending,  $\mathcal{D}_0^0, \mathcal{D}_1^0$  descending. Each of the four sets is then adjusted to match their respective  $C_z^y$  values by either deleting the top elements or iteratively duplicating them. The duplication step puts the respective top element and its copy to the bottom of the list before the next sample is duplicated.

Finally, the pre-processed dataset is the union of the modified sets  $\mathcal{D}_z^y$ .

### 6.3 Fair Representations

The approach of finding fair representations is related to the notion of representation learning [22]. For a biased dataset  $\mathcal{D}$  an alternative, cleaned dataset  $\tilde{\mathcal{D}}$  is constructed which obscures the original bias but still is as similar as possible to the original data  $\mathcal{D}$ . McNamara et al. [137] discuss the costs and benefits of fair representations, showing that “any use of the cleaned data will not be too unfair”.

**Optimized Pre-Processing.** The aim of optimized pre-processing [41] is to transform the dataset  $\mathcal{D} = \{(X_i, Z_i, Y_i)\}_i$  into  $\{(\tilde{X}_i, \tilde{Y}_i)\}_i$  by finding an appropriate, randomised mapping  $p_{\tilde{X}, \tilde{Y}|X, Y, Z}$ . This is achieved by solving the optimisation problem

$$\begin{aligned} \min_{p_{\tilde{X}, \tilde{Y}|X, Y, Z}} \quad & \Delta(p_{\tilde{X}, \tilde{Y}}, p_{X, Y}) \\ \text{s.t.} \quad & D(p_{\tilde{Y}|Z}(y|z), p_{Y_T}(y)) \leq \epsilon_{y, z} \text{ and} \\ & E(\delta((x, y), (\tilde{X}, \tilde{Y})) | Z = z, X = x, Y = y) \leq c_{z, x, y} \forall (x, y, z) \in \mathcal{D}, \\ & p_{\tilde{X}, \tilde{Y}|X, Y, Z} \text{ is a valid distribution} \end{aligned} \quad (27)$$

where  $D(\cdot, \cdot)$  is some distance metric,  $\delta(\cdot, \cdot)$  is a distortion metric, and  $\Delta(\cdot, \cdot)$  is a given dissimilarity measure between probability distributions. Given that  $\Delta$  is (quasi)convex and  $D$  is quasiconvex in their first respective arguments, the optimisation problem in Eq. (27) itself is (quasi)convex [41]. Hence, it can be solved optimally. The thresholds  $\epsilon_{y, z}$  and  $c_{x, y, z}$  are to be chosen by the user. Hereby, individual fairness is promoted due to the distortion control values  $c_{x, y, z}$  being defined pointwise, hence they can depend on  $X, Y, Z$  if so desired.

**Disparate Impact Remover.** Contrary to optimized preprocessing, the disparate impact remover [70] is based on only changing the features  $X$  instead of also the ground-truth  $Y$  of individuals for achieving group fairness.

The goal is to transform a dataset  $\mathcal{D}$  which contains disparate impact into a repaired dataset  $\tilde{\mathcal{D}}$ . This is done by mapping each  $(X, Y, Z) \in \mathcal{D}$  to an associated triple  $(\tilde{X}, Y, Z) \in \tilde{\mathcal{D}}$ .

Let  $X = (X^{(1)}, \dots, X^{(m)})$ . For  $k = 1, \dots, m$  let  $\xi = X^{(k)}$ . Let  $\xi_z = P(\xi | Z = z)$  denote the marginal distribution on  $\xi$  conditioned on  $Z$ , let  $F_z : \xi_z \rightarrow [0, 1]$ , and let  $F_z^{-1} : [0, 1] \rightarrow \xi_z$ .  $F_z$  is a cumulative distribution function over the values for  $\xi$ , whereas  $F_z^{-1}$  is the associated quantile function [70]. For instance,  $F_z^{-1}(\frac{1}{2})$  is the value  $x$  with  $P_z(\xi = x) = \frac{1}{2}$ .

The needed transformation  $\xi \mapsto \tilde{\xi}$  is done by

$$\tilde{\xi} = \text{median}_{\forall z'} F_z^{-1}(F_z(\xi)). \quad (28)$$

**Variational Fair Autoencoder.** By treating  $Z$  as a nuisance variable the pre-processing problem becomes a problem of domain adaption [127, 131]. Removing the domain  $Z$  leads hereby to improved fairness, aiming for a latent representation  $\tilde{X}$  of  $X$  which is minimally informative about  $Z$  yet maximally informative about  $Y$ .

In this regard, Louizos et al. propose their method of employing a variational fair autoencoder [131], based on deep variational autoencoders [119, 157]. This model encourages separation between  $\tilde{X}$  and  $Z$  by using factorised priors  $P(\tilde{X})P(Z)$  and avoiding keeping dependencies in the variational posterior  $q(\tilde{X} | X, Z)$  by employing a maximum mean discrepancy term [86].

In prior work, Li et al. propose a similar method called learning unbiased features [127], employing a single autoencoder, also with maximum mean discrepancy term. The variational fair autoencoder however is semi-supervised and conducted with two deep neural networks [84, 161] building a two-layer pipeline. The first network takes  $X, Z$  as input, yielding an encoding  $\tilde{X}'$  which is invariant to  $Z$ . The second expects  $\tilde{X}', Y$  as input and yields the final encoding  $\tilde{X}$  which has information to its corresponding ground-truth injected.

## 6.4 Further Algorithms

Further algorithms, which we did not summarise above, include rule protection [92, 93], adversarial learned fair representations [61], fairness through optimal transport theory [16], k-NN for discrimination prevention [135], situation testing [21], statistical framework for fair predictive algorithms [134], continuous framework for fairness [91], sensitive information remover [106], fairness through awareness [60], provably fair representations [136], neural styling for interpretable representations [154], and encrypted sensitive attributes [117].

## 7 In-Processing: Discrimination-aware Learners

In-processing techniques consider the training of a fair classifier on a possibly discriminatory dataset. This includes ensembles, novel or adjusted algorithms, or adding a regularization term to the loss function.

We categorised five common categories of in-processing methods:

- adjusted learning algorithms (Section 7.1),
- adapted loss functions (Section 7.2),
- adversarial approaches (Section 7.3),
- optimisation subject to fairness constraints (Section 7.4), and
- compositional approaches (Section 7.5).

## 7.1 Adjusted Learning Algorithms

Algorithms in this category describe changes onto common machine learning algorithms which ensure fairness-awareness.

**Two Naïve Bayes** Early work by Calders and Verwer [39] proposed to train a naïve Bayes classifier for each protected attribute and balance them in order to achieve group fairness. The models  $M_z$  for  $z \in \{0, 1\}$  are trained on  $\mathcal{D}_z = \{(X, Y, Z) \in \mathcal{D} \mid Z = z\}$  only and the overall outcome is determined for an individual by the outcome of the model corresponding to the individual’s respective protected attribute.

Let  $X$  consist of  $m$  features  $X = \langle X^{(1)}, \dots, X^{(m)} \rangle$ . As the overall model depends on  $Z$ , the model can be formalised as

$$P(X, Y, Z) = P(Y \mid Z) \prod_{i=1}^m P(X^{(i)} \mid Y, Z) \quad (29)$$

which is equal to the two different naïve Bayes models for the values of  $Z$  [39]. Hereby, the probability  $P(Y \mid Z)$  is modified as in the authors’ post-processing approach ‘modifying naïve Bayes’ in Section 8.

The authors argue that removing  $Z$  from the feature set results in too big of a loss in accuracy, hence keeping the protected attribute for classification is sensible. This however is unsuitable where any decision making use of the protected attribute is forbidden by law (cf. Section 1.1).

**Naïve Bayes with Latent Variable.** A more complex approach than the balanced bayes ensemble, also proposed by Calders and Verwer [39], is to model the actual class labels  $L$  the dataset would have had if it had been discrimination free to begin with. This is done by treating  $L$  as latent variable under two assumptions:

1.  $L$  is independent from  $Z$ .
2.  $Y$  is determined by discriminating over  $L$  using  $Z$  uniformly at random.

To determine  $L$ , the authors propose a variation of the expectation maximization algorithm (EM) [53] which utilises prior knowledge. In the E-step of EM, the expected values for  $L$  are computed. However, as relabelling individuals with  $Z = 0, Y = 1$  or  $Z = 1, Y = 0$  would only increase discrimination [39] they will stay fixed with  $L = Y$ . Further, the distribution  $P(Y \mid L, Z)$  can be pre-computed as distribution which ensures group fairness as in Eq. (2).

**Discrimination Aware Decision Tree Construction.** Another interesting approach is proposed by Kamiran et al. [110] which alters the splitting heuristic used for learning decision trees [33, 155].

Generally, the decision tree induction iteratively splits the dataset  $\mathcal{D}$  based on the attribute leading to the highest information gain until only leaves in which all datapoints share the same ground-truth remain. Assume a split which divides the data into  $k$  different datasplits  $\mathcal{D}_1, \dots, \mathcal{D}$ . The information gain over the ground-truth is defined as

$$IG_Y = H_Y(\mathcal{D}) - \sum_{i=1}^k \frac{|\mathcal{D}_i|}{|\mathcal{D}|} H_Y(\mathcal{D}_i) \quad (30)$$

where  $H_Y$  denotes the entropy with respect to the ground-truth. By accounting for the entropy  $H_Z$  over the protected attribute, the discrimination gain can be measured by

$$IG_Z = H_Z(\mathcal{D}) - \sum_{i=1}^k \frac{|\mathcal{D}_i|}{|\mathcal{D}|} H_Z(\mathcal{D}_i). \quad (31)$$

For determining the most suitable split during training time, three different combinations of  $IG_Y, IG_Z$  are possible:  $IG_Y - IG_Z$  to only allow non-discriminatory splits,  $IG_Y/IG_Z$  to make a trade-off between accuracy and fairness, and  $IG_Y + IG_Z$  to increase both, accuracy and unfairness. Although the third heuristic actually increases discrimination in the tree, the authors state that it leads to good results in combination with their proposed post-processing technique of discrimination aware decision tree relabelling [110] (Section 8).

## 7.2 Adapted Loss Function

Algorithms described here may depend on certain machine learning algorithms, yet leave their training procedure unchanged. Instead of altering the algorithm, the loss function is adapted to account for fairness, either by swapping the whole loss function altogether or by adding a regularization term.

**Prejudice Remover Regularizer.** Kamishima et al. [113, 114] trained a logistic regression model [50] with a regularizer term which aims to reduce the prejudice learned by the model. Prejudice is hereby divided into direct and indirect prejudice. Direct prejudice [151] occurs when the outcome is correlated with the protected attribute. Indirect prejudice (a.k.a. indirect discrimination [151]) occurs when the outcome is not directly related with the protected attribute, but correlation can be observed given  $X$ .

To measure prejudice, Kamishima et al. defined the (indirect) prejudice index (PI) [113, 114]

$$PI = \sum_{Y,Z} P(Y, Z) \ln \frac{P(Y, Z)}{P(Y)P(Z)}. \quad (32)$$

Let  $\Theta$  denote the parameters of the prediction model  $h$ . Building upon Eq. (32) the prejudice removing regularization term

$$R_{PR}(\mathcal{D}, \Theta) = \sum_{(x,z) \in \mathcal{D}} \sum_{y \in \{0,1\}} h(x; \Theta) \ln \frac{P(\hat{Y} = y | Z = z)}{P(\hat{Y} = y)} \quad (33)$$

is formed [113]. The authors further propose a general framework that utilises two regularizer terms. One to ensure fairness, e.g. Eq. (33), and one standard regularizer to reduce overfitting, e.g.  $L_2$  regularization  $\|\Theta\|_2^2$  [102].

**Learning Fair Representations.** Zemel et al. propose an approach for learning fair representations [208]. This method actually resembles a mix of pre- and in-processing, as the input data is mapped into an intermediate representation over which the decision takes place. As it is possible for this algorithm to learn an appropriate distance function, it is actually well-suited for achieving individual fairness (Section 5.5)

The intermediate representation is that of  $K$  learned prototypes. A vector  $v_k$  is associated to each prototype, which lies in the same feature space as  $X$ . Let  $C$  denote the multinomial random variable representing one of the prototypes. The division of individuals to prototypes is to satisfy statistical parity (cf. Eq. (2))

$$P_1(C = k) = P_0(C = k). \quad (34)$$

Due to the prototypes lying in the same space as  $X$ , they induce a natural probabilistic mapping via the softmax function  $\sigma$  [208]

$$P(C = k | X = x) = \frac{\exp(-d(x, v_k))}{\sum_{j=1}^K \exp(-d(x, v_j))} = \sigma(d(x, v))_k \quad (35)$$

where  $d(\cdot, \cdot)$  is a distance measure (e.g. Euclidean distance). This mapping to prototypes is hence defined as discriminative clustering models in which each prototype acts as a cluster on its own [208].

The learning system for the prototypes should minimise the loss function

$$L = A_C L_C + A_X L_X + A_Y L_Y \quad (36)$$

which is minimised by L-BFGS [146].  $A_C, A_X, A_Y$  are hereby hyperparameters.  $L_C, L_X, L_Y$  are defined as

$$L_C = \sum_{k=1}^K |E(C = k | Z = 1) - E(C = k | Z = 0)|, \quad (37)$$

$$L_X = \sum_{(x,y,z) \in \mathcal{D}} \left( x - \sum_{k=1}^K p_k(x) v_k \right)^2, \quad (38)$$

$$L_Y = \sum_{(x,y,z) \in \mathcal{D}} -y \log(p_k(x) w_k) - (1 - y) \log(1 - p_k(x) w_k) \quad (39)$$

with  $p_k(x) = P(C = k | X = x)$  the probability for  $x$  to correspond to prototype  $k$  and  $w_k \in [0, 1]$  being the probabilistic outcome prediction for prototype  $k$ .  $L_C, L_X, L_Y$  ensure statistical parity, a fitting mapping of  $X$  to  $C$ -space, and an as accurate as possible prediction for the prototypes respectively.

Let  $m$  denote the amount of feature dimensions in  $X$ , i.e.  $x = \langle x_{(1)}, \dots, x_{(m)} \rangle$ . Each feature dimension can be individually weighted by a parameter  $\alpha_i$ , acting as inverse precision value in the distance function of Eq. (35)

$$d(x, v_k, \alpha) = \sum_{i=1}^m \alpha_i (x^{(i)} - v_k^{(i)})^2. \quad (40)$$

By optimizing  $\alpha, \{v_k\}_k, \{w_k\}_k$  jointly, the model learns its own distance function for individual fairness (Section 5.5). By utilising different weights  $\alpha^z$  for each demographic group, this distance metric also addresses the inversion problem [60], where features are of different impact with respect to the classification of the two groups.

### 7.3 Adversarial Approaches

Adversarial training [85] consists of employing two models which play against each other. On one side, a model is trained to accurately predict the ground truth, whereas a second model predicts the protected attribute by considering the first model’s prediction.

**Adversarially Learning Fair Representations.** Beutel et al. [29] propose an adversarial training to prevent biased latent representations closely related to the work of Edwards et al. [61] that allows to achieve the fairness notions of equality of opportunity as in Eq. (8).

This method makes use of two neural networks with a shared hidden layer  $g(X)$ . Assume a subset  $\mathcal{E} = (X_E, Z_E, Y_E) \subseteq \mathcal{D}$  from which  $Z$  can be observed. The goal of the predictor model  $h(g(X))$  is to correctly predict  $Y$  whereas the goal of the discriminator  $a(g(X_E))$  is to correctly predict  $Z$ . The overall model’s objective is defined over two loss functions, one for  $h$  and  $a$  respectively

$$\min \left[ \sum_{(x,y,z) \in \mathcal{D}} L_Y(h(g(x)), y) + \sum_{(x,y,z) \in \mathcal{E}} L_Z(a(J_\lambda(g(x))), z) \right] \quad (41)$$

whereby  $J_\lambda$  is an identity function with negative gradient, i.e.  $J_\lambda(g(X_E)) = g(X_E)$  and  $\frac{dJ_\lambda}{dX_E} = -\lambda \frac{dg(X_E)}{dX_E}$ . Without  $J_\lambda$ ,  $g(\cdot)$  would be encouraged to predict  $Z$  [29]. The  $\lambda$  parameter determines the trade-off between accuracy and fairness.

Beutel et al. observe that “remarkably small datasets”  $\mathcal{E}$  are already effective for more fair representations, and that further a balanced distribution over  $Z$  in  $\mathcal{E}$  yield more fair results [29].

**Adversarial Debiasing.** Similar to the previous method of adversarially learning fair representations, the method of adversarial debiasing [209] by Zhang et al. also makes use of adversarial training. However, their framework follows a different architecture and allows for one the notions of group fairness, equalised odds, and equality of opportunity (Eqs. (2), (8) and (9)).

The approach employs two distinct models, the predictor  $h(X)$  and the adversary  $a(\cdot)$ . Depending on the fairness to achieve, the input given to  $a(\cdot)$  changes. For group fairness, only the prediction score  $S = h(X)$  serves as input. For equalised odd, both the prediction score  $S$  and the ground truth  $Y$  are given. Given a target class  $y$ , restricting the input of the adversary to  $\mathcal{D}^y = \{(X, Y, Z) \mid Y = y\}$  achieves equality of opportunity. Again, the goal of the adversary is to predict  $Z$  correctly.

Assuming the loss functions  $L_P(\hat{y}, y)$  for the predictor and  $L_A(\hat{z}, z)$  for the adversary, and the model parameters  $W$  and  $U$  for predictor and adversary respectively.  $U$  is updated according to the gradient  $\nabla_U L_A$   $W$  is updated according to

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A \quad (42)$$

with  $\alpha$  being a tunable hyperparameter and  $\text{proj}_{\nu} x = 0$  if  $\nu = 0$ .

Remark that this approach is model agnostic, as long as the model is training using a gradient based method [209]. Further, the authors suggest using a simple adversary, whereas the predictor might be arbitrarily complex.

#### 7.4 Optimisation Subject to Fairness Constraints

This set of algorithms leaves the loss function unaltered and instead treats the loss optimisation as a constrained optimisation problem, having the fairness criterion as a constraint.

**Avoiding Disparate Mistreatment.** In their paper, Zafar et al [202] not only propose the notion of disparate mistreatment (a.k.a. equalised odds) (see Eq. (9)) but also propose a method of achieving a classifier which is free of disparate mistreatment [202, 203]. Given a decision boundary-based classifier, goal is to minimise loss subject to posed fairness constraints.

Let  $\Theta$  denote the parameters of such a classifier, and let  $d_{\Theta}(x)$  denote the signed distance of  $x$  to the respective decision boundary (e.g.  $d_{\Theta}(x) = \Theta^T x$  for linear models). The authors propose to measure disparate mistreatment via a tractable proxy over the covariance between  $z$  and  $d_{\Theta}(x)$  for mislabelled individuals  $(x, y, z) \in \mathcal{D}$  with

$$\text{Cov}(z, g_{\Theta}(y, x)) \approx \frac{1}{n} \sum_{(x, y, z) \in \mathcal{D}} (z - \bar{z}) g_{\Theta}(y, x) \quad (43)$$

where  $g_{\Theta}(y, x) = \min(0, y d_{\Theta}(x))$  [202] and  $\bar{z}$  denotes the arithmetic mean over  $(z_i)_{i=1}^n$ . This follows the approach of the disparate impact proxy proposed



in [203]. Hence, the constraints to the loss function formulate as

$$\begin{aligned}
 \min \quad & L(\Theta) \\
 \text{s.t.} \quad & \frac{1}{n} \sum_{(x,y,z) \in \mathcal{D}} (z - \bar{z})g_{\Theta}(y, x) \leq c, \\
 & \frac{1}{n} \sum_{(x,y,z) \in \mathcal{D}} (z - \bar{z})g_{\Theta}(y, x) \geq -c,
 \end{aligned} \tag{44}$$

where  $c \in \mathbb{R}^+$  is the covariance threshold, trading fairness for accuracy. The closer  $c$  is to zero, the higher the fairness, but the larger the potential loss in accuracy [203].

This can be converted into a disciplined convex-concave program [166] which can be solved efficiently for a convex loss  $L(\Theta)$ :

$$\begin{aligned}
 \min \quad & L(\Theta) \\
 \text{s.t.} \quad & \frac{-|\mathcal{D}_1|}{n} \sum_{(x,y) \in \mathcal{D}_0} g_{\Theta}(y, x) + \frac{-|\mathcal{D}_0|}{n} \sum_{(x,y) \in \mathcal{D}_1} g_{\Theta}(y, x) \leq c, \\
 & \frac{-|\mathcal{D}_1|}{n} \sum_{(x,y) \in \mathcal{D}_0} g_{\Theta}(y, x) + \frac{-|\mathcal{D}_0|}{n} \sum_{(x,y) \in \mathcal{D}_1} g_{\Theta}(y, x) \geq -c.
 \end{aligned} \tag{45}$$

Note that for  $Z \notin X$  this not only avoids disparate mistreatment, but also disparate treatment as well. Remark that this approach is restricted to convex margin-based classifiers [203].

**Accuracy Constraints.** Contrary to their previous approach of avoiding disparate mistreatment, Zafar et al. additionally propose a method which, rather than maximising accuracy under fairness constraints, aims to maximise fairness under accuracy constraints [203].

This formulates as

$$\begin{aligned}
 \min \quad & \frac{1}{n} \sum_{(x,y,z) \in \mathcal{D}} (z - \bar{z})\Theta^T x \\
 \text{s.t.} \quad & L(\Theta) \leq (1 + \gamma)L(\Theta^*)
 \end{aligned} \tag{46}$$

where  $L(\Theta^*)$  denotes the optimal loss of the respective unconstrained classifier and  $\gamma$  specifies the maximum additional loss to be accepted. For instance,  $\gamma = 0$  ensures maximally achievable fairness by retaining optimal loss.

Given a loss which is additive over the data points, i.e.  $L(\Theta) = \sum_{i=1}^n L_i(\Theta^*)$  where  $L_i$  is the individual loss of the  $i$ th individual, it is possible to fine-grain the constraints with individual  $\gamma_i$  to be

$$\begin{aligned}
 \min \quad & \frac{1}{n} \sum_{(x,y,z) \in \mathcal{D}} (z - \bar{z})\Theta^T x \\
 \text{s.t.} \quad & L_i(\Theta) \leq (1 + \gamma_i)L_i(\Theta^*) \quad \forall i.
 \end{aligned} \tag{47}$$

Setting  $\gamma_i = 0$  for individuals with ground-truth  $y_i = 1$  ensures that the probability for qualified individuals to be assigned the favourable outcome is at least as high as without the constraints.

**Reduction Approach.** Agarwal et al. provide an algorithm which allows and fairness definition, as long as it can be formalised via linear inequalities on conditional moments [4]. That is, for fairness notions which can be represented as

$$M_\mu(h) \leq c, \quad (48)$$

where  $M \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{J}|}$  is a matrix and  $c \in \mathbb{R}^{|\mathcal{K}|}$  is a vector describing linear constraints indexed by  $k \in \mathcal{K}$ . Further,  $\mu(h) \in \mathbb{R}^{|\mathcal{J}|}$  is a vector of conditional moments having the form

$$\mu_j(h) = E(g_j(X, Y, Z, h(X)) \mid \mathcal{E}_j) \quad \text{for } j \in \mathcal{J}. \quad (49)$$

As an example, the authors formulate group fairness in terms of Eqs. (48) and (49) with  $\mathcal{J} = Z \times \{*\}$ ,  $\mathcal{E}_z = \{Z = z\}$ ,  $\mathcal{E}_* = \{True\}$ , and  $g_j(X, Y, Z, h(X)) = h(X) \forall j$ , hence  $\mu_j(h) = E(h(X))$ , leading to

$$\begin{aligned} \mu_z(h) - \mu_*(h) &\leq 0 \\ -\mu_z(h) + \mu_*(h) &\leq 0 \end{aligned}$$

and with  $\mathcal{K} = Z \times \{+, -\}$  finally to  $M_{(z,+),z'} = \mathbf{1}\{z' = z\}$ ,  $M_{(z,+),*} = -1$ ,  $M_{(z,-),z'} = -\mathbf{1}\{z' = z\}$ ,  $M_{(z,-),*} = 1$ , and  $c = 0$ .

Let  $\mathcal{H}$  denote a family of classifiers and let  $Q$  be a randomised classifier which makes predictions by first sampling a  $h$  from  $\mathcal{H}$  and then returning  $h(x)$ . The reduction approach [4] interprets the classification problem as saddle point problem by posing an additional  $L_1$  norm constraint onto it and considering also its dual

$$\min_Q \max_{\lambda, \|\lambda\|_1 \leq B} L(Q, \lambda) \quad (50)$$

$$\max_{\lambda, \|\lambda\|_1 \leq B} \min_Q L(Q, \lambda) \quad (51)$$

with  $L(Q, \lambda) = \text{err}(Q) + \lambda^T (M\mu(Q) - c)$  where  $\lambda \in \mathbb{R}_+^{|\mathcal{K}|}$  is a Lagrange multiplier. The problem is solved by the standard scheme of Freund and Schapire [73], which finds an equilibrium in a zero-sum game. Input to the algorithm hereby is the training data  $\mathcal{D}$ , the fairness constraint expressed by  $g_j, \mathcal{E}_j, M, c$ , bound  $B$ , learning rate  $\eta$ , and a minimum accuracy  $\nu$ . If a deterministic classifier is preferred, the found saddle point yields a suitable set of candidates.

## 7.5 Compositional Approaches

Algorithms in this category train a set of models with one classifier per group in  $Z$ . Thus the subgroup accuracy is kept high while the overall classifier achieves fair results.

**Decoupled Classifiers.** Under the assumption that it is ‘legal and ethical’, as the authors put it, Dwork et al. propose the use of decoupled classifiers [59].

Reasoning that a single classifier might lead to too much of an accuracy trade-off over certain groups, the proposal is to use decoupled classification systems. This means to train a separate classifier for each group.

The framework starts by obtaining a set of classifiers  $C_z = \{C_z^{(1)}, \dots, C_z^{(k)}\}$  for each group  $z \in Z$ , in which the  $C_z^{(j)}$  differ in the number of positively classified individuals from the group  $z$ . The decoupled training step outputs a single element of  $C_0 \times \dots \times C_{|Z|}$  ( $C_0 \times C_1$  for a binary protected attribute) yielding one classifier per group by minimising a joint loss function. The joint loss needs to penalise unfairness as well as model the explicit trade-off between accuracy and fairness.

Let the profile of a decoupled classifier denote the vector  $\langle p_1, \dots, p_{|Z|} \rangle$  with  $p_z, z \in Z$  denoting the number of positively classified individuals of group  $z$ . The authors observe that the most accurate classifier for a given profile also minimises false positives and false negatives. Hence, joint loss determines the profile to choose.

Note that as long as the loss is weakly monotone, any off-the-shelf classifier can be used for determining a decoupled solution.

## 7.6 Further Algorithms

Further in-processing techniques include integrating different counterfactual assumptions [163], confidence based approach [72], unfairness penaliser [18], empirical risk minimization under fairness constraints [55], and meta-algorithm for fair classification [43].

## 8 Post-Processing: Correcting Biased Classifiers

Post-processing techniques work by taking a trained classifier which is possibly biased and correcting this bias dependent on the protected attribute. While this is not always legally feasible (cf. Section 1.1) it is shown that subgroup specific thresholding leads to the fairest yet accurate results [49, 138].

We categorised the presented algorithm into

- output correction (Section 8.1),
- input correction (Section 8.2), and
- classifier correction (Section 8.3).

### 8.1 Output correction

Techniques in this category alter the predicted outputs, either by changing the prediction threshold or deterministic labels.

**Discrimination Aware Decision Tree Relabelling** As an alternative to their discrimination aware decision tree construction algorithm, Kamiran et al. further propose the method of decision tree relabelling [110]. Typically, in a decision tree  $T$  the predictions are done via majority votes in the leaves. This approach however flips the labels of certain leaves to achieve a less discriminative prediction.

Consider the following contingency tables for the dataset  $\mathcal{D}$  and a leaf  $l$  respectively, taken from Kamiran et al. [110]:

$\mathcal{D}$	$Y = 0$	$Y = 1$	
$\hat{Y} \rightarrow$	0/1	0/1	
$Z = 1$	$A_0/A_1$	$B_0/B_1$	$z$
$Z = 0$	$C_0/C_1$	$D_0/D_1$	$\bar{z}$
	$N_0/N_1$	$P_0/P_1$	

Leaf $l$	$Y = 0$	$Y = 1$	
$Z = 1$	$a$	$b$	$z$
$Z = 0$	$c$	$d$	$\bar{z}$
	$n$	$p$	

The accuracy and discrimination of the decision tree  $T$  before the label of  $l$  is changed can be calculated by

$$\begin{aligned} \text{acc}_T &= N_0 + P_1 \\ \text{disc}_T &= \frac{C_1 + D_1}{\bar{z}} - \frac{A_2 + B_2}{z} \end{aligned}$$

Let the impact of relabelling leaf  $l$  on accuracy and discrimination be defined as

$$\begin{aligned} \Delta \text{acc}_l &= \begin{cases} n - p & p > n \\ p - n & p < n \end{cases} \\ \Delta \text{disc}_l &= \begin{cases} \frac{a+b}{z} - \frac{c+d}{\bar{z}} & p > n \\ -\frac{a+b}{z} + \frac{c+d}{\bar{z}} & p < n \end{cases} \end{aligned}$$

Note that  $\Delta \text{acc}_l$  is always negative. Given a set of leaves  $L$ , define

$$\text{rem\_disc}(L) = \text{disc}_T + \sum_{l \in L} \Delta \text{disc}_l.$$

The proposed algorithm now aims to find the minimal subset of leaves  $L \subseteq \mathcal{L}$  which need to be relabelled to achieve a discrimination smaller than  $\epsilon \in [0, 1]$ . This is done by defining  $\mathcal{I} = \{l \in \mathcal{L} \mid \Delta \text{disc}_l < 0\}$ , which is the set of all leaves which reduce the discrimination upon relabelling, and then iteratively constructing  $L$ . Hereby  $\arg \max_{l \in \mathcal{I} \setminus L} (\Delta \text{disc}_l) / \Delta \text{acc}_l$  is added to the initially empty set  $L$  as long as  $\text{rem\_disc}(L) > \epsilon$ .

Note that this relabelling problem is equivalent to the NP problem KNAPSACK [110] and actually NP-complete.

**Reject Option Based Classification.** This approach alters the prediction of samples for which  $h(x) = P(Y \mid X)$  is close to the decision boundary, usually 0.5 in binary classification, by introducing a rejection option [111].

For  $\theta$  with  $0.5 < \theta < 1$  let  $[1 - \theta, \theta]$  be the interval further denoted as the critical region. As an example, for  $\theta = 0.7$  the critical region would be  $[0.3, 0.7]$ .

If the prediction score  $h(x)$  lies in the critical region, the actual prediction is dependent on  $Z$ : unprivileged individuals receive the favourable outcome  $\hat{y} = 1$  whereas privileged individuals receive the unfavourable outcome  $\hat{y} = 0$ . For prediction scores outside the critical region, predictions remain unchanged. That is, for  $h(x) \in [0, 1 - \theta[$  the prediction is  $\hat{y} = 0$ , and for  $h(x) \in ]\theta, 1]$  it is  $\hat{y} = 1$ .

This can be interpreted as a cost-based prediction method [111], as the loss for misclassifying a qualified unprivileged individual is  $\theta/(1 - \theta)$  times as high as the loss of misclassifying an unqualified individual.

**Correcting Rule-based Classifiers.** Concerning themselves with association rule learning [5], Pedreschi et al. give a correction framework of classification rules of the form  $r = A, B \rightarrow C$  [152] in rule-based classifiers such as CPAR [200]. This corresponds to  $r = X, Z \rightarrow Y$  in our introduced notation from Section 4, meaning  $Y$  is a consequence of  $(X, Z)$ .

In rule-based classification, the predicted class for an individual  $x, z$  is determined by

$$h(x, z) = \arg \max_{y \in Y} \frac{|\{(X, Y, Z) \in \mathcal{D} \mid X = x, Y = y, Z = z\}|}{|\{(X, Y, Z) \in \mathcal{D} \mid X = x, Z = z\}|},$$

which is the class with the highest confidence given the training data  $\mathcal{D}$ .

Define  $a_{x,z} = |\{(X, Y, Z) \in \mathcal{D}^1 \mid X = x\}|$  as the number of qualified individuals with  $X = x, Z = z$ , and  $n_{x,z} = |\{(X, Y, Z) \in \mathcal{D}^1 \mid X = x\}|$  as the number of qualified individuals with  $X = x$ , disregarding their protected attribute. The confidence for the rule  $X, Z \rightarrow Y$  is then given by  $p_{x,z \rightarrow y}$  with

$$\begin{aligned} p_{x,z \rightarrow 1} &= a_{x,z} / n_{x,z}, \\ p_{x,z \rightarrow 0} &= 1 - p_{x,z \rightarrow 1}. \end{aligned} \tag{52}$$

Let  $f$  denote an approximative fairness measure and let  $a \in \mathbb{R}$  be a fixed threshold. The classification rule  $r = X, Z \rightarrow Y$  is said to be  $a$ -discriminatory with respect to  $f$  [151, 152] if  $f(r) \geq a$ .

To rectify an  $a$ -discriminatory rule-based classifier, the authors adapt the confidences  $p_{x,0 \rightarrow y}$  from Eq. (52) to

$$\begin{aligned} p'_{x,0 \rightarrow 1} &= (a_{x,0} + \Delta) / n_{x,0}, \\ p'_{x,0 \rightarrow 0} &= 1 - p'_{x,0 \rightarrow 1} \end{aligned} \tag{53}$$

with  $\Delta \in \mathbb{Z}$  such that  $|\Delta|$  is the minimum integer resulting in  $f(r) < a$ .

Note that this is only for correcting direct discrimination [151, 152]. However, the authors further give a correction method for indirect discrimination [151, 152], which we omit here.

**Plugin Approach.** Menon and Williamson propose a plugin approach for correcting outputs of a classifier by thresholding of the class probability function on an instance-basis [138]. This approach assumes the classification problem to be cost-sensitive.

Two logistic regression models are trained,  $\eta_A, \eta_F$ . Hereby,  $\eta_Y$  is trained on  $X, Y$  and estimates the probability of  $x \in X$  to be qualified for the favourable outcome  $\eta_A(x) = P(Y = 1 \mid X = x)$ , i.e.  $\eta_A = h$ . The second model,  $\eta_F(x)$  estimates  $P(Z = 1 \mid X = x)$ , i.e. the probability an individual belongs to the privileged group. Note that his definition of  $\eta_F$  is meant to achieve group fairness. If  $\eta'_F(x) = P(Z = 1 \mid X = x, Y = 1)$  is estimated instead, equality of opportunity is the target fairness instead.

Given two cost parameters  $c_A, c_F$ , define  $s : x \mapsto \eta_A(x) - c_A - \lambda(\eta_F(x) - c_F)$  with the tradeoff parameter  $\lambda$ .

The final classification happens by  $h(x) = H_\alpha(s(x))$ , where  $H_\alpha$  for  $\alpha \in [0, 1]$  is the modified Heaviside step function  $H_\alpha(s) = \mathbf{1}_{\{s > 0\}} + \alpha \mathbf{1}_{\{s = 0\}}$ .

**Discrimination-Aware Ensemble.** Another approach presented by Kamiran et al. is that of the discrimination-aware ensemble [111]. Here, an ensemble of classifiers  $h^{(1)}, \dots, h^{(k)}$  is corrected in their predictions by altering decisions made in the disagreement region.

Let  $\hat{y}^{(i)}$  denote the prediction made by the  $i$ -th classifier  $h^{(i)}$ . If all classifiers agree in their predictions, i.e.  $\hat{y}^{(i)} = \hat{y}^{(j)} \forall i, j \in \{1, \dots, k\}$ , then this unanimous prediction is used as final prediction  $\hat{y}$ . If otherwise at least one classifier disagrees, i.e.  $\exists i, j \in \{1, \dots, k\} \cdot \hat{y}^{(i)} \neq \hat{y}^{(j)}$ , then the prediction is made as in the case of the critical region for reject option based classification [111]: unprivileged individuals receive the favourable outcome  $\hat{y} = 1$  whereas privileged individuals receive the unfavourable outcome  $\hat{y} = 0$ .

The disagreement of the ensemble over a data set  $\mathcal{D}$  can be measured by

$$\text{disag}(\mathcal{D}) = \frac{|\{(X, Y, Z) \in \mathcal{D} \mid \exists i, j \cdot \hat{y}^{(i)} \neq \hat{y}^{(j)}\}|}{|\mathcal{D}|}.$$

The authors remark that the accuracy drops as the disagreement increases. The discrimination of the ensemble can ultimately be controlled by the selection of its member classifiers [111].

## 8.2 Input correction

These approaches are related to pre-processing from Section 6, but add a pre-processing layer in front of an already trained algorithm.

**Gradient Feature Auditing.** The gradient feature auditing method proposed by Adler et al. [3] works by obscuring the (indirect) influences of  $Z$  on  $X^{(i)}$  and finding a minimal perturbation of  $X$  over which the classifier  $h$  yields more fair predictions.

The  $\epsilon$ -obscure version of  $X$  with respect to feature  $i$  is denoted as  $X \setminus_{\epsilon} X_i$ , if  $X^{(i)}$  cannot be predicted from  $X \setminus_{\epsilon} X_i$ . That is

$$\text{BER}(X \setminus_{\epsilon} X_i, X^{(i)}, f) > \epsilon \quad \forall f : X \setminus X_{(i)} \rightarrow X^{(i)} \quad (54)$$

where BER denotes the balanced error rate of  $f$  (a.k.a. half total error rate [175]). By considering the difference in accuracy of on  $h(X)$  and  $h(X \setminus_{\epsilon} X_i)$ , the (indirect) influence of feature  $i$  is measured.

The features are ordered by their (indirect) influences. To remove a feature  $i$  (e.g.  $Z$ ) from  $X$ , compute  $X \setminus_{\epsilon} X_i$  by applying the following obscuring procedure to each feature  $j \neq i$ . For a numerical feature  $W = X^{(j)}$  let  $W_x = P(W \mid X^{(i)} = x)$  and  $F_x(w) = P(W \geq w \mid X^{(i)} = x)$  denote the marginal distribution and cumulative distribution conditioned on  $X^{(i)} = i$ . Consider the median distribution  $A$ , which has its cumulative distribution  $F_A$  given by  $F_A^{-1}(u) = \text{median}_{x \in X^{(i)}} F_x^{-1}(u)$ . It was already shown for the disparate impact remover that a distribution  $\tilde{W}$  which is minimally changed to mimic the distribution of  $A$  maximally obscures  $X^{(j)}$  [70].

### 8.3 Classifier correction

Algorithms in this category take a predictor and construct a related predictor from it which yields fairer decisions.

**Derived Predictor.** Alongside proposing the notions of equality of opportunity and equalised odds (Eqs. (8) and (9)), Hardt et al. presented a framework for achieving those notions [97]. Given a possibly unfair predictor  $\hat{Y}$ , the goal is to derive a predictor  $\tilde{Y}$  which satisfies the respective fairness notion. This can be calculated by only considering the joint distribution  $\hat{Y}, Z, Y$ .

For  $z \in Z$ , let

$$\gamma_z(\hat{Y}) = (P_z(\hat{Y} = 0 \mid Y = 0), P_z(\hat{Y} = 1 \mid Y = 1)) \quad (55)$$

denote the tuple of false and true positive rates for  $\hat{Y}$  and define the two-dimensional convex polytope

$$\mathcal{P}_z = \text{convhull}\{(0, 0), \gamma_z(\hat{Y}), \gamma_z(1 - \hat{Y}), (1, 1)\} \quad (56)$$

where  $\text{convhull}$  denotes the convex hull of the given set. Given a loss function  $L$ ,  $\tilde{Y}$  can be derived via the optimisation problem

$$\begin{aligned} \min_{\tilde{Y}} \quad & E(L(\tilde{Y}, Y)) \\ \text{s.t.} \quad & \gamma_z(\tilde{Y}) \in \mathcal{P}_z(\hat{Y}) \quad \forall z \in Z, \\ & \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}). \end{aligned} \quad (57)$$

This ensures  $\tilde{Y}$  to minimise loss whilst satisfying equalised odds. For achieving equality of opportunity, the second condition weakens to only require equal true positive rates over the groups [97].

**Calibration via Information Withholding.** Pleiss et al. propose an algorithm for achieving a relaxed version of equalised odds and calibration [153] by adjusting the predictions of a classifier  $h$  made over the unprivileged group.

Let  $h$  be split into two classifiers  $h_0, h_1$  with  $h_z$  giving predictions over the group with protected attribute  $Z = z$ .

Let  $g_z(h_z)$  define a cost function over the false positive and false negative rates of  $h_z$

$$g_z(h_z) = a_z E(h_z(X) | Y = 0) + b_z E(1 - h_z(X) | Y = 1)$$

with  $a_z, b_z \geq 1$  being group dependent constants of which at least one is non-zero. We say a classifier  $h = (h_0, h_1)$  satisfies *relaxed* equalised odds with calibration if  $g_0(h_0) = g_1(h_1)$ . Hereby the notion of equalised odds is relaxed such that it no longer contradicts the notion of calibration.

Assume  $h_0, h_1$  with  $g_1(h_1) < g_0(h_0) < g_1(h^{\mu_1})$  and a randomly sampled subset  $\mathcal{E}_1 \subseteq \mathcal{D}_1$ . Let  $h^{\mu_1}(x) = |\{(X, Y, Z) \in \mathcal{E}_1 | Y = 1\}| / |\mathcal{E}_1|$  denote the trivial classifier which always returns the prediction score equal to the base positive rate of  $\mathcal{E}_1$ . Define  $\alpha = \frac{g_0(h_0) - g_1(h_1)}{g_1(h^{\mu_1}) - g_1(h_1)}$ .

The corrected classification can now be constructed by setting

$$\tilde{h}_1(x) = \begin{cases} h^{\mu_1}(x) & \text{with probability } \alpha \\ h_1(x) & \text{with probability } 1 - \alpha \end{cases} \quad (58)$$

and hence achieving a calibrated classifier  $h_0, \tilde{h}_1$  with  $g_0(h_0) = g_1(\tilde{h}_1)$ .

The authors remark that, as the output of  $\tilde{h}_1$  is achieved by withholding information over a randomly selected subset, the outcome becomes inequitable within the group [153]. Due to the algorithm being optimal however, any other algorithm would at least yield as many false positives or negatives as  $h_0, h_1$  does.

## 8.4 Further Algorithms

Further post-processing algorithms include naïve bayes correction [39], learning non-discriminatory predictors [198], fair causal reasoning [130], counterfactually fair deconvolution [122], and deep weighted averaging classifiers [42].

## 9 Fairness Toolkits

Given the recent popularity of fairness-aware machine learning, it is no surprise that there exists a number of tools which help developers to achieve and ensure non-discriminatory systems. The most commonly used datasets in literature are listed in Section 9.1. In Section 9.2 we will then list a number of tools and frameworks.



## 9.1 Datasets

**ProPublica Recidivism.** The ProPublica recidivism data set<sup>6</sup> includes data from the COMPAS risk assessment tool and was analysed by Angwin et al. [7] to show COMPAS’ race bias. It contains 7214 individuals and encodes the sensitive attributes race, sex, and age. The prediction cast outcome is whether an individual was rearrested within two years of the first arrest.

A *violent recidivism* version exists where the outcome is a rearrest within two years on basis of a violent crime.

**German Credit.** The German Credit Data [57] consist of 1000 samples spanning 20 features each. Containing features such as employment time, current credits, or marital status it provides the prediction task to determine whether an individual has good or bad credit risk. Sensitive attributes are sex and age.

**Adult Income.** The Adult Data (a.k.a. Census Income Dataset) [57] consists of 48842 samples. It’s 14 features include information such as relationship status, education level, or occupation, as well as the sensitive attributes race, sex, and age. The associated prediction task is to classify whether an individual’s income exceeds \$50,000 annually.

**Dutch Virtual Census.** The Dutch Virtual Census data are released by the Dutch Central Bureau for Statistics [44, 45] and consist of two sets, one from 2001,<sup>7</sup> the other from 1971.<sup>8</sup> The data contains 189,725 and 159,203 samples respectively with the classification objective whether an individual has a prestigious occupation or not, providing the sensitive attribute sex.

As Kamiran et al. [110] pointed out, these two sets are unique in a way that the sex discrimination has decreased from 1971 to 2001 as seen in the data. Hence, a classifier can be trained on the 2001 data and then be evaluated on the discriminatory 1971 data.

## 9.2 Frameworks

**FairML.** In his master’s thesis, Adebayo developed FairML [1, 2], an end-to-end toolbox for quantifying the relative significance of the feature dimensions of a given model. FairML is written in Python and available on GitHub.<sup>9</sup>

It employs four different ranking algorithms with which the final combined scores for each feature are determined. Those algorithms are the iterative orthogonal feature projection algorithm [1], minimum redundancy maximum relevance

<sup>6</sup> <https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>

<sup>7</sup> <https://microdata.worldbank.org/index.php/catalog/2102>

<sup>8</sup> <https://microdata.worldbank.org/index.php/catalog/2101>

<sup>9</sup> <https://github.com/adebayoj/fairml>

feature selection [54], the lasso regression algorithm [180], and the random forest feature selection algorithm [34, 132].

Having the relative significance of each feature eases reasoning about the problem domain and validating potential discrimination of the audited system.

**FairTest.** FairTest is an implementation of the unwarranted associations framework [181] which detects indirect discriminations present in an algorithm. It is written in Python and available on GitHub.<sup>10</sup>

The framework identifies sets of features and resulting predictions of interest, integrates further user or application requirements which might justify certain kinds of indirect discrimination, estimates strength of unfairness over an appropriate metric, and tests for indirect discrimination over meaningful subpopulations. The authors emphasise the repeatability of those steps for fairness-aware debugging.

**Themis.** Themis [77] was developed by the Laboratory for Advanced Software Engineering Research at the University of Massachusetts Amherst and is a tool for testing software for discrimination. Themis is written in Python and available on GitHub.<sup>11</sup>

The tool can be used in three different ways. Firstly, it can be used for generating a test suite to compute the discrimination scores for a set of features. Secondly, it can compute all feature subsets against which the software discriminates more than a specified threshold. Thirdly, given a test suite, the apparent discrimination scores for a feature set are calculated. The discrimination scores hereby consist of group discrimination scores, as well as causal discrimination scores.

**Themis-ML.** Besides the similar names, Themis-ML [12] is unrelated to the previously presented Themis testing tool. Themis-ML titles itself as fairness-aware machine learning interface, giving access to a number fairness metrics and algorithms from literature. It is written in Python and available on GitHub.<sup>12</sup>

The set of implemented measures consists of group fairness and normalised difference. Further it provides the massaging [38] pre-processing technique, the counterfactually fair deconvolution [122] in-processing algorithm, as well as the post-processing of reject option based classification [111]. Additionally, it incorporates access to the German Credit as well as the Census Income Data sets.

**Fairness Measures.** Fairness Measures [206] is a tool for evaluating a dataset on a selected fairness measure. It is implemented in Python, yet serves as a command line tool expecting a CSV file and can be used language-independently

<sup>10</sup> <https://github.com/columbia/fairtest>

<sup>11</sup> <https://github.com/LASER-UMASS/Themis>

<sup>12</sup> <https://github.com/cosmicboy/themis-ml>

for any project. It is available over <http://www.fairness-measures.org> and also on GitHub.<sup>13</sup>

The implemented measures are taken from Žliobaitė’s 2017 survey [212] and comprise statistical tests, absolute measures, conditional measures, and situation measures.

**Fairness Comparison.** Friedler et al. provide with their Fairness Comparison [75] a strong benchmarking tool for different fairness-aware algorithms. It is implemented in Python and available on GitHub.<sup>14</sup>

The benchmarker works by using a pre-processed version of a given dataset. This does not already correspond to a pre-processing for achieving algorithmic fairness, but rather ensures that most algorithms in the benchmark suit can handle the input. On the pre-processed data the actual algorithms are run and finally evaluated based on both, accuracy measures and fairness measures.

**Aequitas.** The Center for Data Science and Public Policy of the University of Chicago published Aequitas [165], a bias and fairness audit toolkit and produces detailed bias reports over the input data. It is usable as a Python library or as a standalone command line utility and available on GitHub.<sup>15</sup>

The tool checks for six different fairness notions: group fairness, disparate impact, predictive parity, predictive equality, false omission rate parity, and equality of opportunity. The authors further provided a *fairness tree*, guiding the choice of which notion to use via a decision tree. Given a set of notions to check for, the tool outputs a bias report which captures which notions were violated and to what extent on a per-subgroup base.

**AIF360.** IBM Research released AI Fairness 360 (AIF360) [20], an extensible Python toolkit comprising multiple fairness-aware algorithms known from literature. It is available on GitHub<sup>16</sup> and provides a website with interactive tutorials and a gentle introduction to its concepts and capabilities.<sup>17</sup>

The algorithms it comprises are optimized pre-processing [41], the disparate impact remover [70], the derived predictor [97], reweighing [109], reject option based classification [111], the prejudice remover regularizer [113], calibration via information withholding [153], learning fair representations [208] adversarial debiasing [209], and the meta-algorithm for fair classification [43]. Further, it contains more than 70 different metrics for individual fairness (i.e. euclidean distance, mahalanobis distance), group fairness (i.e. equal opportunity difference, disparate impact), and also general classification measurements (i.e. true/false positive/negative rates, precision, recall) which allow for further construction of

<sup>13</sup> [https://github.com/megantosh/fairness\\_measures\\_code](https://github.com/megantosh/fairness_measures_code)

<sup>14</sup> <https://github.com/algofairness/fairness-comparison>

<sup>15</sup> <https://github.com/dssg/aequitas>

<sup>16</sup> <https://github.com/ibm/aif360>

<sup>17</sup> <https://aif360.mybluemix.net>

so far unimplemented fairness notions. Finally, it provides easy access to prominent data sets, such as the Adult Data, German Credit Data, or the ProPublica recidivism data set.

**FairSearch.** Zehlike et al. [207] published FairSearch, a tool for fairness in ranked search results. It implements two previously published algorithms, FA\*IR [204] and DELTR [205], and is available as Python or Java library, as well as an Elasticsearch plugin.<sup>18</sup>

This tool is concerned with the fairness of exposure in rankings [169]. As users of such systems usually only glance at the first few results, an average group exposure is needed for a fairer ranking. This can be achieved either by FA\*IR, a reranking algorithm, or DELTR, which is a learn-to-rank framework. Both algorithms aim to optimise rankings subject to fairness criteria.

**FAT Forensics.** FAT Forensics [171] is a Python toolbox for inspecting fairness, accountability, and transparency of all aspects of a machine learning system. It was started as an academic collaboration between the University of Bristol and the Thales Group, and is available on GitHub.<sup>19</sup>

FAT Forensics works throughout the whole machine learning pipeline, consisting of data, models, and predictions. Besides only concerning itself with fairness, it further implements algorithms regarding accountability and transparency as well, giving reports on, i.e., neighbour-based density estimations regarding prediction robustness, inspection of counterfactual fairness, or finding explanations for blackbox models via local interpretable model-agnostic explanations [158].

## 10 Further Discussion and Final Remarks

In this survey, we surveyed a multitude of fairness notions as well as algorithms to achieve bias-free prediction and classification. However, we considered but a part of the corresponding literature and by no means claim to deliver an exhaustive overview. By only considering the fairness-aware machine learning literature, we knowingly left out three other, related aspects: accountability, transparency, ethics.

In this section, we will point to existing literature concerned with those three aspects and finally will conclude with some final remarks about fairness-aware machine learning.

### 10.1 Accountability and Transparency

While avoiding discrimination is the overall goal of the field, there are other important aspects to consider. The legal texts mentioned in Section 1.1 for example

<sup>18</sup> <https://github.com/fair-search>

<sup>19</sup> <https://github.com/fat-forensics/fat-forensics>

are also concerned with the question of accountability. If employers of discriminatory systems can be held (legally) accountable, accountability serves as an important tool towards ensuring a more responsible use of machine learning, which ultimately leads already to an improvement of society [65,145]. On a similar note, Taylor [174] makes a case for requiring *data justice*, based on the three pillars of (in)visibility, (dis)engagement with technology, and anti-discrimination.

Transparency is an important factor for individuals affected by decisions of an AI system. Transparency of a decision making process allows a user to actually inspect and understand why a specific decision was made, which ultimately also allows the developers of such systems to more easily check for discrimination. This feature can also be a legal requirement. The European Union states in the General Data Protection Regulation [69] (effective since 2018) that “[the user] should have the right [...] to obtain an explanation of the decision reached after such assessment [...]”, hence stating a *right to explanation* for the affected users. In the U.S.A., the Equal Credit Opportunity Act [191] states the same right for the credit system. Creditors must provide specific reasoning to the applicants for why a credit action was taken. The European Union identifies accountability and transparency as tools to reach fairer systems [65].

Unfortunately, in current practice many of the popular machine learning algorithms such as neural networks [84] or random forests [34], provide only blackboxes. This is on one hand a transparency issue, as the concrete process of how the system derived a decision is unknown. On the other hand this also impacts accountability, as it is harder to prove true discrimination inside an algorithm (although statistical measures as listed in Section 5 can still be tested for). There was and is a lot of research conducted concerned with symbolic learning for interpretability such as neural-symbolic integration [11, 79, 95] or explainable AI systems which can deliver an explanation for their decision process [56, 90, 100, 139, 158, 196].

Another addition to this topic is recourse by Ustun et al. [192], which denotes the ability of an individual to change the output of a given, fixed classifier by influencing the inputs. This can be understood as the ability of an individual to adapt it’s profile (i.e., its feature vector) in such a way, that the favourable outcome is guaranteed. While this does not imply transparency or accountability and vice-versa the authors emphasise that this gives some kind of agency to the individuals which hence could increase the perceived fairness of the algorithms [31]

## 10.2 Ethics and the Use of Protected Attributes

While AI offers a broad range of different ethical problems (e.g. AI warfare [88], AI and robot ethics [37, 89, 128], harms via unethical use [183], or accountability of autonomous systems [63]), we restrict this section only to the ethical questions regarding application of machine learning for decision processes over individuals.

Goodall [83] discusses whether an autonomous vehicle can be programmed with a set of ethics to act by. As an example they consider a vehicle having to decide whether to hit a pedestrian or swerve around and hence cause other

kind of property damage. The vehicle’s system could factor in the estimated financial cost of hitting the pedestrian and weight against the estimated property damage. Assuming the vehicle does the first estimate over historical data of civil settlements in the respective neighbourhood, this could lead to a higher crash risk of pedestrians in poorer areas, where lower settlements might be more common. The vehicle would then discriminate against social status. Keeping in mind that the neighbourhood may be correlated with race, this can also lead to racial discrimination.

In “The Authority of ‘Fair’ in Machine Learning” [170], Skirpan and Gorelick discuss whether the employment of decision making systems is fair in the first place and propose three guiding questions for developers to take into account, and hence to take more responsibility for the implementation. The proposed questions are whether it is fair to implement the decision system in the first place, whether a fair technical approach exists, and whether it provides fair results after it has been implemented. While the latter two questions correspond to methods covered in Sections 5 to 8, the first question poses another problem not yet covered: how do we determine whether the employment of an automated decision system is fair in the first place?

This is related to the feature selection for the task at hand: are the features upon which the decision shall be conducted a fair set? The answer to this begins with whether protected attributes should be considered as features or not, which still remains a topic of open discussion. As already pointed out, inclusion of the protected attributes into the feature set might not always be legally feasible. A notable exception from the law is business necessity [15], i.e. an employer can prove that a decision over a protected attribute is actually justified in context of the business.

Dwork et al. [59] argue that consideration of the protected attribute “may increase accuracy for all groups and may avoid biases”, yet constrain this to cases in which inclusion is “legal and ethical”. This is in line with the results of Berk et al. [23], which reported better overall results by taking race into account in their experiments. Žliobaitė [212] argued that a model which considers the protected attribute would not treat individuals with otherwise identical attributes similarly and hence would practice direct discrimination. However, he still emphasises that utilising the protected attribute still aids in enforcing non-discrimination.

A generalisation of this to the fairness of the whole feature selection problem is the notion of process fairness proposed by Grgić-Hlača et al. [87], where the users are given some kind of agency over the employed features. Proposed are three notions of process fairness: feature-apriori fairness, measuring the percentage of users which assume a given feature to be fair without further knowledge, feature-accuracy fairness, measuring the percentage of users which assume a given feature fair given the information that it increases accuracy, and feature-disparity fairness, considering the percentage of users which assume a feature to be fair even if it creates disparity in the prediction. Employing a study to find a process-fair feature set before the predictor is implemented might lead to predictors which are perceived as more fair in the eyes of affected individuals.

Process fairness allows for affected individuals to decide themselves whether the protected attribute (any attribute, really) is unfair or not.

Overall, ethics questions do not only concern the machine learning community but software engineering as a whole. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems published their vision for an ethically aligned design [176] in which they discuss general principles and methods towards an ethically responsible use of AI systems for increasing human well-being. While this includes topics of fairness, accountability, and transparency, the article spans over topics such as human rights, data agency, or awareness of misuse as well. Aydemir and Dalpiaz [9] proposed a set of seven research questions to formulate a roadmap towards ethics-aware software engineering. The proposed research includes characterising relevant ethics issues for software development and corresponding notations and models, as well as determining analysis and verification approaches to show that software keeps to imposed ethics specifications.

### 10.3 Critiques and Perspectives on Current Practices

As already indicated above, one of the main concerns for training a fairness-aware classifier is the data it is trained on. Friedler et al. [75] have shown in their study that the eventual fairness of the trained systems is strongly dependent on the initial data split, as demonstrated by their cross-validation approach. This indicates that fairness methods “[are] more brittle than previously thought”, as stated by them. In their conclusion, they provide three recommendations of how to approach future contributions to the fairness research to increase quality and close possible gaps. These recommendations are: emphasising pre-processing requirements, i.e. providing multiple performance metrics if the training data can be processed in different ways, avoiding proliferation of measures, i.e. new notions should be introduced only if fundamentally different than existing ones, and accounting for training instability, i.e. providing performances on multiple training splits. Hence, they provide a means for more robust result replication for readers of such papers, as well as a more unified ground of comparison of provided performances between different papers.

A similar critique was given by Gebru et al. [80], who criticise the missing documentation of datasets used in machine learning. They propose a unified approach to give crucial information, answering questions such as why a dataset was created, by whom it was funded, whether pre-processing took place, or whether affected individuals knew about assembly of their data into the set. Ambition is also to have a protocol for potentially protected attributes or correlations a dataset user should be aware of.

Žliobaitė [211] points out the need of a unified theory on how to assess fairness, as it is hard to argue whether a system is fair or not, as it depends on the choice of fairness conditions which partially contradict each other. A further problem he considers is that most solutions are targeted at a specific fairness notion tied to a specific set of machine learning algorithms. He thus proposes three research tasks which build upon one another: consolidated algorithmic fairness measures, empirical and theoretical analysis of how an algorithm can

become discriminatory, and fairness-aware optimisation criteria. Another point he emphasises is the need of interdisciplinary research. The machine learning community’s primary goal is not to decide which attributes are to be predicted but rather to implement a solution for non-discriminatory decision procedures *given* a set of already determined predicted attributes. For the decision on protected attributes and in which areas it is to be employed, the expertise of the law and social sciences is strictly necessary [211].

Similarly, Barocas et al. [13] list five areas of concern for a fairness-aware research agenda: determining whether a model exhibits objectionable bias, increasing awareness of the subject matter, improvements to transparency and control of algorithmic decision procedures, studying fundamental sources of unfairness in the whole machine learning pipeline, and supporting interdisciplinary scholarship.

Regarding the machine learning part, it appears that the common notion will converge towards a unified process framework for fairness-aware predictors, which enhances the common machine learning pipeline by further pre-processing and evaluation steps. This presumably would span the monitoring and justification of the initial data collection techniques, study-driven feature development, application of a fairness-aware algorithm in the sense of Sections 6 to 8 and rigorous evaluation thereof, a (preferably externally conducted) auditing of the processes for achieving the predictor, a formal justification as of why these systems can be assumed to act in a non-discriminatory manner, as well as an accompanied documentation of the whole process (which is preferably open to the public). A formal definition of such a framework is of course non-trivial and under our current understanding certain steps are even application dependent. For instance, how much agency should be given to affected individuals during feature development? While giving more agency to the people over how they will finally be treated is overall a noble and desirable goal, the question arises how fair or ethical this level of agency is for the overall population. How much agency should a criminal have regarding his recidivism scores? How much agency should an applicant hold over how credit actions are determined for them?

**The Need for Proof.** We conclude this survey by pointing into another direction, which should be considered in future research. While most fairness evaluations are based on the performance on a designated test-set, the problem arises whether the test-set is a good representation of the eventual individuals over which the predictions are conducted. If not, there are no strong guarantees whether the system will act reliably fair in real-world scenarios. A formal proof of system-fairness would need to be conducted to show real fairness of a predictor, independent of training and test set.

McNamara et al. [136] presented an approach of provably fair representation learning. However, while being a big step into the right direction and proving that the inferred representation function indeed provides fair results, this proves are still restricted on probability guarantees over the test set.



Tramèr et al. [181] referred to unfairness in predictors as *fairness bugs* or *association bugs*. In a sense, labelling discriminatory misbehaviour as bugs captures the problem quite well. It further allows us to look into other computer science areas where the absence of bugs, i.e. the absence of programmatic misbehaviour, is crucial: safety-critical systems. Here, programs need to be rigorously proven in a formal, mathematical manner, before they are put into production. A means to do this are formal methods [78, 197].

Proof of machine learning algorithms (with a focus of guaranteed safety) is part of recent and current research. For instance, it is known that image detection neural networks can be manipulated in their output by simply changing certain pixels in the input, unnoticeable for the human eye [140]. As reaction, the community started to develop proof techniques to verify that the expectable input space of neural networks is safe from such adversarial perturbations [103, 115, 123, 199]. Proof-carrying code [71, 94, 144] is a mechanism in which a piece of software is bundled with a formal proof which can be redone by the host system for verification purposes. This could be used for the aforementioned unified framework so that the prediction model ultimately is always accompanied with its formal proof of non-discrimination. A similar idea was presented by Ramadan et al. [156], who outlined a UML-based workflow which allows for automated discrimination analysis.

If we relate discrimination to software bugs and fairness to software safety, the intersection of the formal methods community and the fairness community could actually give field to novel perspectives, algorithms, and applications which ultimately can benefit not only both research groups, but also the individuals affected by a more and more digital world, which we shape together to be safer and fairer for everyone.

## References

1. Adebayo, J., Kagal, L.: Iterative orthogonal feature projection for diagnosing bias in black-box models. arXiv preprint arXiv:1611.04967 (2016)
2. Adebayo, J.A.: FairML: ToolBox for diagnosing bias in predictive modeling. Master's thesis, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, United States (2016)
3. Adler, P., Falk, C., Friedler, S.A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., Venkatasubramanian, S.: Auditing black-box models for indirect influence. *Knowledge and Information Systems* **54**(1), 95–122 (2018)
4. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) *Proceedings of the 35th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 80, pp. 60–69. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018)
5. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. *VLDB '94*, vol. 1215, pp. 487–499. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)

6. Ahlman, L.C., Kurtz, E.M.: The APPD randomized controlled trial in low risk supervision: The effects on low risk supervision on rearrest. Philadelphia Adult Probation and Parole Department (Oct 2008)
7. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. (May 2016)
8. Arrow, K.: The theory of discrimination. *Discrimination in Labor Markets* **3**(10), 3–33 (1973)
9. Aydemir, F.B., Dalpiaz, F.: A roadmap for ethics-aware software engineering. In: *Proceedings of the International Workshop on Software Fairness*. pp. 15–21. ACM (2018)
10. Ayres, I.: Outcome tests of racial disparities in police practices. *Justice Research and policy* **4**(1-2), 131–142 (2002)
11. Bader, S., Hitzler, P.: Dimensions of neural-symbolic integration – a structured survey. arXiv preprint cs/0511042 (2005)
12. Bantilan, N.: Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. *Journal of Technology in Human Services* **36**(1), 15–30 (2018)
13. Barocas, S., Bradley, E., Honavar, V., Provost, F.: Big data, data science, and civil rights. arXiv preprint arXiv:1706.03102 (2017)
14. Barocas, S., Hardt, M., Narayanan, A.: *Fairness and Machine Learning*. fairml-book.org (2019), <http://www.fairmlbook.org>
15. Barocas, S., Selbst, A.D.: Big data's disparate impact. *Calif. L. Rev.* **104**, 671 (2016)
16. Barrio, E.D., Fabrice, G., Gordaliza, P., Loubes, J.M.: Obtaining fairness using optimal transport theory. In: Chaudhuri, K., Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 97, pp. 2357–2365. PMLR, Long Beach, California, USA (09–15 Jun 2019)
17. Barth, J.R., Cordes, J.J., Yezer, A.M.: Financial institution regulations, redlining and mortgage markets. *The regulation of financial institutions* **21**, 101–143 (1979)
18. Bechavod, Y., Ligett, K.: Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044 (2017)
19. Becker, G.S., et al.: *The economics of discrimination*. University of Chicago Press Economics Books (1957)
20. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. arXiv preprint arXiv:1810.01943 (2018)
21. Bendick, M.: Situation testing for employment discrimination in the United States of America. *Horizons stratégiques* (3), 17–39 (2007)
22. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35**(8), 1798–1828 (2013)
23. Berk, R.: The role of race in forecasts of violent crime. *Race and Social Problems* **1**(4), 231 (Nov 2009)
24. Berk, R.: *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. Springer Science & Business Media (2012)
25. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018)

26. Berk, R., Sherman, L., Barnes, G., Kurtz, E., Ahlman, L.: Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(1), 191–211 (2009)
27. Berkovec, J.A., Canner, G.B., Gabriel, S.A., Hannan, T.H.: Race, redlining, and residential mortgage loan performance. *The Journal of Real Estate Finance and Economics* **9**(3), 263–294 (1994)
28. Berliant, M., Thomson, W., Dunz, K.: On the fair division of a heterogeneous commodity. *Journal of Mathematical Economics* **21**(3), 201–216 (1992)
29. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. In: *Proceedings of 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning. FAT/ML* (2017)
30. Biddle, D.: *Adverse Impact and Test Validation*. Gower Publishing, Ltd., 2 edn. (2006)
31. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: ‘it’s reducing a human being to a percentage’; perceptions of justice in algorithmic decisions. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. p. 377. ACM (2018)
32. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: *Advances in neural information processing systems*. pp. 4349–4357 (2016)
33. Breiman, L., Friedman, J., Olshen, R., Stone, C.: *Classification and regression trees* (1984)
34. Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (Oct 2001)
35. Brescia, R.H.: Subprime communities: Reverse redlining, the fair housing act and emerging issues in litigation regarding the subprime mortgage crisis. *Albany Government Law Review* **2**, 164 (2009)
36. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning*, vol. 81, pp. 77–91. PMLR, New York, NY, USA (2018)
37. Burton, E., Goldsmith, J., Mattei, N.: Teaching AI ethics using science fiction. In: *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015)
38. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: *2009 IEEE International Conference on Data Mining Workshops*. pp. 13–18. IEEE (2009)
39. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* **21**(2), 277–292 (2010)
40. Calders, T., Žliobaitė, I.: Why unbiased computational processes can lead to discriminative decision procedures. In: *Discrimination and Privacy in the Information Society*, pp. 43–57. Springer (2013)
41. Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: *Advances in Neural Information Processing Systems*. pp. 3992–4001 (2017)
42. Card, D., Zhang, M., Smith, N.A.: Deep weighted averaging classifiers. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. pp. 369–378. ACM (2019)

43. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 319–328. ACM (2019)
44. Centraal Bureau voor de Statistiek: Volkstelling (1971)
45. Centraal Bureau voor de Statistiek: Volkstelling (2001)
46. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **5**(2), 153–163 (2017)
47. Chouldechova, A., G’Sell, M.: Fairer and more accurate, but for whom? arXiv preprint arXiv:1707.00046 (2017)
48. Citron, D.K., Pasquale, F.: The scored society: Due process for automated predictions. *Wash. L. Rev.* **89**, 1 (2014)
49. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 797–806. ACM (Aug 2017)
50. Cramer, J.S.: The origins of logistic regression (2002)
51. Danner, M.J., VanNostrand, M., Spruance, L.: Risk-based pretrial release recommendation and supervision guidelines (Aug 2015)
52. Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies* **2015**(1), 92–112 (2015)
53. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22 (1977)
54. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology* **3**(2), 185–205 (2005)
55. Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J.S., Pontil, M.: Empirical risk minimization under fairness constraints. In: Advances in Neural Information Processing Systems. pp. 2791–2801 (2018)
56. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? a new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (Oct 2017)
57. Dua, D., Graff, C.: UCI machine learning repository (2017)
58. Duivesteijn, W., Feelders, A.: Nearest neighbour classification with monotonicity constraints. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 301–316. Springer (2008)
59. Dwork, C., Immorlica, N., Tauman Kalai, A., Leiserson, M.: Decoupled classifiers for fair and efficient machine learning. arXiv e-prints (Jul 2017)
60. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226. ACM (2012)
61. Edwards, H., Storkey, A.J.: Censoring representations with an adversary. *CoRR abs/1511.05897* (2015)
62. Eisenhauer, E.: In poor health: Supermarket redlining and urban nutrition. *GeoJournal* **53**(2), 125–133 (Feb 2001)
63. Etzioni, A., Etzioni, O.: AI assisted ethics. *Ethics and Information Technology* **18**(2), 149–156 (Jun 2016)
64. European Parliament: Legislative resolution of 2 april 2009 on the proposal for a council directive on implementing the principle of equal treatment between

- persons irrespective of religion or belief, disability, age or sexual orientation (com(2008)0426 – c6-0291/2008 – 2008/0140(cns)), 2008/0140(APP)
65. European Parliamentary Research Service, Panel for the Future of Science and Technology: A governance framework for algorithmic accountability and transparency (Apr 2019), PE 624.262
  66. European Union Legislation: Council directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Official Journal of the European Communities **L 180/22** (2000)
  67. European Union Legislation: Council directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Official Journal of the European Communities **L 303/16** (Nov 2000)
  68. European Union Legislation: Council directive 2006/54/EC of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation. Official Journal of the European Communities **L 204/23** (2006)
  69. European Union Legislation: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). Official Journal of the European Union **L 119/1** (Apr 2016)
  70. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 259–268. ACM (2015)
  71. Feng, X., Ni, Z., Shao, Z., Guo, Y.: An open framework for foundational proof-carrying code. In: Proceedings of the 2007 ACM SIGPLAN international workshop on Types in languages design and implementation. pp. 67–78. ACM (2007)
  72. Fish, B., Kun, J., Lelkes, Á.D.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining. pp. 144–152. SIAM (2016)
  73. Freund, Y.: Game theory, on-line prediction and boosting. In: Proceedings of the Ninth Annual Conference on Computational Learning Theory. pp. 325–332 (1996)
  74. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im)possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)
  75. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 329–338. ACM (2019)
  76. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017)
  77. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: testing software for discrimination. In: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. pp. 498–510. ACM (2017)
  78. Garavel, H., Graf, S.: Formal methods for safe and secure computer systems. Federal Office for Information Security (2013)
  79. Garcez, A.d., Besold, T.R., De Raedt, L., Földiák, P., Hitzler, P., Icard, T., Kühnberger, K.U., Lamb, L.C., Miikkulainen, R., Silver, D.L.: Neural-symbolic learning and reasoning: Contributions and challenges. In: 2015 AAAI Spring Symposium Series (2015)

80. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé III, H., Crawford, K.: Datasheets for datasets. In: The 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning. Proceedings of Machine Learning, PMLR (2018)
81. Goel, S., Rao, J.M., Shroff, R.: Personalized risk assessments in the criminal justice system. *American Economic Review* **106**(5), 119–23 (2016)
82. Goel, S., Rao, J.M., Shroff, R.: Precinct or prejudice? understanding racial disparities in New York City’s stop-and-frisk policy. *The Annals of Applied Statistics* **10**(1), 365–394 (2016)
83. Goodall, N.J.: Can you program ethics into a self-driving car? *IEEE Spectrum* **53**(6), 28–58 (2016)
84. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press (2016), <http://www.deeplearningbook.org>
85. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in neural information processing systems*. pp. 2672–2680 (2014)
86. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: *Advances in neural information processing systems*. pp. 513–520 (2007)
87. Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: Feature selection for fair decision making. In: *NIPS Symposium on Machine Learning and the Law*. vol. 1, p. 2 (2016)
88. Guizzo, E., Ackerman, E.: When robots decide to kill. *IEEE Spectrum* **53**(6), 38–43 (2016)
89. Gunkel, D.J.: *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. MIT Press (2012)
90. Gurumoorthy, K.S., Dhurandhar, A., Cecchi, G.A., Aggarwal, C.: Efficient data representation by selecting prototypes with importance weights. (Aug 2019), to be published at ICDM’19
91. Hacker, P., Wiedemann, E.: A continuous framework for fairness. arXiv preprint arXiv:1712.07924 (2017)
92. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* **25**(7), 1445–1459 (2012)
93. Hajian, S., Domingo-Ferrer, J., Martinez-Balleste, A.: Rule protection for indirect discrimination prevention in data mining. In: *International Conference on Modeling Decisions for Artificial Intelligence*. pp. 211–222. Springer (2011)
94. Hamid, N.A., Shao, Z., Trifonov, V., Monnier, S., Ni, Z.: A syntactic approach to foundational proof-carrying code. *Journal of Automated Reasoning* **31**(3-4), 191–229 (2003)
95. Hammer, B., Hitzler, P.: *Perspectives of Neural-Symbolic Integration*, vol. 77. Springer (2007)
96. Hand, D.J.: Classifier technology and the illusion of progress. *Statistical science* pp. 1–14 (2006)
97. Hardt, M., Price, E., Srebro, N., et al.: Equality of opportunity in supervised learning. In: *Advances in neural information processing systems*. pp. 3315–3323 (2016)
98. Harris, R., Forrester, D.: The suburban origins of redlining: A Canadian case study, 1935-54. *Urban Studies* **40**(13), 2661–2686 (2003)

99. Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. In: European Conference on Computer Vision. pp. 793–811. Springer (2018)
100. Hind, M., Wei, D., Campbell, M., Codella, N.C., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., Varshney, K.R.: TED: Teaching AI to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 123–129. ACM (2019)
101. Hoadley, B.: Comment on ‘statistical modeling: The two cultures’ by I. Breiman. *Statistical Science* **16**(3), 220–224 (2001)
102. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
103. Huang, X., Kroening, D., Kwiatkowska, M., Ruan, W., Sun, Y., Thamo, E., Wu, M., Yi, X.: Safety and trustworthiness of deep neural networks: A survey. arXiv preprint arXiv:1812.08342 (2018)
104. Hunt, D.B.: Redlining. *Encyclopedia of Chicago* (2005)
105. Ingold, D., Soper, S.: Amazon does not consider the race of its customers. should it? *Bloomberg* (Apr 2016)
106. Johndrow, J.E., Lum, K., et al.: An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* **13**(1), 189–220 (2019)
107. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication. pp. 1–6. IEEE (2009)
108. Kamiran, F., Calders, T.: Classification with no discrimination by preferential sampling. In: Proc. 19th Machine Learning Conf. Belgium and The Netherlands. pp. 1–6. Citeseer (2010)
109. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* **33**(1), 1–33 (2012)
110. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE International Conference on Data Mining. pp. 869–874. IEEE (2010)
111. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: 2012 IEEE 12th International Conference on Data Mining. pp. 924–929. IEEE (2012)
112. Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems* **35**(3), 613–644 (2013)
113. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 35–50. Springer (2012)
114. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 643–650. IEEE (2011)
115. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Towards proving the adversarial robustness of deep neural networks. arXiv preprint arXiv:1709.02802 (2017)
116. Kay, M., Matuszek, C., Munson, S.A.: 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. pp. 3819–3828. ACM (Apr 2015)

117. Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., Weller, A.: Blind justice: Fairness with encrypted sensitive attributes. In: Proceedings of the 35th International Conference on Machine Learning (ICML 2018). vol. 80, pp. 2635–2644. International Machine Learning Society (IMLS) (2018)
118. Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning. In: Advances in Neural Information Processing Systems. pp. 656–666 (2017)
119. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *stat* **1050**, 10 (2014)
120. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: Papadimitriou, C.H. (ed.) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Leibniz International Proceedings in Informatics (LIPIcs), vol. 67, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2017)
121. Kuhn, P.: Sex discrimination in labor markets: The role of statistical evidence. *The American Economic Review* pp. 567–583 (1987)
122. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems. pp. 4066–4076 (2017)
123. Kwiatkowska, M.Z.: Safety verification for deep neural networks with provable guarantees (invited paper). In: Fokink, W., van Glabbeek, R. (eds.) 30th International Conference on Concurrency Theory (CONCUR 2019). Leibniz International Proceedings in Informatics (LIPIcs), vol. 140, pp. 1:1–1:5. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2019)
124. LaCour-Little, M.: Discrimination in mortgage lending: A critical review of the literature. *Journal of Real Estate Literature* **7**(1), 15–49 (Jan 1999)
125. Leadership Conference on Civil and Human Rights: Civil rights principles for the era of big data. <https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/> (2014)
126. Lerman, J.: Big data and its exclusions. *Stanford Law Review Online* **66**, 55 (2013)
127. Li, Y., Swersky, K., Zemel, R.: Learning unbiased features. arXiv preprint arXiv:1412.5244 (2014)
128. Lin, P., Abney, K., Bekey, G.A.: Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press (2014)
129. Lippert-Rasmussen, K.: “we are all different”: Statistical discrimination and the right to be treated as an individual. *The Journal of Ethics* **15**(1-2), 47–59 (2011)
130. Loftus, J.R., Russell, C., Kusner, M.J., Silva, R.: Causal reasoning for algorithmic fairness. arXiv preprint arXiv:1805.05859 (2018)
131. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R.: The variational fair autoencoder. *stat* **1050**, 12 (2015)
132. Louppe, G., Wehenkel, L., Sutura, A., Geurts, P.: Understanding variable importances in forests of randomized trees. In: Advances in Neural Information Processing Systems. pp. 431–439 (2013)
133. Lum, K., Isaac, W.: To predict and serve? *Significance* **13**(5), 14–19 (2016)
134. Lum, K., Johndrow, J.E.: A statistical framework for fair predictive algorithms. *stat* **1050**, 25 (2016)
135. Luong, B.T., Ruggieri, S., Turini, F.: 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. pp. 502–510. ACM (2011)
136. McNamara, D., Ong, C.S., Williamson, R.C.: Provably fair representations. arXiv preprint arXiv:1710.04394 (2017)



137. McNamara, D., Ong, C.S., Williamson, R.C.: Costs and benefits of fair representation learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 263–270. ACM (2019)
138. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Conference on Fairness, Accountability and Transparency. pp. 107–118 (2018)
139. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 279–288. ACM (2019)
140. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1765–1773 (2017)
141. Mouzannar, H., Ohannessian, M.I., Srebro, N.: From fair decision making to social equality. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 359–368. ACM (2019)
142. Nabi, R., Shpitser, I.: Fair inference on outcomes. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
143. Nash Jr, J.F.: The bargaining problem. *Econometrica: Journal of the Econometric Society* pp. 155–162 (1950)
144. Necula, G.C.: Proof-carrying code. In: Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages. pp. 106–119. ACM (1997)
145. Nissenbaum, H.: Computing and accountability. *Communications of the ACM* **37**(1), 72–81 (1994)
146. Nocedal, J., Wright, S.: Numerical optimization. Springer Science & Business Media (2006)
147. Paaßen, B., Bunge, A., Hainke, C., Sindelar, L., Vogelsang, M.: Dynamic fairness – breaking vicious cycles in automatic decision making. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (Apr 2019)
148. Page Scott, E.: The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton University Press (2007)
149. Pearl, J.: Causality: models, reasoning and inference, vol. 29. Springer (2000)
150. Pearl, J.: Direct and indirect effects. In: Proceedings of the seventeenth conference on uncertainty in artificial intelligence. pp. 411–420. Morgan Kaufmann Publishers Inc. (2001)
151. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 560–568. ACM (2008)
152. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 581–592. SIAM (2009)
153. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Advances in Neural Information Processing Systems. pp. 5680–5689 (2017)
154. Quadrianto, N., Sharmanska, V., Thomas, O.: Neural styling for interpretable fair representations. arXiv preprint arXiv:1810.06755 (2018)
155. Quinlan, J.R.: C 4.5: Programs for machine learning. The Morgan Kaufmann Series in Machine Learning, San Mateo, CA: Morgan Kaufmann (1993)
156. Ramadan, Q., Ahmadian, A.S., Strüber, D., Jürjens, J., Staab, S.: Model-based discrimination analysis: A position paper. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 22–28. IEEE (2018)

157. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning. pp. 1278–1286 (2014)
158. Ribeiro, M.T., Singh, S., Guestrin, C.: “why should i trust you?”: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144. ACM (2016)
159. Robinson, D., Yu, H., Rieke, A.: Civil rights, big data, and our algorithmic future. In: Leadership Conference on Civil and Human Rights. vol. 1 (2014), available at: <https://bigdata.fairness.io>
160. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* **29**(5), 582–638 (2014)
161. Rosenblatt, F.: Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Tech. rep., Cornell Aeronautical Lab Inc Buffalo NY (1961)
162. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **4**(2), 9 (May 2010)
163. Russell, C., Kusner, M.J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: Advances in Neural Information Processing Systems. pp. 6414–6423 (2017)
164. Ryu, H.J., Adam, H., Mitchell, M.: Inclusivefacenet: Improving face attribute detection with race and gender diversity. 2018 ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (2017)
165. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577 (2018)
166. Shen, X., Diamond, S., Gu, Y., Boyd, S.: Disciplined convex-concave programming. In: 2016 IEEE 55th Conference on Decision and Control (CDC). pp. 1009–1014. IEEE (2016)
167. Shpitser, I.: Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive science* **37**(6), 1011–1035 (2013)
168. Simoiu, C., Corbett-Davies, S., Goel, S., et al.: The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* **11**(3), 1193–1216 (2017)
169. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2219–2228. KDD '18, ACM, New York, NY, USA (2018)
170. Skirpan, M., Gorelick, M.: The authority of “fair” in machine learning. arXiv preprint arXiv:1706.09976 (2017)
171. Sokol, K., Santos-Rodriguez, R., Flach, P.: FAT forensics: A python toolbox for algorithmic fairness, accountability and transparency. arXiv preprint arXiv:1909.05167 (2019)
172. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Australian conference on artificial intelligence. vol. 4304, pp. 1015–1021 (2006)
173. Sweeney, L.: Discrimination in online ad delivery. *Commun. ACM* **56**(5), 44–54 (Jan 2013)
174. Taylor, L.: What is data justice? the case for connecting digital rights and freedoms globally. *Big Data & Society* **4**(2) (2017)
175. Tharwat, A.: Classification assessment methods. *Applied Computing and Informatics* (2018)

176. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. 1 edn. (2019)
177. The White House. Executive Office of the President: Big data: Seizing opportunities and preserving values (May 2014)
178. The White House. Executive Office of the President: Big data: Seizing opportunities and preserving values: Interim progress report. (Feb 2015)
179. The White House. Executive Office of the President: Big data: A report on algorithmic systems, opportunity, and civil rights (May 2016)
180. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288 (1996)
181. Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.P., Humbert, M., Juels, A., Lin, H.: FairTest: Discovering unwarranted associations in data-driven applications. In: 2017 IEEE European Symposium on Security and Privacy (EuroS&P). pp. 401–416. IEEE (2017)
182. Trenkler, G., Stahlecker, P.: Dropping variables versus use of proxy variables in linear regression. *Journal of Statistical Planning and Inference* **50**(1), 65–75 (1996), *econometric Methodology, Part III*
183. Tufekci, Z.: Algorithmic harms beyond facebook and google: Emergent challenges of computational agency. *Colo. Tech. LJ* **13**, 203 (2015)
184. Turner, M.A., Skidmore, F.: Mortgage lending discrimination: A review of existing evidence (1999)
185. UK Legislation: Sex discrimination act 1975 c. 65 (1975)
186. UK Legislation: Race relations act 1976 c. 74 (1976)
187. UK Legislation: Disability discrimination act 1995 c. 50 (1995)
188. UK Legislation: Equality act 2010 c. 15 (Oct 2000)
189. U.S. Federal Legislation: The equal pay act of 1963, pub.l. 88–38, 77 stat. 56 (Apr 1963)
190. U.S. Federal Legislation: Civil rights act of 1968, pub.l. 90–284, 82 stat. 73 (Apr 1968)
191. U.S. Federal Legislation: Equal credit opportunity act of 1974, 15 u.s.c. § 1691 et seq. (Oct 1974)
192. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 10–19. ACM (2019)
193. Varian, H.R.: Equity, envy, and efficiency (1973)
194. Verma, S., Rubin, J.: Fairness definitions explained. In: FairWare’18: IEEE/ACM International Workshop on Software Fairness. ACM (May 2018)
195. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* **12**(4), 5–33 (1996)
196. Wei, D., Dash, S., Gao, T., Gunluk, O.: Generalized linear rule models. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6687–6696. PMLR, Long Beach, California, USA (09–15 Jun 2019)
197. Woodcock, J., Larsen, P.G., Bicarregui, J., Fitzgerald, J.: Formal methods: Practice and experience. *ACM computing surveys (CSUR)* **41**(4), 19 (2009)
198. Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. In: Kale, S., Shamir, O. (eds.) Proceedings of the 2017 Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 65, pp. 1920–1953. PMLR, Amsterdam, Netherlands (07–10 Jul 2017)

199. Wu, M., Wicker, M., Ruan, W., Huang, X., Kwiatkowska, M.: A game-based approximate verification of deep neural networks with provable guarantees. *Theoretical Computer Science* (2019)
200. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In: *Proceedings of the 2003 SIAM International Conference on Data Mining*. pp. 331–335. SIAM (2003)
201. Zadrozny, B.: Learning and evaluating classifiers under sample selection bias. In: *Proceedings of the Twenty-First International Conference on Machine Learning*. p. 114. ACM (2004)
202. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: *Proceedings of the 26th International Conference on World Wide Web*. pp. 1171–1180. International World Wide Web Conferences Steering Committee (2017)
203. Zafar, M.B., Valera, I., Roriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: *Artificial Intelligence and Statistics*. pp. 962–970 (2017)
204. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: FA\*IR: A fair top-k ranking algorithm. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. pp. 1569–1578. CIKM '17, ACM, New York, NY, USA (2017)
205. Zehlike, M., Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. *arXiv preprint arXiv:1805.08716* (2018)
206. Zehlike, M., Castillo, C., Bonchi, F., Baeza-Yates, R., Hajian, S., Megahed, M.: Fairness measures: A platform for data collection and benchmarking in discrimination-aware ML. <http://fairness-measures.org> (Jun 2017)
207. Zehlike, M., Sühr, T., Castillo, C., Kitanovski, I.: FairSearch: A tool for fairness in ranked search results (2019)
208. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: *International Conference on Machine Learning*. pp. 325–333 (2013)
209. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. pp. 335–340. ACM (2018)
210. Žliobaitė, I.: On the relation between accuracy and fairness in binary classification. In: *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15* (2015)
211. Žliobaitė, I.: Fairness-aware machine learning: a perspective. *arXiv preprint arXiv:1708.00754* (2017)
212. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* **31**(4), 1060–1089 (Jul 2017)