Fairness in Machine Learning: A Survey

A Preprint

Simon Caton
University College Dublin
Dublin, Ireland
simon.caton@ucd.ie

Christian Haas
University of Nebraska at Omaha
Omaha, US
christianhaas@unomaha.edu

October 9, 2020

Abstract

As Machine Learning technologies become increasingly used in contexts that affect citizens, companies as well as researchers need to be confident that their application of these methods will not have unexpected social implications, such as bias towards gender, ethnicity, and/or people with disabilities. There is significant literature on approaches to mitigate bias and promote fairness, yet the area is complex and hard to penetrate for newcomers to the domain. This article seeks to provide an overview of the different schools of thought and approaches to mitigating (social) biases and increase fairness in the Machine Learning literature. It organises approaches into the widely accepted framework of pre-processing, in-processing, and post-processing methods, subcategorizing into a further 11 method areas. Although much of the literature emphasizes binary classification, a discussion of fairness in regression, recommender systems, unsupervised learning, and natural language processing is also provided along with a selection of currently available open source libraries. The article concludes by summarising open challenges articulated as four dilemmas for fairness research.

Keywords fairness, accountability, transparency, machine learning

1 Introduction

Machine Learning (ML) technologies solve challenging problems which often have high social impact, such as examining re-offence rates (e.g. [27, 43, 98, 213, 11, 25]), automating chat and (tech) support, and screening job applications (see [226, 267]). Yet, approaches in ML have "found dark skin unattractive",[1] claimed that "black people reoffend more",[2] and created a Neo-Nazi sexbot.[3] With the increasingly widespread use of automated decision making and ML approaches in general, fairness considerations in ML have gained significant attention in research and practice in the 2010s. However, from a historical perspective these modern approaches often build on prior definitions, concepts, and considerations that have been suggested and developed over the past five decades. Specifically, there is a rich set of fairness-related work in a variety of disciplines, often with concepts that are similar or equal to current ML fairness research [137]. For example, discrimination in hiring decisions has been examined since the 1960s [122]. Research into (un)fairness, discrimination, and bias emerged after the 1964 US Civil Rights act, making it illegal to discriminate based on certain criteria in the context of government agencies (Title VI) and employment (Title VII). Two initial foci of fairness research were unfairness of standardized tests in higher education/university contexts [69, 70] as well as discrimination in employment-based concepts [122]. The first years after the Civil Rights act saw the emergence of a variety of definitions, metrics, and scholarly disputes about the applicability of various definitions and fairness concepts as well as the realizations that some concepts (such as group-based vs individual notions of fairness) can be incompatible.

When comparing current work in ML with initial work in fairness, it is noteworthy that much of the early literature considers regression settings as well a correlation-based definition of fairness properties of an underlying mechanism

[1] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-peo
[2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[3] https://www.technologyreview.com/s/610634/microsofts-neo-nazi-sexbot-was-a-great-lesson-for-makers-of-ai-assis

due to the focus on test fairness with a continuous target variable (see e.g., [77]). However, the general notions transfer to (binary) classification settings and thus define essential concepts such as protected/demographic variables (e.g. [69, 70]), notions of group vs individual fairness (e.g. [260, 242]), impossibility conditions between different fairness conditions ([77]), and fairness quantification based on metrics (e.g., true positive rates, [71]).

Despite the increased discussion of different aspects and viewpoints of fairness in the 1970s as well as the founding of many modern fairness concepts, no general consensus as to what constitutes fairness or if it can/should be quantified emerged based on this first wave of fairness research. As [137] note, some of the related discussions resonate with current discussions in ML, e.g., the difficulty that different notions can be incompatible with each other, or the fact that each specific quantified measurement of fairness seems to have particular downsides.

More recently, many researchers (e.g. [118, 269, 49, 72, 15, 33, 96, 217, 297, 41, 42, 203, 181]), governments (e.g. the EU in [2, 1], and the US in [224, 207]), policies like the General Data Protection Regulation (GDPR), NGOs (e.g. the Association of Internet Researchers [197]) and the media have fervently called for more societal accountability and social understanding of ML. There is a recognition in the literature that often data is the problem, i.e. intrinsic biases in the sample will manifest themselves in any model built on the data [42, 33], inappropriate uses of data leading to (un)conscious bias(es) [41, 42], data veracity and quality [297], data relativity and context shifts [42, 233, 114], and subjectivity filters [41]. Even for skilled ML researchers, the array of challenges can be overwhelming and current ML libraries often do not yet accommodate means to ascertain social accountability. Note that data is not the only source of bias and discrimination, here we refer to [201] for a general discussion on the main types of bias and discrimination in ML.

ML researchers have responded to this call, developing a large number of metrics to quantify fairness in decisions (automated or otherwise) and mitigate any bias and unfairness issues in ML. Figure 1 shows the number of papers, starting in 2010, that have been published in the fairness in ML domain. These numbers are based on the articles referenced in our survey. The figure shows a clear uptick in papers starting in 2016 and 2017.[4]

In this respect, this article aims to provide an entry-level overview of the current state of the art for fairness in ML. This article builds on other similarly themed reviews that have focused on the history of fairness in ML [137], a multidisciplinary survey of discrimination analysis [231], a discussion on key choices and assumptions [205] and finally [201, 256] who review different types of biases, [201] also introduce a number of methods to mitigate these. In this article we assume that the reader has a working knowledge of applied ML, i.e. they are familiar with the basic structure of data mining methodologies such as [95] and how to apply and evaluate "standard" ML methods.

Upon this basis, this article aims to: 1) Provide an easy introduction into the area fairness in ML (Section 2); 2) Summarise the current approaches to measure fairness in ML within a standardised notation framework discussing the various trade-offs of each approach as well as their overarching objectives (Section 3); 3) Define a two-dimensional taxonomy of approach categories to act as a point of reference. Within this taxonomy, we highlight the main approaches, assumptions, and general challenges for binary classification (Section 4) and beyond binary classification (Section 5); 4) Highlight currently available toolkits for fair ML (Section 6); and 5) Outline the dilemmas for fairness research as avenues of future work to improve the accessibility of the domain (Section 7).

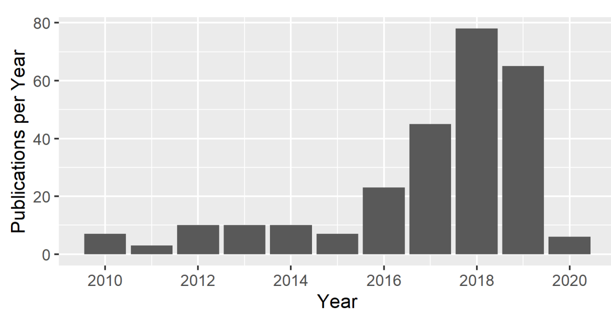

Figure 1: Number of Papers related to Fairness in ML research based on cited articles in this survey.

[4]The numbers for 2020 are incomplete as the survey was submitted during the year.

2 Fairness in Machine Learning: key methodological components

Much of the related literature focuses on either the technical aspects of bias and fairness in ML, or theorizing on the social, legal, and ethical aspects of ML discrimination [116]. Technical approaches are typically applied prior to modelling (pre-processing), at the point of modelling (in-processing), or after modelling (post-processing), i.e. they emphasize intervention [33]. In this paper we focus on technical approaches, and in this section give a high-level overview of the framework for an intervention-based methodology for fairness in ML; see Figure 2 for a graphical representation. Whilst not all approaches for fair ML fit into this framework, it provides a well-understood point of reference and acts as one dimension in a taxonomy of approaches for fairness in ML.

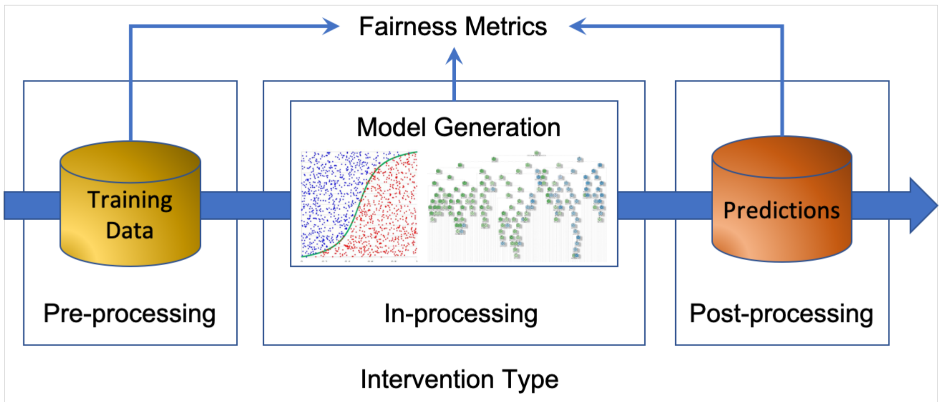

Figure 2: High Level Illustration of Fairness in ML. Note we omit standard practices common within methodologies like KDD [95] to prepare and sample data for ML methods and their evaluation for visual simplicity.

2.1 Sensitive and Protected Variables and (Un)privileged Groups

Most approaches to mitigate unfairness, bias, or discrimination are based on the notion of protected or sensitive variables (we will use the terms interchangeably) and on (un)privileged groups: groups (often defined by one or more sensitive variables) that are disproportionately (less) more likely to be positively classified. Before discussing the key components of the fairness framework, a discussion on the nature of protected variables is needed. Protected variables define the aspects of data which are socioculturally precarious for the application of ML. Common examples are gender, ethnicity, and age (as well as their synonyms). However, the notion of a protected variable can encompass any feature of the data that involves or concerns people [17].

The question of which variables should be protected quickly arises. We note that many variables are explicitly defined as “sensitive” by specific legal frameworks, see [262, 180, 182, 283, 120, 119, 25, 196, 251] and the references therein. While there are some readily available sets of declared sensitive variables, there are relatively few works that actively seek to determine whether other variables or rare (i.e., minority) combinations should be protected or not. [101, 4] both present approaches specifically looking at the variable importance (or model influence) of sensitive variables, and could act as a means to identify potentially problematic variables. Yet, there is still the question of variables that are not strictly sensitive, but have a relationship with one or more sensitive variables. [63] notes that many definitions of fairness express model output in terms of sensitive variables, without considering “related” variables. Not considering these related variables could erroneously assume a fair ML model has been produced. Not considering correlated variables has been shown to increase the risk of discrimination (e.g., redlining[5]) [217, 231, 285, 268, 87, 54, 86, 191, 87, 185].

Understanding “related” variables in a general sense is a well studied area, especially in the privacy and data archiving literature, where overlooked variable relationships can enable the deanonymization of published data (see [295]). The fairness literature, however, often overlooks these effects on fairness, although the relationship between discrimination and privacy was noted in [87]. In particular, sensitive data disclosure is a long-standing challenge in protecting citizen anonymity when data is published and/or analysed [8, 105, 184]. Key approaches in this area (e.g. [257, 193, 183]) seek to protect specific individuals and groups from being identifiable within a given dataset, i.e. minimize disclosure risk. Yet, these approaches can still struggle to handle multiple sensitive attributes at once [183]. Whilst

[5]The term redlining stems from the United States and describes maps that were color coded to represent areas a bank would not invest in, e.g. give loans to residents of these areas [144].

these approaches (and many others) have been successful in anonymizing datasets, they still often require a list of features to protect. For explicit identifiers (such as name, gender, zip etc.) such lists exist as already discussed. For correlated or otherwise related variables (often referred to as proxies or quasi-identifiers), much of the literature assumes a priori knowledge of the set of quasi-identifiers [105], or seeks to discover them on a case-by-case basis (e.g. [206, 139]), and moves towards a notion of privacy preserving data mining as introduced by [8]. [123] also discuss the notion of proxy groups, a set of "similar" instances of the data that could correspond to a protected group (e.g. young women).

More recently, fairness researchers have begun to investigate graph- and network-based methods for discovering proxies either with respect to anonymity criteria (e.g. [279]) or specific notions of fairness (we introduce these approaches in subsection 4.2). [236] provides a brief overview of different theoretical applications to algorithmic fairness, with [110] noting how different causal graph-based models can help use variable relationships to distill different biases in the model and/or data. Table 1 provides some examples of sensitive variables and potential proxies. Ultimately, users need to thoroughly consider how they will identify and define the set of protected variables.

Sensitive Variable	Example Proxies
Gender	Education Level, Income, Occupation, Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), University Faculty, Working Hours
Marital Status	Education Level, Income
Race	Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), Zipcode
Disabilities	Personality Test Data

Table 1: Example Proxy relationships based on findings from [101, 25, 236, 37, 245, 127, 282, 244, 274, 198]

2.2 Metrics

Underpinning the intervention-based approaches are an ever increasing array of fairness measures seeking to quantify fairness. The implication of "measurement" is, however, precarious as it implies a straightforward process [17]. Aside from the philosophical and ethical debates on defining fairness (often overlooked in the ML literature), creating generalized notions of fairness quantification is challenging. Metrics usually either emphasize individual (e.g. everyone is treated equal), or group fairness, where the latter is further differentiated to within group (e.g. women vs. men) and between group (e.g. young women vs. black men) fairness. Currently, combinations of these ideals using established definitions have been shown to be mathematically intractable [170, 66, 27]. Quantitative definitions allow fairness to become an additional performance metric in the evaluation of an ML algorithm. However, increasing fairness often results in lower overall accuracy or related metrics, leading to the necessity of analyzing potentially achievable trade-offs in a given scenario [124].

2.3 Pre-processing

Pre-processing approaches recognize that often an issue is the data itself, and the distributions of specific sensitive or protected variables are biased, discriminatory, and/or imbalanced. Thus, pre-processing approaches tend to alter the sample distributions of protected variables, or more generally perform specific transformations on the data with the aim to remove discrimination from the training data [155]. The main idea here is to train a model on a "repaired" data set. Pre-processing is argued as the most flexible part of the data science pipeline, as it makes no assumptions with respect to the choice of subsequently applied modeling technique [86].

2.4 In-processing

In-processing approaches recognize that modeling techniques often become biased by dominant features, other distributional effects, or try to find a balance between multiple model objectives, for example having a model which is both accurate and fair. In-processing approaches tackle this by often incorporating one or more fairness metrics into the model optimization functions in a bid to converge towards a model parameterization that maximizes performance and fairness.

2.5 Post-processing

Post-processing approaches recognize that the actual output of an ML model may be unfair to one or more protected variables and/or subgroup(s) within the protected variable. Thus, post-processing approaches tend to apply transformations to model output to improve prediction fairness. Post-processing is one of the most flexible approaches as it only needs access to the predictions and sensitive attribute information, without requiring access to the actual algorithms and ML models. This makes them applicable for black-box scenarios where not the entire ML pipeline is exposed.

2.6 Initial Considerations: pre-processing vs. in-processing vs. post-processing

A distinct advantage of pre- and post-processing approaches is that they do not modify the ML method explicitly. This means that (open source) ML libraries can be leveraged unchanged for model training. However, they have no direct control over the optimization function of the ML model itself. Yet, modification of the data and/or model output may have legal implications [16] and can mean models are less interpretable [182, 191], which may be at odds with current data protection legislation with respect to explainability. Only in-processing approaches can optimize notions of fairness during model training. Yet, this requires the optimization function to be either accessible, replaceable, and/or modifiable, which may not always be the case.

3 Measuring Fairness and Bias

Behind intervention-based approaches are a myriad of definitions and metrics (e.g. [16, 27, 66, 129, 168, 276, 284]) to mathematically represent bias, fairness, and/or discrimination; but they lack consistency in naming conventions [72] and notation. More so, there are many different interpretations of what it means for an algorithm to be "fair". Several previous publications provide a (limited) overview of multiple fairness metrics and definitions, e.g., [270, 256, 222, 230]. We extend these prior summaries by including additional perspectives for types of biases and a larger set of metrics and definitions that are included as compared to previous publications.

Typically, metrics fall under several categories, for example: 1) *statistical parity*: where each group receives an equal fraction of possible [decision] outcomes [96, 158, 288]; 2) *disparate impact*: a quantity that captures whether wildly different outcomes are observed in different [social] groups [96, 284]; 3) *equality of opportunity* [129], 4) *calibration* [168]: where false positive rates across groups are enforced to be similar (defined as *disparate mistreatment* by [284] when this is not the case), 5) *counterfactual fairness* which states that a decision is fair towards an individual if it coincides with one that would have been taken were the sensitive variable(s) different [62].

Although the literature has defined a myriad of notions to quantify fairness, each measures and emphasizes different aspects of what can be considered "fair". Many are difficult / impossible to combine [170, 66], but ultimately, we must keep in mind (as noted in [65]) there is no universal means to measure fairness, and also at present no clear guideline(s) on which measures are "best". Thus, in this section we provide an overview of fairness measures and seek to provide a lay interpretation to help inform decision making. Table 2 presents an overview of the categories of fairness metrics presented with Table 3 introducing key notation.

	Group-based Fairness				Individual and Counterfactual Fairness
	Parity-based Metrics	Confusion Matrix-based Metrics	Calibration-based Metrics	Score-based Metrics	Distribution-based Metrics
Concept	Compare predicted positive rates across groups	Compare groups by taking into account potential underlying differences between groups	Compare based on predicted probability rates (scores)	Compare based on expected scores	Calculate distributions based on individual classification outcomes
Abstract Criterion	Independence	Separation	Sufficiency	-	-
Examples	Statistical Parity, Disparate Impact	Accuracy equality, Equalized Odds, Equal Opportunity	Test fairness, Well calibration	Balance for positive and negative class, Bayesian Fairness	Counterfactual Fairness, Generalized Entropy Index

Table 2: Overview of suggested fairness metrics for binary classification

3.1 Abstract Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier: First, the sensitive variable S that defines the groups for which we want to measure fairness. Second, the target variable Y. In binary classification, this represents the two classes that we can predict: $Y = 0$ or $Y = 1$. Third, the classification score R, which represents the predicted score (within $[0, 1]$) that a classifier yields for each observation. Using these properties, general fairness desiderata are categorized into three "non-discrimination" criteria [17]:

Symbol	Description
$y \in 0, 1$	Actual value / outcome
$\hat{y} \in 0, 1$	Predicted value / outcome
$s = Pr(\hat{y}_i = 1)$	Predicted score of an observation i. Probability of $y = 1$ for observation i
g_i, g_j	Identifier for groups based on protected attribute

Table 3: Notation for Binary Classification

Independence aims for classifiers to make their scoring independent of the group membership:

$$R \perp S \tag{1}$$

An example group fairness metric focusing on independence is Statistical/Demographic Parity. Independence does not take into account that the outcome Y might be correlated with the sensitive variable S. I.e., if the separate groups have different underlying distributions for Y, not taking these dependencies into account can lead to outcomes that are considered fair under the Independence criterion, but not for (some of the) groups themselves. Hence, an extension of the Independence property is the **Separation** criterion which looks at the independence of the score and the sensitive variable conditional on the value of the target variable Y:

$$R \perp S|Y \tag{2}$$

Example metrics that target the Separation property are Equalized Odds and Equal Opportunity. The third criterion commonly used is **Sufficiency**, which looks at the independence of the target Y and the sensitive variable S, conditional for a given score R:

$$Y \perp S|R \tag{3}$$

As [17] point out, Sufficiency is closely related to some of the calibration-based metrics. [17] also discuss several impossibility results with respect to these three criteria. For example, they show that if S and Y are not independent, then Independence and Sufficiency cannot both be true. This falls into a more general discussion on impossibility results between fairness metrics.

3.2 Group Fairness Metrics

Group-based fairness metrics essentially compare the outcome of the classification algorithm for two or more groups. Commonly, these groups are defined through the sensitive variable as described in subsection 2.1. Over time, many different approaches have been suggested, most of which use metrics based on the binary classification confusion matrix to define fairness.

3.2.1 Parity-based Metrics

Parity-based metrics typically consider the predicted positive rates, i.e., $Pr(\hat{y} = 1)$, across different groups. This is related to the Independence criterion that was defined in subsection 3.1.

Statistical/Demographic Parity: One of the earliest definitions of fairness, this metric defines fairness as an equal probability of being classified with the positive label [288, 158, 96, 73]. I.e., each group has the same probability of being classified with the positive outcome. A disadvantage of this notion, however, is that potential differences between groups are not being taken into account.

$$Pr(\hat{y} = 1|g_i) = Pr(\hat{y} = 1|g_j) \tag{4}$$

Disparate Impact: Similar to statistical parity, this definition looks at the probability of being classified with the positive label. In contrast to parity, Disparate Impact considers the ratio between unprivileged and privileged groups. Its origins are in legal fairness considerations for selection procedures which sometimes use an 80% rule to define if a process has disparate impact (ratio smaller than 0.8) or not [96].

$$\frac{Pr(\hat{y} = 1|g_1)}{Pr(\hat{y} = 1|g_2)} \tag{5}$$

While often used in the (binary) classification setting, notions of Disparate Impact are also used to define fairness in other domains, e.g., dividing a finite supply of items among participants [219].

3.2.2 Confusion Matrix-based Metrics

While parity-based metrics typically consider variants of the predicted positive rate $Pr(\hat{y} = 1)$, confusion matrix-based metrics take into consideration additional aspects such as True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR). The advantage of these types of metrics is that they are able to include underlying differences between groups who would otherwise not be included in the parity-based approaches. This is related to the Separation criterion that was defined in subsection 3.1.

Equal Opportunity: As parity and disparate impact do not consider potential differences in groups that are being compared, [129, 223] consider additional metrics that make use of the FPR and TPR between groups. Specifically, an algorithm is considered to be fair under equal opportunity if its TPR is the same across different groups.

$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \tag{6}$$

Equalized Odds (Conditional procedure accuracy equality [27]): Similarly to equal opportunity, in addition to TPR equalized odds simultaneously considers FPR as well, i.e., the percentage of actual negatives that are predicted as positive.

$$Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) \;\; \& \;\; Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 0\&g_j) \tag{7}$$

Overall accuracy equality [27]: Accuracy, i.e., the percentage of overall correct predictions (either positive or negative), is one of the most widely used classification metrics. [27] adjusts this concept by looking at relative accuracy rates across different groups. If two groups have the same accuracy, they are considered equal based on their accuracy.

$$Pr(\hat{y} = 0|y = 0\&g_i) + Pr(\hat{y} = 1|y = 1\&g_i) = Pr(\hat{y} = 0|y = 0\&g_j) + Pr(\hat{y} = 1|y = 1\&g_j) \tag{8}$$

Conditional use accuracy equality [27]: As an adaptation of the overall accuracy equality, the following conditional procedure and conditional use accuracy do not look at the overall accuracy for each subgroup, but rather at the positive and negative predictive values.

$$Pr(y = 1|\hat{y} = 1\&g_i) = Pr(y = 1|\hat{y} = 1\&g_j) \;\; \& \;\; Pr(y = 0|\hat{y} = 0\&g_i) \qquad = Pr(y = 0|\hat{y} = 0\&g_j) \tag{9}$$

Treatment equality [27]: Treatment equality considers the ratio of False Negative Predictions (FNR) to False Positive Predictions.

$$\frac{Pr(\hat{y} = 1|y = 0\&g_i)}{Pr(\hat{y} = 0|y = 1\&g_i)} = \frac{Pr(\hat{y} = 1|y = 0\&g_j)}{Pr(\hat{y} = 0|y = 1\&g_j)} \tag{10}$$

Equalizing disincentives [148]: The Equalizing disincentives metric compares the difference of two metrics, TPR and FPR, across the groups and is specified as:

$$Pr(\hat{y} = 1|y = 1\&g_i) - Pr(\hat{y} = 1|y = 0\&g_i) = Pr(\hat{y} = 1|y = 1\&g_j) - Pr(\hat{y} = 1|y = 0\&g_j) \tag{11}$$

Conditional Equal Opportunity [30]: As some metrics can be dependent on the underlying data distribution, [30] provide an additional metric that specifies equal opportunity on a specific attribute a our of a list of attributes A, where τ is a threshold value:

$$Pr(\hat{y} \geq \tau|g_i\&y < \tau\&A = a) = Pr(\hat{y} =\geq \tau|g_j\&y < \tau\&A = a) \tag{12}$$

3.2.3 Calibration-based Metrics

In comparison to the previous metrics which are defined based on the predicted and actual values, calibration-based metrics take the predicted probability, or score, into account. This is related to the Sufficiency criterion that was defined in Section 3.1.

Test fairness/ calibration / matching conditional frequencies ([66], [129]): Essentially, test fairness or calibration wants to guarantee that the probability of $y = 1$ is the same given a particular score. I.e., when two people from different groups get the same predicted score, they should have the same probability of belonging to $y = 1$.

$$Pr(y = 1|S = s\&g_i) = Pr(y = 1|S = s\&g_j) \tag{13}$$

Well calibration [168]: An extension of regular calibration where the probability for being in the positive class also has to equal the particular score.

$$Pr(y = 1|S = s\&g_i) = Pr(y = 1|S = s\&g_j) = s \tag{14}$$

3.2.4 Score-based Metrics

Balance for positive and negative class [168]: The expected predicted score for the positive and negative class has to be equal for all groups:

$$E(S = s|y = 1\&g_i) = E(S = s|y = 1\&g_j), E(S = s|y = 0\&g_i) = E(S = s|y = 0\&g_j) \tag{15}$$

Bayesian Fairness [83] extend the balance concept from [168] when model parameters themselves are uncertain. Bayesian fairness considers scenarios where the expected utility of a decision maker has to be balanced with fairness of the decision. The model takes into account the probability of different scenarios (model parameter probabilities) and the resulting fairness / unfairness.

3.3 Individual and Counterfactual Fairness Metrics

As compared to group-based metrics which compare scores across different groups, individual and counterfactual fairness metrics do not focus on comparing two or more groups as defined by a sensitive variable, but consider the outcome for each participating individual. [173] propose the concept of counterfactual fairness which builds on causal fairness models and is related to both individual and group fairness concepts. [252] proposes a generalized entropy index which can be parameterized for different values of α and measures the individual impact of the classification outcome. This is similar to established distribution indices such as the Gini Index in economics.

Counterfactual Fairness: Given a causal model (U, V, F), where U are latent (background) variables, $V = S \cup X$ are observable variables including the sensitive variable S, and F is a set of functions defining structural equations such that V is a function of U, counterfactual fairness is:

$$P(\hat{y}_{A \leftarrow a}(U) = y|X = x, A = a) = P(\hat{y}_{A \leftarrow a'}(U) = y|X = x, A = a) \tag{16}$$

Essentially, the definition ensures that the prediction for an individual coincides with the decision if the sensitive variable would have been different.

Generalized Entropy Index: [252] defines the Generalized Entropy Index (GEI) which considers differences in an individual's prediction (b_i) to the average prediction accuracy (μ). It can be adjusted based on the parameter α, where $b_i = \hat{y}_i - y_i + 1$ and $\mu = \frac{\sum_i b_i}{n}$:

$$GEI = \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^{n} \left[(\frac{b_i}{\mu})^{\alpha} - 1 \right] \tag{17}$$

Theil Index: a special case of the GEI for $\alpha = 1$. In this case, the calculation simplifies to:

$$Theil = \frac{1}{n} \sum_{i=1}^{n} (\frac{b_i}{\mu}) log(\frac{b_i}{\mu})) \tag{18}$$

3.4 Summary

The literature is at odds with respect to whether individual or group fairness should be prioritized. [252] note that many approaches to group fairness tackle only between-group issues, as a consequence they demonstrate that within-group issues are worsened through this choice. Consequently, users must decide on where to place emphasis, but be mindful of the trade off between any fairness measure and model accuracy [26, 87, 73, 129, 296, 54]. With a reliance on expressing fairness and bias mathematically, [72, 118] argue that these definitions often do not map to normative social, economic, or legal understandings of the same concepts. This is corroborated by [249] who note an over emphasis in the literature on disparate treatment. [5, 54, 252] criticize ad hoc and implicit choices concerning distributional assumptions or realities of relative group sizes.

4 Binary Classification Approaches

Building on the metrics discussed in Section 3, fairness in ML researchers seek to mitigate unfairness by "protecting" sensitive sociodemographic attributes (as introduced in subsection 2.1). The literature is dominated by approaches for mitigating bias and unfairness in ML within the problem class of binary classification [26]. There are many reasons for this, but most notably: 1) many of the most contentious application areas that motivated the domain are binary

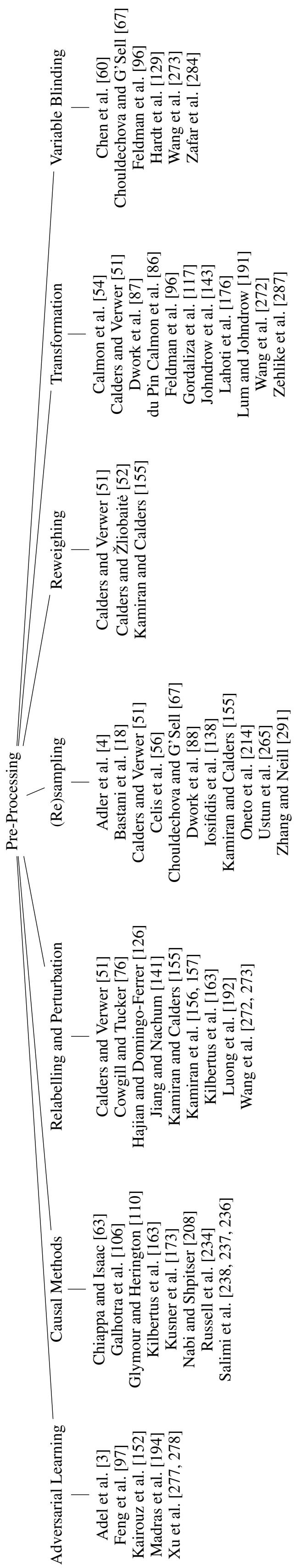

Figure 3: Pre-processing Methods

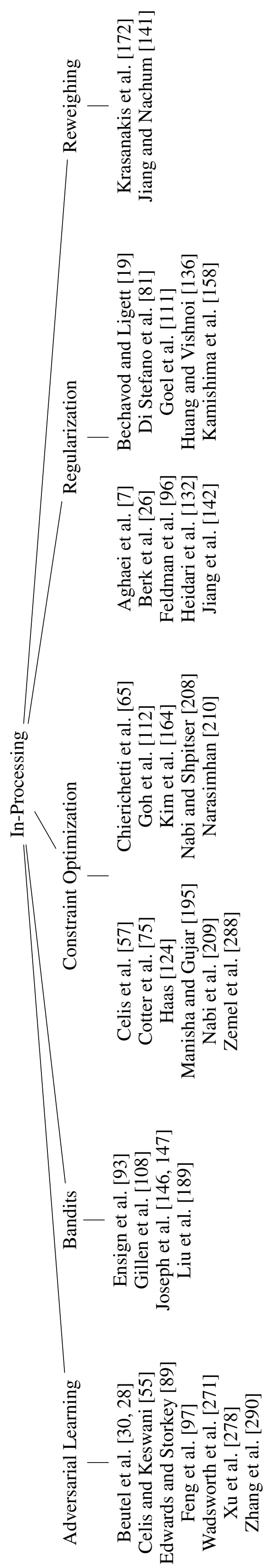

Figure 4: In-processing Methods

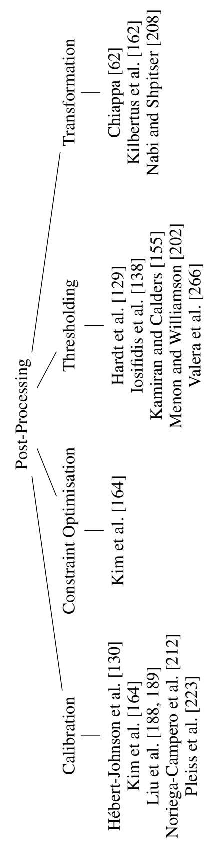

Figure 5: Post-processing methods

decisions (hiring vs. not hiring; offering a loan vs. not offering a loan; re-offending vs. not re-offending etc.); 2) quantifying fairness on a binary dependent variable is mathematically more convenient; addressing multi-class problems would at the very least add terms in the fairness quantity.

In this section, we discuss the main approaches for tackling fairness in the binary classification case. We begin by arranging mitigation methods into a visual taxonomy according to the location in the ML framework (Figure 2), i.e. pre-processing: Figure 3, in-processing: Figure 4, and post-processing: Figure 5. We note an abundance of pre- and in-processing vs. post-processing methods and that method families, e.g. methods leveraging adversarial learning, can belong to multiple stages (pre- and in-processing in this case). It is also noteworthy that many of the approaches listed (i.e. the overall mitigation strategy applied by researchers) do not belong to a single category or stage, but several: approaches tend to be hybrid and this is becoming more prevalent in more recent approaches. However, we are yet to find an approach using methods from all three stages, even if there are papers comparing methods from multiple stages. Finally, we also note that we do not comment on the advantages of specific approaches over others, yet where relevant we outline challenges researchers must navigate. The literature has an urgent need for a structured meta review of approaches to fairness. Whilst many papers compare specific subsets of the approaches in this section, they do not and realistically cannot offer a holistic comparison.

4.1 Blinding

Blinding is the approach of making a classifier "immune" to one or more sensitive variables [289]. A classifier is, for example, race blind if there is no observable outcome differentiation based on the variable race. For example, [129] seek to train a race blind classifier (among others) in that each of the 4 race groups have the same threshold value (see subsection 4.11), i.e. the provided loan rate is equal for all races. Other works have termed the omission of sensitive variables from the training data as blinding. However, we distinguish **immunity** to sensitive variables as distinct from **omission of** sensitive variables. Omission has been shown to decrease model accuracy [125, 60, 238] and not improve model discrimination [54, 155, 87]. Both omission and immunity overlook relationships with proxy variables (as discussed in subsection 2.1, we note [273] as an exception here who omit proxies), which can result in increasing instead of decreasing bias and discrimination [170], or indirectly concealing discrimination [86]. It also ignores that discrimination may not be one variable in isolation, but rather the result of several joint characteristics [217] and as such determining which combination(s) of variables to blind is non-trivial and the phenomenon of omitted variable bias should not be downplayed [68, 150].

Approaches in subgroup analysis (see subsection 4.3) have used statistical techniques to determine when variable immunity does not adversely affect fairness (e.g. [67]). Similarly, researchers still use omission in their evaluation methodologies to compare to earlier works and act as a baseline. Omission can also be used in specific parts of the fairness methodology, for example [96] temporarily omit sensitive variables prior to transforming (see subsection 4.4) the training data. Blinding (or partial blinding) has also been used as a fairness audit mechanism [4, 133, 78]. Specifically, such approaches explore how partially blinding features (sensitive or otherwise) affect model performance. This is similar to the idea of causal models (subsection 4.2) and can help identify problematic sensitive or proxy variables with black-box-like analysis of a ML model.

4.2 Causal Methods

Approaches using causal methods recognise that the data upon which ML models are trained often reflect some form of underlying discrimination. A key objective is to uncover causal relationships in the data and find dependencies between sensitive and non-sensitive variables [106, 163, 173, 63, 110, 208, 234]. Thus, causal methods are specifically well suited to identifying proxies of sensitive variables as discussed in subsection 2.1 as well as subgroup analyses of which subgroups are most (un)fairly treated and differentiate the types of bias exhibited [110]. Alternatively, casual methods can be employed to provide transparency with respect to how (classification) decisions were made [133]. In both scenarios, they provide visual descriptions of (un)fairness in the dataset upon the basis of bias in terms of causation (see [162, 173, 174, 133]). Directed acyclic graphs (DAGs) are a common means to represent conditional independence assumptions between variables [144].

Extensions have also leveraged causal dependencies to "repair" training data [238, 237, 236] using dependency information to repair (insert, modify, and remove) samples from the training data in accordance to satisfying fairness-specific constraints and conditional independence properties of the training data. Initial results with data repair methods have shown to result in "debiased" classifiers which are robust to unseen test data, yet require significant computational resources. The main challenge for causal models is that they require background information and context regarding the causal model which may not always be accessible [238]. They have also been criticized for not well examining how they would be applied in practice [238].

4.3 Sampling and Subgroup Analysis

Sampling methods have two primary objectives: 1) to create samples for the training of robust algorithms (e.g. [56, 284, 88, 265, 9, 138]), i.e. seek to "correct" training data and eliminate biases [138]; and 2) to identify groups (or subsamples) of the data that are significantly disadvantaged by a classifier, i.e. as a means to evaluate a model (e.g. [67, 4, 291]) via subgroup analysis. [56] articulate the challenge of sampling for fairness as the following question: how are samples selected from a (large) dataset that are both diverse in features and fair to sensitive attributes? Without care, sampling can propagate biases within the training data [213, 88], as ensuring diversity in the data used to train the model makes no guarantees of producing fair models [213]. As such, approaches that seek to create fair training samples include notions of fairness in the sampling strategy. [155], propose to preferentially sample (similar to oversampling) instances "close" to a decision boundary (based on an initial model prototype to approximate a decision boundary) as these are most likely to be discriminated or favored due to underlying biases within the training data.

Within the sampling approaches, the application of decoupled classifiers and multitask learning has emerged (see [284, 88, 265, 9, 214, 51]). Here, the training data is split into subgroups (decoupled classifiers), i.e. combinations of one or more sensitive variables (e.g. [old, males]), or these groupings are learned in a preprocessing stage (multitask learning). Thus, such approaches seek to make the most accurate models for given subgroups (decoupled classifiers) or considering the observation of different subgroups (multitask learning).[6] [138] have taken this approach a little further and create an ensemble of ensembles where each base level ensemble operates on a given protected group. Note that sufficient data is required for each subgroup for this method of sampling to not negatively affect performance and fairness, as shown by [9], and that outliers can be problematic [45, 211]. For each subgroup, an individual classifier is trained.

A key challenge in decoupled classifiers is the selection of groups: some can be rarer than others [265] and as such a balance is needed to ensure groups are as atomic as possible but robust against gerrymandering [130, 161]. Thus, different candidate groupings are often evaluated via in- or post-processing methods to inhibit overfitting, maximize some fairness metric(s), and/or prevent other theoretical violations. Common approaches in group formation are recursive partitioning (e.g. [276, 161]) and clustering (e.g.: [67, 138]) as (good) clusters well approximate underlying data distributions and subgroups. [138] used clustering as a means to build stratified samples of different subgroups within the data as an exercise in bagging for the training of fair ensembles.

Subgroup analysis can also be a useful exercise in model evaluation [67, 4, 291] often defining quantities to measure how models affect different subgroups. For example, to analyze if one model is more discriminatory than another to some observed subgroup, or to identify how variable omission (see subsection 4.1) affects model fairness. Statistical hypothesis testing is employed to reveal whether models are significantly different with respect to fairness quantities or denote variable instability, i.e. when a model is not robustly fair when a given variable or set of variables are included. These methods can also treat previously trained ML models as a black-box [67, 4]. See [200] for an example set of statistical tests to indicate the likelihood of fair decisions. Probabilistic verification (e.g. [18]) of fairness measures via the sampling of sensitive variables has also been proposed to evaluate a trained model within some (small) confidence bound. Similar to other evaluation approaches (e.g. [10]), these approaches present fairness as a dichotomous outcome: a model is fair, or it isn't, as opposed to quantifying how (un)fair a model is. Yet, this is still a useful (and scalable) means to quickly evaluate different models against a number of fairness metrics.

4.4 Transformation

Transformation approaches learn a new representation of the data, often as a mapping or projection function, in which fairness is ensured, but still preserving the fidelity of the ML task [96]. Current transformation approaches operate mainly on numeric data, which is a significant limitation [96]. There are different perspectives to transforming the training data: operating on the dependent variable (e.g. [86]), operating on numeric non-sensitive variables (e.g. [96, 191, 51]), mapping individuals to an input space which is independent of specific protected subgroupings (e.g. [87, 288, 54, 176, 117, 143, 287]), and transforming the distribution of model predictions in accordance to specific fairness objectives (e.g. [142, 272]). There are parallels between blinding (in the immunity sense) and independence mappings, as in many ways these two approaches share a common goal: creating independence from one or more specific sensitive variables. Other forms of transformation include relabelling and perturbation, but we consider these a class of their own (see: subsection 4.5).

To provide an example for transformation, [96] discuss transforming the distribution of SAT scores towards the median to "degender" the original distribution into a distribution which retains only the rank order for individuals independent of gender. This is essentially removing information about protected variables from a set of covariates. Transformation approaches often seek to retain rank orders within transformed variables in order to preserve predictive ability. [51,

[6] Whilst not a sampling method, learning proxy groupings [123] is similar in intent to applications of multitask learning.

191] define a similar approach yet model the transformation process with different assumptions and objectives. An alternative to retaining rank order is the use of distortion constraints (e.g. [86]) which seek to prevent mapping "high" values to "low" values and vice versa. There is an inherent trade-off between the degree of transformation (fairness repair) and the effect on classifier performance [96]. Approaches to combat this often resort to partial repair: transforming the data towards some target distribution, but not in its entirety seeking to balance this trade-off (e.g. [96, 117]).

Although largely a pre-processing method, transformation can also been applied within a post-processing phase. [62, 162, 208] transform the output of a classifier in accordance to the identification of unfair causal pathways, either by averaging [208], constraining the conditional distribution of the decision variable [162] or through counterfactual correction [62]. As an approach, this is similar to the idea of thresholding (see subsection 4.11) and calibration (see subsection 4.10).

There are a number of challenges to consider when applying transformation techniques: 1) the transformed data should not be significantly different from the original data, otherwise the extent of "repair" can diminish the utility of the produced classifier [191, 96, 86] and, in general, incur data loss [117]; 2) understanding the relationships between sensitive and potential proxy variables [96], as such ML researchers may wish to use causal methods to first understand these relationships prior to the application of transformation methods; 3) the selection of "fair" target distributions is not straightforward [86, 296, 117]; 4) finding an "optimal" transformation often requires an optimisation step, which under high dimensionality can be computationally expensive, even under assumptions of convexity [86]; 5) missing data provides specific problems for transformation approaches, as it is unclear how to deal with such data samples. Many handle this by simply removing these samples, yet this may raise other methodological issues; 6) transformation makes the model less interpretable [182, 191], which may be at odds with current data protection legislation; and 7) there are no guarantees that a transformed data set has "repaired" potentially discriminatory latent relationships with proxy variables [54]. In this case, causal methods (see subsection 4.2) may help.

4.5 Relabelling and Perturbation

Relabelling and perturbation approaches are a specific subset of transformation approaches: they either flip or modify the dependent variable (relabelling; e.g.[76, 156, 155, 157, 192, 158, 51]), or otherwise change the distribution of one or more variables in the training data directly (perturbation; e.g. [125, 141, 273]). Referred to as data-massaging by [288, 155], relabelling involves the modification of the labels of training data instances so that the proportion of positive instances are equal across all protected groups. It can also be applied to the test data upon the basis of strategies or probabilities learned on the training data. Often, but not always, approaches seek to retain the overall class distribution, i.e. the number of positive and negative instances remains the same. For example, [192] relabel the dependent variable (flip it from positive to negative or vice versa) if the data instance is determined as being discriminated against with respect to the observed outcome. Relabelling is also often part of counterfactual studies (e.g. [131, 151, 273]) which seek to investigate if flipping the dependent variable or other categorical sensitive variables affect the classification outcome.

Perturbation often aligns with notions of "repairing" some aspect(s) of the data with regard to notions of fairness. Applications in perturbation-based data repair [125, 163, 141, 96, 143, 117] have shown that accuracy is not significantly affected through perturbation. Whilst there are a number of papers that harness perturbation (it is not always referred to as perturbation) in the ML literature, this approach appears more prevalent in the discrimination-aware data mining literature where it is often used as a means of privacy preservation. Often perturbation-based approaches are applied as a pre-processing step to prepare for an in-processing approach; often reweighing (e.g.[141]), and/or regularisation / optimization (e.g. [273, 265]). It has also been proposed as a mechanism to detect proxy variables and variable influence [272] and counterfactual distributions [273].

Closely related to perturbation as a means to "repair" data is the use of sensitivity analysis (see [239]) to explore how various aspects of the feature vector affect a given outcome. This is a relatively under-addressed area in the fairness literature, yet it has been well motivated (although perhaps indirectly): [116] called for a better understanding of bias stemming from uncertainty, and [129] who stressed that assessment of data reliability is needed. The application of sensitivity analysis in ML is often to measure the stability of a model [227]. Whilst a number of approaches exist to determine model stability [44, 228, 175, 246, 39, 160], it has rarely been applied to ML research beyond notions of model convergence and traditional performance measures with similar objectives to cross-validation. Yet, relabelling and perturbation are not far from the principals of sensitivity analysis. [76] proposed the perturbation of feature vectors to measure the effect on model performance of specific interventions. [199, 107] investigated visual mechanisms to better display "issues" with data to users, yet these approaches generally lack support for novice users [24]. [151, 149] used sensitivity analysis to evaluate sensitive variables and their relationship(s) with classification outcomes. Whilst sensitivity analysis is not a method to improve fairness (thus omitted from Figures 3-5) it can help to better understand uncertainty with respect to fairness.

As for transformation, modification of the data via relabelling and perturbation is not always legally permissible [16] and changes to the data should be minimised [192, 125]. [172] also note that some classifiers may be unaffected by the presence or specific nuances of some biases, and others may be negatively affected by altering the training data in an attempt to mitigate them. Thus, it is important to continuously (re)assess any fairness and methodological decisions made.

4.6 Reweighing

Unlike transformation, relabelling, and perturbation approaches which alter (certain instances of) the data, reweighing assigns weights to instances of the training data while leaving the data itself unchanged. Weights can be introduced for multiple purposes: 1) to indicate a frequency count for an instance type (e.g. [51]), 2) to place lower/higher importance on "sensitive" training samples (e.g. [52, 155, 141]), or 3) to improve classifier stability (e.g. [172]). Reweighing as an approach straddles the boundary between pre-processing and in-processing. For example, [155] seek to assign weights that take into consideration the likelihood of an instance with a specific class and sensitive value pairing (pre-processing). Whereas, [172] first build an unweighted classifier, learn the weights of samples, then retrain their classifier using these weights (in-processing). A similar approach is taken by [141] who identify sensitive training instances (pre-processing), but then learn weights for these instances (in-processing) to optimize for the chosen fairness metric.

With appropriate sampling (see subsection 4.3), reweighing can maintain high(er) accuracy when compared to relabelling and blinding (omission) approaches [155]. However, as [172, 109] note, classifier stability and robustness can be an issue. Thus ML researchers need to carefully consider how reweighing approaches are applied and appropriately check for model stability. Reweighing also subtly changes the data composition making the process less transparent [182, 191].

4.7 Regularization and Constraint Optimisation

Classically, regularization is used in ML to penalize the complexity of the learned hypothesis seeking to inhibit overfitting. When applied to fairness, regularization methods add one or more penalty terms which penalize the classifier for discriminatory practices [158]. Thus, it is not hypothesis (or learned model) driven, but data driven [19] and based upon the notion(s) of fairness considered. Much of the literature extends or augments the (convex) loss function of the classifier with fairness terms usually trying to find a balance between fairness and accuracy (e.g. [158, 111, 142, 132, 51, 57, 195, 290, 26, 7, 96]). Some notable exceptions are approaches which emphasise empirical risk subject to fairness constraints or welfare conditions (e.g. [85, 131]), TPR/FPR of protected groups (e.g. [19]), stability of fairness (e.g. [136]), or counterfactual terms (e.g. [81]).

[136, 103] note that often approaches for fair ML are not stable, i.e. subtle changes in the training data significantly affect performance (comparatively high standard deviation). [136] argue that stability of fairness can be addressed through regularization and present corresponding empirical evidence through extensions of [158, 284]. Aside from this, [288, 109] note that regularization as a mechanism is fairly generic, and can lead to a lack of model robustness and generalizability.

In-processing (constraint) optimization approaches (e.g. [5, 57, 65, 124, 195, 209, 210, 288, 112, 208, 75, 284, 164]) have similar objectives to fairness regularization approaches, and hence we present them together. Constraint optimization approaches often include notions of fairness[7] in the classifier loss function operating on the confusion matrix during model training. [209] also approached this via reinforcement learning. Yet, these approaches can also include other constraints, and/or reduce the problem to a cost-sensitive classification problem (e.g. [112, 5, 210]). Similarly, a multi-fairness metric approach has been proposed by [164] where adaptions to stochastic gradient descent optimize weighted fairness constraints as an in-processing, or post-processing (when an pre-trained classifier is used) scenario. [210, 112, 75] summarizes a number of additional constraints types as precision or budget constraints to address the accuracy fairness trade off (often expressed as utility or risk functions e.g. [73, 124]); quantification or coverage constraints to capture disparities in class or population frequencies; churn constraints capturing online learning scenarios and enforcing that classifiers do not differ significantly from their original form as defined by the initial training data; and, stability constraints akin to the observations of [136, 103].

Key challenges for regularization approaches are: 1) they are often non-convex in nature or achieve convexity at the cost of probabilistic interpretation [111]; 2) not all fairness measures are equally affected by the strength of regular-

[7] We also note that many constraint optimisation papers often define new notions of fairness.

ization parameters [81, 26]; and 3) different regularization terms and penalties have diverse results on different data sets, i.e. this choice can have qualitative effects on the trade-off between accuracy and fairness [26]. For constraint optimization it can be difficult to balance conflicting constraints leading to more difficult or unstable training [75].

4.8 Adversarial Learning

In adversarial learning the objective is for an adversary to try and determine whether a model training algorithm is robust enough. The framework of [115] helped popularize the approach through the process of detecting falsified data samples [55]. When applied to applications of fairness in ML, an adversary instead seeks to determine whether the training process is fair, and when not, feedback from the adversary is used to improve the model [55]. Most approaches in this area use notions of fairness within the adversary to apply feedback for model tuning as a form of in-processing where the adversary penalizes the model if a sensitive variable is predictable from the dependent variable (e.g. [271, 28, 290, 89, 278, 55, 30]). This is often formulated as a multi-constraint optimization problem considering many of the constraints as discussed in subsection 4.7. There has also been work proposing the use of an adversary as a pre-processing transformation process on the training data (e.g. [277, 194, 97, 278, 3, 152]) with similar objectives to transformation as discussed in subsection 4.4 yet often moving towards a notion of "censoring" the training data with similar objectives to variable blinding as discussed in subsection 4.1. Work has also started applying the notions of causal and counterfactual fairness to adversarial learning (e.g. [278]). Here, the causal properties of the data prior to and after intervention are modeled with the adversarial intention to optimize a set of fairness constraints towards improved interventions.

An advantage of adversarial approaches is that they can consider multiple fairness constraints [271], often treating the model as a blackbox [194]. However, adversarial approaches have been reported to often lack stability which can make them hard to train reliably [30, 97] and also specifically in some transfer learning scenarios [194] when, for example, the protected variable is known only for a small number of samples. Additional forms of regularization have been proposed to try and address these issues (e.g. [30]). The use of generative adversarial networks (GAN) with fairness considerations also permit applications within unstructured (for example multimedia) data or more generally as a generative process of creating an "unbiased" dataset using a number of samples. [241] illustrate this by using a GAN with fairness constraints to produce "unbiased" image datasets and [278] have evidenced similar results for structured data.

4.9 Bandits

The application of bandits to ML fairness [93, 146, 147, 189, 108] is nascent. As yet, papers have proofs of their work but lack general evaluation against specific data. As a reinforcement learning framework, bandits are motivated on the need for decisions to be made in an online manner [145], and that decision makers may not be able to define what it means to be "fair" but that they may recognize "unfairness" when they see it [108]. Approaches that use bandits do so on the basis of [87]'s notion of individual fairness, i.e. that all similar individuals should be treated similarly. Thus, bandit approaches frame the fairness problem as a stochastic multi-armed bandit framework, assigning either individuals to arms, or groups of "similar" individuals to arms, and fairness quality as a reward represented as regret [145, 189]. The two main notions of fairness that have emerged from the application of bandits are meritocratic fairness [145, 147] (group agnostic) and subjective fairness [189] (emphasises fairness in each time period t of the bandit framework).

4.10 Calibration

Calibration is the process of ensuring that the proportion of positive predictions is equal to the proportion of positive examples [79]. In the context of fairness, this should also hold for all subgroups (protected or otherwise) in the data [223, 291, 60]. Calibration is particularly useful when the output is not a direct decision but used to inform human judgement when assessing risks (e.g. awarding a loan) [212]. As such, a calibrated model does not inhibit biases of decision makers, but rather ensure that risk estimates for various (protected) subgroups carry the same meaning [223]. However calibrating for multiple protected groups and/or using multiple fairness criteria at once has been shown to be impossible [66, 167, 188, 130, 189, 223, 212, 164]. [223] even note that the goals of low error and calibration are competing objectives for a model. This occurs as calibration has limited flexibility [141]. [276] also evidenced that decoupling the classifier training from the means to increase fairness, i.e. post-processing, is provably sub-optimal.

The literature has proposed various approaches to handle the impasse of achieving calibration and other fairness measures. One approach has been to apply a randomization post-processing process to try and achieve a balance between accuracy and fairness, yet [167, 72, 129, 223, 212] discuss a number of shortcomings of this approach. Notably, the individuals who are randomized are not necessarily positively impacted, and the overall accuracy of the model can be adversely affected. [212] also note that this approach is Pareto sub-optimal, and instead propose a cost-based approach to balance calibration and error parity. [188] suggest that calibration is sufficient as a fairness criterion if the model is unconstrained. [130] instead seek to achieve approximate calibration (i.e. to guarantee calibration with high probability) using a multi-calibration approach that operates on identifiable subgroups to balance individual and group fairness measures even for small samples: a specific challenge for achieving calibration [188]. [161, 165] undertake a similar approach under different settings. [189] have proposed a bandit-based approach to calibration.

4.11 Thresholding

Thresholding is a post-processing approach which is motivated on the basis that discriminatory decisions are often made close to decision making boundaries because of a decision maker's bias [157] and that humans apply threshold rules when making decisions [169]. Thresholding approaches often seek to find regions of the posterior probability distribution of a classifier where favored and protected groups are both positively and negatively classified.[8] Such instances are considered to be ambiguous, and therefore potentially influenced by bias [157]. To handle this, researchers have devised approaches to determine threshold values via measures such as equalized odds specifically for different protected groups to find a balance between the true and false positive rates to minimize the expected classifier loss [129]. The underlying idea here is to incentivize good performance (in terms of both fairness and accuracy) across all classes and groups.

Computing the threshold value(s) can be undertaken either by hand to enable a user to assert preferences with respect to the fairness accuracy trade-off or use other statistical methods. [202] estimate the thresholds for each protected group using logistic regression, then use a fairness frontier to illustrate disalignment between threshold values. [157] use an ensemble to identify instances in an uncertainty region and assist in setting a threshold value. [100] propose a method to shift decision boundaries using a form of post-processing regularization. [266] use posterior sampling to maximize a fairness utility measure. [138] learn a threshold value after training an ensemble of decoupled ensembles (a pre-processing sampling approach, see subsection 4.3) such that the discrepancy between protected and non-protected groups is less than some user specified discrimination threshold value. Thus, one of the key challenges for thresholding approaches is to determine preferences with respect to a tolerance for unfairness. [138] note that this is often undertaken with respect to accuracy, but in many cases class imbalance would invalidate such decisions. Thresholding is often argued as a potential human-in-the-loop mechanism, yet in the absence of appropriate training programs (for the human-in-the-loop) this can introduce new issues [268]. This stems from fairness typically not being representable as a monotone function, therefore assigning a threshold value may be quite arbitrary [252]. Thresholding approaches often claim compelling notions of equity, however, only when the threshold is correctly chosen [72].

5 Beyond Binary Classification

The bulk of the fairness literature focuses on binary classification [26]. In this section, we provide an overview and discussion beyond approaches for binary classification (albeit less comprehensive) and note that there is a sufficient need for fairness researchers to also focus on other ML problems.

5.1 Fair Regression

The main goal of fair regression is to minimize a loss function $l(Y, \hat{Y})$, which measures the difference between actual and predicted values, while also aiming to guarantee fairness. The general formulation is similar to the case of (binary) classification, with the difference that Y and $\hat{Y}$ are continuous rather than binary or categorical. Fairness definitions for regressions adapt principles defined in Section 3. For example, parity-based metrics aim to make the loss function equal for different groups [6]. With respect to defining fairness metrics or measurements, [27] suggest several metrics that can be used for general regression models. [6] define both statistical parity and bounded-group-loss metrics to measure fairness in regression, the latter providing a customizable maximum allowable loss variable that defines a specific trade-off between fairness and loss (and thus predictor performance). [50] consider biases in linear regression as measured by the effects of a sensitive attribute on Y through the mean difference (difference of mean target variable between groups) and AUC metrics. They suggest the use of propensity modeling as well as additional constraints (e.g., enforcing a mean difference of zero) to mitigate biases in linear regression.

The effect of bias on parameter estimates and coefficients in multiple linear regression is discussed by [144], who also suggest a post-processing approach to make parameter estimates impartial with respect to a sensitive attribute. [171] include fairness perspectives in non-convex optimization for (linear) regression using the coefficient of determination

[8] We note that there is a fine line between thresholding and calibration approaches and that they often overlap.

between the predictions $\hat{y}$ and the sensitive attribute(s) as additional constraints in their (constrained) linear least squares model that generates a solution for a user-selected maximum level for the coefficient of determination.

[218] propose methods for fair regression as well as fair dimensionality reduction using a Hilbert Schmidt independence criterion and a projection-based methodology that is able to consider multiple sensitive attributes simultaneously. [158] suggest a regularization approach that can be applied to general prediction algorithms. [104] define the concept of μ-neutrality that measures if probabilistic models are neutral with respect to specific variables and show that this definition is equivalent to statistical parity. [26] propose a family of regularization approaches for fair regression that works with a variety of group and individual fairness metrics. Through a regularization weight, the proposed method is able to calculate accuracy-fairness trade-offs and evaluate the efficient frontier of achievable accuracy-fairness trade-offs. [102] consider group-based fairness metrics and their inclusion in kernel regression methods such as decision tree regression while keeping efficiency in computation and memory requirements.

5.2 Recommender Systems

Considerations of fairness have been actively studied in the context of rankings and recommender systems. For rankings in general, [280], [286], and [32] define different types of fairness notions such as group-based fairness in top-k ranking ([280, 286]), an individual fairness measure in rankings following concepts similar to [129] and the equality of opportunity in binary classification [32], and unfairness of rankings over time through a dynamic measure called amortized fairness [32].

For recommender systems in particular, [179] consider fairness-aware loan recommendation systems and argue that fairness and recommendation are two contradicting tasks. They measure fairness as the standard deviation of the top-N recommendations, where a low standard deviation signifies a fair recommendation without compromising accuracy. Subsequent publications expanded this view of recommender fairness by proposing additional metrics as well as algorithms [281, 294, 254, 29, 58]. These include a set of ML inspired group-based fairness metrics that address different forms of unfairness to address potential biases in collaborative filtering recommender systems stemming from a population imbalance or observation bias [281], fairness goals for recommender systems as overcoming algorithmic bias and making neutral recommendations independent of group membership (e.g., based on gender or age) [294], recommendation calibration, i.e., the proportional representation of items in recommendations, [254], pairwise fairness as well as a regularization approach to improve model performance [29], and two fairness measures in top-k recommendations, proportional representation and anti-plurality [58]. Further approaches include tensor-based recommendations have been proposed that take statistical parity into account [294], and a mechanism design approach for fairly dividing a set of goods between groups using disparate impact as fairness measure and a recommender system as evaluation use case [219].

An aspect that sets fairness considerations in recommender systems apart from binary classification in ML is that fairness can be seen as multi-sided concept that can be relevant for both users (who consume the recommendations) and items. [47, 48] introduce the notion of “C-fairness” for fair user/consumer recommendation (user-based), and “P-fairness” for fairness of producer recommendation (item-based) to address this multi-sided aspect, showing that defining generalized approaches to multi-sided fairness is hard due to the domain specificity of the multi-stakeholder environment. [91] presents an empirical analysis of P-fairness for several collaborative filtering algorithms. Similarly, [292] aim to find an optimal trade-off between the utilities of the multiple stakeholders. Other work considering fairness from either the consumer or provider side include the analysis of different recommendation strategies for a variety of (fairness) metrics [140], subset-based evaluation metrics to measure the utility of recommendations for different groups (e.g., based on demographics) [90], and a general framework to optimize utility metrics under fairness of exposure constraints [247, 248]. Besides the previous work on metrics and algorithms, several authors have also proposed bias mitigation strategies. This includes a pre-processing approach (recommendation independence) to make recommendations independent of a specific attribute [159], adding specifically designed “antidote” data to the input instead of manipulating the actual input data to improve the social desirability of the calculated recommendations [229], and a post-processing algorithm to improve user-based fairness through calibrated recommendations [254].

5.3 Unsupervised Methods

Currently, unsupervised methods fall into three distinct areas: 1) fair clustering (e.g. [243, 64, 22, 232, 14, 23, 22, 61]); 2) investigating the presence and detection of discrimination in association rule mining (e.g. [217, 216, 125]); and 3) transfer learning (e.g. [88, 74]).

Fair clustering started with the initial work of [64] who introduced the idea of micro-cluster fairlet decomposition as a pre-processing stage applied prior to standard centroid-based methods like k-means and k-medians. Thus far, clustering approaches have mostly operating on [96]’s introduction of disparate impact introducing this as cluster balance, where balance pertains to uniformity of distribution over k clusters of belonging to some protected group. [64] use color to represent belonging to the protected group or not. When multiple protected groups are in place this means optimising for both the number of clusters, and the number as well as spread of colors. This is undertaken by [22] who extend the work of [64] to allow for more than two colors and fuzzy cluster membership functions arguing that otherwise the approach is too stringent and brittle. Yet, there is a cost here, unlike other approaches to fairness in ML, fair clustering has significant computational costs associated to it. However, methods have emerged to handle this via coresets [243] and approximate fairlet decomposition [14]. Fair clustering has also seen applications in the the discovery of potentially protected groups (e.g. [21, 65]).

Approaches that utilize transfer learning do so in combination with other methods. The motivation for using transfer learning is typically in response to an observable covariate shift between the source (training) and target distributions. This can often occur in real world application settings and requires that the model is trained on a different probability distribution to that which the model will ultimately be tested (and later deployed) on [31, 255, 225]. Here, transfer learning acts as an unsupervised domain adaption technique to account for such covariate shifts [113, 263, 88]. In this, transfer learning approaches are somewhat analogous to reweighing approaches in that they seek to determine weights for each training example that account for a covariate shift optimized using regularization techniques (e.g. [74]) or forms of joint loss functions (e.g. [88]).

5.4 Natural Language Processing

Natural language processing (NLP) is an area of machine learning which operates on text data, e.g. document classification, sentiment analysis, text generation, translation, etc. Unintended biases have also been noticed in NLP; these are often gender or race focused [38, 215, 82, 84, 166, 293, 53, 293]. [38] highlighted that word embeddings (often used for various tasks ranging from Google news sorting to abusive content filtering) can be biased against women. [166] notes that sentiment analysis systems can discriminate against races and genders noting that specific race or gender references in text can result negative sentiment whereas different race or gender references the same text is noted as positive. Unintended bias can be introduced to the data by personal biases during manual data labelling [84] as well as biases that occur in language use for example through non-native speaker errors [293, 258, 35] or how the text data is prepared as well as the general model architecture [215]. In terms of approaches, combating biases in NLP occur mostly in the pre-processing stage, such as removing or replacing specific words (e.g. [82]), dictionary modification (e.g. [293]), and unsupervised balancing of the training dataset [84]. However, in some cases mitigation of biases is not attempted due to the complexity of the task [293, 166].

6 Current Platforms

While many researchers publish their individual approaches on github and similar platforms, a few notable projects have emerged that address fairness in ML from a more general perspective. Table 4 describes some of these approaches. We note the existence of proprietary software, yet here emphasize readily tools to discuss the current state of the art for ML researchers and practitioners.

Project	Features
AIF360 [20]	Set of tools that provides several pre-, in-, and post-processing approaches for binary classification as well as several pre-implemented datasets that are commonly used in Fairness research
Fairlean[9]	Implements several parity-based fairness measures and algorithms [129, 5, 6] for binary classification and regression as well as a dashboard to visualize disparity in accuracy and parity.
Aequitas [235]	Open source bias audit toolkit. Focuses on standard ML metrics and their evaluation for different subgroups of a protective attribute.
Responsibly [190]	Provides datasets, metrics, and algorithms to measure and mitigate bias in classification as well as NLP (bias in word embeddings).
Fairness[10]	Tool that provides commonly used fairness metrics (e.g., statistical parity, equalized odds) for R projects.
FairTest [264]	Generic framework that provides measures and statistical tests to detect unwanted associations between the output of an algorithm and a sensitive attribute.
Fairness Measures[11]	Project that considers quantitative definitions of discrimination in classification and ranking scenarios. Provides datasets, measures, and algorithms (for ranking) that investigate fairness.
Audit AI[12]	Implements various statistical significance tests to detect discrimination between groups and bias from standard machine learning procedures.
Dataset Nutrition Label [134]	Generates qualitative and quantitative measures and descriptions of dataset health to assess the quality of a dataset used for training and building ML models.
ML Fairness Gym	Part of Google's Open AI project, a simulation toolkit to study long-run impacts of ML decisions.[13] Analyzes how algorithms that take fairness into consideration change the underlying data (previous classifications) over time (see e.g. [187, 92, 94, 135, 204]).

Table 4: Overview of projects addressing Fairness in Machine Learning.

7 Concluding Remarks: The Fairness Dilemmas

In this article, we have provided an introduction to the domain of fairness in Machine Learning (ML) research. This encompasses a general introduction (Section 2), different measures of fairness for ML (Section 3), methods to mitigate bias and unfairness in binary classification problems (Section 4) as well as beyond binary classification (Section 5) and listed some open source tools (Section 6) to assist researchers and practitioners seeking to enter this domain or employ state of the art methods within their ML pipelines. For specific methods, we have noted the key challenges of their deployment. Now, we focus on the more general challenges for the domain as a whole as a set of four dilemmas for future research (the ordering is coincidental and not indicative of importance): **Dilemma-1:** Balancing the trade-off between fairness and model performance (subsection 7.1); **Dilemma-2:** Quantitative notions of fairness permit model optimization, yet cannot balance different notions of fairness, i.e individual vs. group fairness (subsection 7.2); **Dilemma-3:** Tensions between fairness, situational, ethical, and sociocultural context, and policy (subsection 7.3); and **Dilemma-4:** Recent advances to the state of the art have increased the skills gap inhibiting "man-on-the-street" and industry uptake (subsection 7.4).

7.1 Dilemma 1: Fairness vs. Model Performance

A lack of consideration for the sociocultural context of the application can result in ML solutions that are biased, unethical, unfair, and often not legally permissible [36, 283]. The ML community has responded with a variety of mechanisms to improve the fairness of models as outlined in this article. However, when implementing fairness measures, we must emphasize either fairness or model performance as improving one can often detriment the other [27, 87, 73, 129, 296, 54, 124]. [96] do, however, note that a reduction in accuracy may in fact be the desired result, if it was discrimination in the first place that raised accuracy. Note that even prior to recognizing this trade-off, we need to be cautious in our definition of model performance. ML practitioners can measure performance in a multitude of ways, and there has been much discussion concerning the choice of different performance measures and approaches [80, 250]. The choice of performance measure(s) itself may even harbor, disguise, or create new underlying ethical concerns. We also note that currently, there is little runtime benchmarking of methods outside of clustering approaches (see [243, 14]). This is an observation as opposed to a criticism, but we note that potential users of fairness methods will likely concern themselves with computational costs, especially if they increase.

7.2 Dilemma 2: (Dis)agreement and Incompatibility of "Fairness"

On top of the performance trade-off, there is no consensus in literature whether individual or group fairness should be prioritized. Fairness metrics usually either emphasize individual or group fairness, but cannot combine both [170, 66]. [252] also note that many approaches to group fairness often tackle between-group issues, as a consequence they demonstrate that within-group issues are worsened through this choice. To further complicate things, [72, 118] argue that with a reliance on expressing fairness mathematically these definitions often do not map to normative social, economic, or legal understandings of the same concepts. This is corroborated by [249] who note an over-emphasis in the literature on specific measures of fairness and insufficient dialogue between researchers and affected communities. Thus, improving fairness in ML is challenging and simultaneously there are many different notions for researchers and practitioners to navigate. Further adding to this discussion is the notion of differing views of the root(s) of fairness and bias. [221, 220, 121, 253] study the differing views of people in this regard and observe that this is not a trivial challenge to address. E.g., [221] notes that women have differing views in the inclusion / exclusion of gender as a protected variable to men. [137] note that a similar discussion was left unresolved in the early days of fairness research in the context of test scores and employment/hiring practices, indicating that this is one of the main challenges of ML fairness research in the future. [154] have noted that this dilemma can be articulated as a bias in, bias out property of ML: i.e. addressing one form of bias results in another.

Thus, the community as articulated in [60, 130, 46] needs to explore ways to either handle combinations of fairness metrics, even if only approximately due to specific incompatibilities, or implement a significant meta review of measures to help categorise specific differences, ideological trade-offs, and preferences. This will enable researchers and practitioners to consider a balance of the fairness measures they are using. This is a challenging undertaking and

[8] `https://github.com/fairlearn/fairlearn`

[9] `https://github.com/kozodoi/Fairness`

[10] `https://github.com/google-research/tensorflow_constrained_optimization`

[11] `https://github.com/pymetrics/audit-ai`

[12] `http://www.fairness-measures.org/`, `https://github.com/megantosh/fairness_measures_code/tree/master`

[13] `https://github.com/google/ml-fairness-gym`

whilst the tools discussed in Section 6 go some way to facilitate this, there is a need for more general toolkits and methodologies for comparing fairness approaches. We note a number of comparative studies, i.e. [103, 106, 264], but these only scratch the surface.

7.3 Dilemma 3: Tensions with Context and Policy

The literature typically hints toward "optimizing" fairness without transparency of the root(s) of (un)fairness [182] rarely extending beyond "(un)fair" [76, 252] typically to mirror current legal thought [96]. This is true for both metrics and methods. As such, platforms are needed to assist practitioners in ascertaining the cause(s) of unfairness and bias. However, beyond this, critics of current research [49, 249, 268, 269, 181, 283, 275, 275] argue that efforts will fail unless contextual, sociocultural, and social policy challenges are better understood. Thus, there is an argument that instead of striving to "minimize" unfairness, more awareness of context-based aspects of discrimination is needed. There is the prevalent assumption that "unfairness" has a uniform context-agnostic egalitarian valuation function for decision makers when considering different (sub)populations [33, 73, 72]. This suggests a disconnect between organizational realities and current research, which undermines advancements [269, 186]. Other suggestions have been for ML researchers and practitioners to better understand the limitations of human decision making [230].

It is easy to criticize, however, the underlying challenge is availability of realistic data. Currently, the literature relies unilaterally on convenience datasets (enabling comparative studies), often from the UCI repository [12] or similar with limited industry context and engagement [269, 268, 181]. [163, 276, 177, 178, 153] note that there is an additional challenge in the datasets used to train models: data represent past decisions, and as such inherent bias(es) in these decisions are amplified. This is a problem referred to as selective labels [178]. Similarly, there may be differences in the distribution(s) of the data between the data the model is trained on, and deployed on: dataset shift as discussed by [225]. As such, data context cannot be disregarded.

Thus, researchers need to better engage with (industry) stakeholders to study models in vivo and engage proactively in open debate on policy and standardization. This is a hard problem to solve: companies cannot simply hand out data to researchers and researchers cannot fix this problem on their own. There is a tension here between advancing the fairness state of the art, privacy [251, 105], and policy. [269] notes that policy makers are generally not considered or involved in the ML fairness domain. We are seeing an increasing number of working groups on best practices for ethics, bias, and fairness, where Ireland's NSAI/TC 002/SC 18 Artificial Intelligence working group, the IEEE P7003 standardization working group on algorithmic bias, and the Big Data Value Association are just three examples of many, but this needs to be pushed harder at national and international levels by funding agencies, policy makers, and researchers themselves.

7.4 Dilemma 4: Democratisation of ML vs the Fairness Skills Gap

Today, ML technologies are more accessible than ever. This has occurred through a combination of surge in third level courses and the wide availability of ML tools such as WEKA [128], RapidMiner[15], and SPSS Modeler[16]. Alternatively, Cloud-based solutions such as Google's Cloud AutoML [34], Uber AI's Ludwig,[17] and Baidu's EZDL[18] remove the need to even run models locally. The no-/low-code ML movement is arguably enabling more companies to adopt ML technologies. In addition, there is a growing trend in the use of Automated Machine Learning (AutoML) [261, 99] to train ML models. AutoML abstracts much of the core methodological expertise (e.g., KDD [95], and CRISP-DM [59]) by automated feature extraction and training multiple models often combining them into an ensemble of models that maximizes a set of performance measures. Collectively, each of these advancements positively democratizes ML, as it means lower barriers of use: "push button operationalization" [13] with online marketplaces[19] selling third party ML solutions.

Lowering the entry barrier to ML through democratization will (if it hasn't already) mean an increase in (un)intentional socially insensitive uses of ML technologies. The challenge is that ML application development follows a traditional software development model: it is modular, sequential, and based on large collections of (often) open source libraries, but methods to highlight bias, fairness, or ethical issues assume high expertise in ML development and do not consider "on-the-street" practitioners [269, 182]. This was our motivation in writing this survey. However, the fairness domain is only just starting to provide open source tools available for practitioners (Section 6). Yet, in general

[15]`https://rapidminer.com`

[16]`https://www.ibm.com/ie-en/products/spss-modeler`

[17]https://uber.github.io/ludwig/

[18]`https://ai.baidu.com/ezdl/`

[19]E.g.: Amazon's `https://aws.amazon.com/marketplace/solutions/machinelearning/` and Microsoft's `https://gallery.azure.ai` ML Marketplaces.

there is little accommodation for varying levels of technical proficiency, and this undermines current advancement [249, 116, 269, 268, 49]. There is a tension between educational programs (as called for in [49]) and the degree of proficiency needed to apply methods and methodologies for fair ML. [268, 72] have advocated this as the formalization of exploratory fairness analysis: similar to exploratory data analysis, yet for informed decision making with regard to "fair" methodological decisions. Similarly, [240] call for core ML educational resources and courses to better include ethical reasoning and deliberation and provide an overview of potential materials. Thus, the fourth dilemma currently facing the fair ML domain is its own democratization to keep up with the pace of ML proliferation across sectors. This means a shift in terms of scientific reporting, open source comprehensive frameworks for repeatable and multi-stage (i.e., pipelined models) decision making processes where one model feeds into another [40, 116, 259]. Currently under-addressed is bias and fairness transitivity: where one ML model is downstream to another.

7.5 Concluding Remarks

The literature almost unilaterally focuses on supervised learning with an overwhelming emphasis on binary classification [26]: diversification is needed. With very few exceptions, the approaches discussed in this article operate on the assumption of some set of (usually a priori known) "protected variables". This doesn't help practitioners. Tools potentially based on causal methods (subsection 4.2) are needed to assist in the identification of protected variables and groups as well as their proxies.

More realistic datasets are needed: [227] argue that approaches tend to operate on too small a subset of features raising stability concerns. This should go hand in hand with more industry-focused training. Tackling fairness from the perspective of protected variables or groups needs methodological care, as "fixing" one set of biases may inflate another [33, 72] rendering the model as intrinsically discriminatory as a random model [223, 88]. There is also the risk of redlining, where although the sensitive attribute is "handled" sufficiently, correlated variables are still present [217, 231, 285, 268, 87, 54], amplifying instead of reducing unfairness [87].

We also note specific considerations of pre-processing vs. in-processing vs. post-processing interventions. Pre-processing methods, which modify the training data, are at odds with policies like GDPR's right to an explanation, and can introduce new subjectivity biases [268]. They also assume sufficient knowledge of the data, and make assumptions over its veracity [72]. Uptake of in-processing approaches requires better integration with standard ML libraries to overcoming porting challenges. [276] noted that generally post-processing methods have suboptimal accuracy compared to other "equally fair" classifiers, with [5] noting that often test-time access to protected attributes is needed, which may not be legally permissible, and have other undesirable effects [60].

As a closing thought many approaches to reduce discrimination may themselves be unethical or impractical in settings where model accuracy is critical such as in healthcare, or criminal justice scenarios [60]. This is not to advocate that models in these scenarios should be permitted to knowingly discriminate, but rather that a more concerted effort is needed to understand the roots of discrimination. Perhaps, as [60, 187, 94, 276] note, it may often be better to fix the underlying data sample (e.g. collect more data, which better represents minority or protected groups and delay the modeling phase) than try to fix a discriminatory ML model.

References

[1] TENDER SPECIFICATIONS: Study on Algorithmic Awareness Building SMART 2017/0055, 2017.

[2] Communication Artificial Intelligence for Europe, apr 2018.

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

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

[5] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[6] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. *arXiv preprint arXiv:1905.12843*, 2019.

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

[8] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 439–450, 2000.

[9] Daniel Alabi, Nicole Immorlica, and Adam Kalai. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*, pages 2043–2066, 2018.

[10] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori. Fairsquare: probabilistic verification of program fairness. *Proceedings of the ACM on Programming Languages*, 1(OOPSLA):1–30, 2017.

[11] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[12] Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

[13] AzureML Team. AzureML: Anatomy of a machine learning service. In *Conference on Predictive APIs and Apps*, pages 1–13, 2016.

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

[15] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76, 2018.

[16] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019.

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

[19] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv:1707.00044*, 2017.

[20] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943*, oct 2018.

[21] Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 289–298, 2019.

[22] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems 32*, pages 4954–4965. 2019.

[23] Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. *arXiv:1811.10319*, 2018.

[24] Bettina Berendt and Sören Preibusch. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2):175–209, 2014.

[25] Richard Berk. Accuracy and fairness for juvenile justice risk assessments. *Journal of Empirical Legal Studies*, 16(1):175–194, 2019.

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

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

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

[29] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, and Et al. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2212–2220. ACM, 2019.

[30] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.

[31] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

[32] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 405–414, New York, NY, USA, 2018. Association for Computing Machinery.

[33] Reuben Binns. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*, volume abs/1712.0, pages 149–159, 2018. URL `http://arxiv.org/abs/1712.03586`.

[34] Ekaba Bisong. Google automl: Cloud vision. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 581–598. Springer, 2019.

[35] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*, 2017.

[36] Paula Boddington. *Towards a code of ethics for artificial intelligence*. Springer, 2017.

[37] Matthew T Bodie, Miriam A Cherry, Marcia L McCormick, and Jintong Tang. The law and policy of people analytics. *U. Colo. L. Rev.*, 88:961, 2017.

[38] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016.

[39] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2 (Mar):499–526, 2002.

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

[41] Danah Boyd and Kate Crawford. Six provocations for big data. In *A decade in internet time: Symposium on the dynamics of the internet and society*, volume 21. Oxford Internet Institute Oxford, UK, 2011.

[42] Danah Boyd and Kate Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[43] Tim Brennan and William L Oliver. Emergence of machine learning techniques in criminology: implications of complexity in our data and in research questions. *Criminology & Pub. Pol'y*, 12:551, 2013.

[44] Bénédicte Briand, Gilles R Ducharme, Vanessa Parache, and Catherine Mercat-Rommens. A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, 53(4):1208–1217, 2009.

[45] J Paul Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations research*, 59 (2):467–479, 2011.

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

[47] Robin Burke. Multisided Fairness for Recommendation, jul 2017.

[48] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214, New York, NY, USA, 2018. PMLR.

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

[50] T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 71–80, 2013.

[51] Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, sep 2010. ISSN 1384-5810.

[52] Toon Calders and Indrė Žliobaitė. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In *Discrimination and Privacy in the Information Society*, pages 43–57. Springer, Berlin, Heidelberg, 2013.

[53] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, aug 2017. ISSN 1095-9203.

[54] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[55] L Elisa Celis and Vijay Keswani. Improved adversarial learning for fair classification. *arXiv:1901.10443*, 2019.

[56] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *arXiv preprint arXiv:1610.07183*, 2016.

[57] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328. ACM, ACM, jun 2019.

[58] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 129–138, New York, New York, USA, 2019. ACM Press. ISBN 9781450361255.

[59] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

[60] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3539–3550. Curran Associates, Inc., 2018.

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

[62] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[63] Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.

[64] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pages 5029–5037, 2017.

[65] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. Matroids, matchings, and fairness. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2212–2220, 2019.

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

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

[68] Kevin A Clarke. The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science*, 22(4):341–352, 2005.

[69] T Cleary. Test bias: Validity of the scholastic aptitude test for Negro and White students in integrated colleges. *ETS Research Bulletin Series*, 1966, 1966. doi: 10.1002/j.2333-8504.1966.tb00529.x.

[70] T Anne Cleary. Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968. ISSN 00220655, 17453984. URL `http://www.jstor.org/stable/1434406`.

[71] Nancy S Cole. Bias in Selection. *Journal of Educational Measurement*, 10(4):237–255, 1973. ISSN 00220655, 17453984.

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

[73] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, New York, New York, USA, 2017. ACM, ACM Press.

[74] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.

[75] Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Seungil You, Karthik Sridharan, and Maya R Gupta. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019.

[76] Bo Cowgill and Catherine Tucker. Algorithmic bias: A counterfactual perspective. *NSF Trustworthy Algorithms*, 2017.

[77] Richard B Darlington. Another Look At “Cultural Fairness”. *Journal of Educational Measurement*, 8(2):71–82, 1971.

[78] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

[79] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[80] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[81] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774*, 2020.

[82] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 412. ACM, 2018.

[83] Christos Dimitrakakis, Yang Liu, David C Parkes, and Goran Radanovic. Bayesian fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 509–516, 2019.

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

[85] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[86] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1106–1119, 2018.

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

[88] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

[89] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[90] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 2018.

[91] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18*, pages 242–250, New York, New York, USA, sep 2018. ACM Press.

[92] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 170–179, 2019.

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

[94] Danielle Ensign, Sorelle A Friedler, Scott Nevlle, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, 2018.

[95] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[96] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, New York, New York, USA, 2015. ACM, ACM Press. ISBN 9781450336642.

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

[98] Andrew Guthrie Ferguson. Big data and predictive reasonable suspicion. *University of Pennsylvania Law Review*, pages 327–410, 2015.

[99] Matthias Feurer and Others. Efficient and robust automated machine learning. In *Advances in neural information processing systems*, pages 2962–2970, 2015.

[100] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[101] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint arXiv:1801.01489*, 2018.

[102] Jack Fitzsimons, AbdulRahman Al Ali, Michael Osborne, and Stephen Roberts. A general framework for fair regression. *Entropy*, 21(8):741, 2019.

[103] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019.

[104] Kazuto Fukuchi, Jun Sakuma, and Toshihiro Kamishima. Prediction with model-based neutrality. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 499–514, 2013.

[105] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur)*, 42(4):1–53, 2010.

[106] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510, 2017.

[107] Bo Gao. *Exploratory Visualization Design Towards Online Social Network Privacy and Data Literacy*. PhD thesis, Ku Leuven, 2015.

[108] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.

[109] Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360. ACM, 2006.

[110] Bruce Glymour and Jonathan Herington. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 269–278, 2019.

[111] Naman Goel, Mohammad Yaghini, and Boi Faltings. Non-discriminatory machine learning through convex fairness criteria. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[112] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[113] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073. IEEE, 2012.

[114] Sandra González-Bailón and Others. Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27, 2014.

[115] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[116] Bryce W Goodman. Economic models of (algorithmic) discrimination. In *29th Conference on Neural Information Processing Systems*, volume 6, 2016.

[117] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

[118] Ben Green. "Fair" Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018.

[119] Ben Green. The false promise of risk assessments: Epistemic reform and the limits of fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'20). ACM. https://doi.org/10.1145/3351095.3372869*, 2020.

[120] Ben Green and Salomé Viljoen. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

[121] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912, 2018.

[122] Robert M Guion. Employment tests and discriminatory hiring. *Industrial Relations: A Journal of Economy and Society*, 5(2):20–37, 1966.

[123] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

[124] Christian Haas. The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness. In *International Conference on Information Systems (ICIS) 2019*, 2019.

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

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

[127] Margeret Hall and Simon Caton. Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. *PloS one*, 12(9):e0184417, 2017.

[128] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[129] Moritz Hardt, Eric Price, Nati Srebro, and Others. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[130] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

[131] Hoda Heidari and Andreas Krause. Preventing disparate treatment in sequential decision making. In *IJCAI*, pages 2248–2254, 2018.

[132] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems 31*, pages 1265–1276. Curran Associates, Inc., 2018.

[133] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery*, 28(5-6):1503–1529, 2014.

[134] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

[135] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 259–268. ACM, 2019. ISBN 9781450361255.

[136] Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890, 2019.

[137] Ben Hutchinson and Margaret Mitchell. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 49–58. ACM, 2019.

[138] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. Fae: A fairness-aware ensemble framework. *arXiv preprint arXiv:2002.00695*, 2020.

[139] Yasser Jafer, Stan Matwin, and Marina Sokolova. Privacy-aware filter-based feature selection. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 1–5. IEEE, 2014.

[140] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, 25(5):427–491, dec 2015.

[141] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. *arxiv*, 2019. URL `https://arxiv.org/pdf/1901.04966.pdf`.

[142] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

[143] James E Johndrow, Kristian Lum, et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

[144] Kory D Johnson, Dean P Foster, and Robert A Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*, 2016.

[145] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair algorithms for infinite and contextual bandits. *arXiv preprint arXiv:1610.09559*, 2016.

[146] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[147] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

[148] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh M Pai, Aaron Roth, and Rakesh Vohra. Fair prediction with endogenous behavior. Technical report, arXiv. org, 2020.

[149] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. Simple rules for complex decisions. *Available at SSRN 2919024*, 2017.

[150] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2018.

[151] Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868*, 2018.

[152] Peter Kairouz, Jiachun Liao, Chong Huang, and Lalitha Sankar. Censored and fair universal representations using generative adversarial models. *arXiv*, pages arXiv–1910, 2019.

[153] Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pages 8895–8906, 2018.

[154] Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887*, 2018.

[155] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, oct 2012. ISSN 0219-1377.

[156] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

[157] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision Theory for Discrimination-Aware Classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, dec 2012. ISBN 978-1-4673-4649-8.

[158] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, Springer, Berlin, Heidelberg, 2012.

[159] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation Independence. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 187–201. PMLR, 2018.

[160] Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999.

[161] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[162] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning, 2017.

[163] Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. *arXiv preprint arXiv:1902.02979*, 2019.

[164] Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31*, pages 4842–4852. 2018.

[165] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[166] Svetlana Kiritchenko and Saif Mohammad. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, 2018.

[167] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[168] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*, 2017.

[169] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[170] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[171] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2737–2746. PMLR, 10–15 Jul 2018.

[172] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, pages 853–862, 2018.

[173] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual Fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4069–4079, 2017. ISBN 978-1-5108-6096-4.

[174] Matt J Kusner, Chris Russell, Joshua R Loftus, and Ricardo Silva. Causal interventions for fairness. *arXiv:1806.02380*, 2018.

[175] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 275–282, 2002.

[176] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345, 2019.

[177] Himabindu Lakkaraju and Cynthia Rudin. Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, pages 166–175, 2017.

[178] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[179] Eric L. Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. Fairness-Aware Loan Recommendation for Microfinance Services. In *Proceedings of the 2014 International Conference on Social Computing - SocialCom '14*, pages 1–4, New York, New York, USA, 2014.

[180] Nicol Turner Lee. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 2018.

[181] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. The tyranny of data? the bright and dark sides of data-driven decision-making for social good. In *Transparent data mining for big and small data*, pages 3–24. Springer, 2017.

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

[183] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[184] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Annual International Cryptology Conference*, pages 36–54. Springer, 2000.

[185] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31*, pages 8125–8135. 2018.

[186] Zachary C Lipton. The Mythos of Model Interpretability. *Queue*, 16(3):30, 2018.

[187] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[188] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060, 2019.

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

[190] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks, 2016.

[191] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv:1610.08077*, 2016.

[192] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 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*, pages 502–510. ACM, 2011.

[193] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1): 3–es, 2007.

[194] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning*, pages 3381–3390, 2018.

[195] P Manisha and Sujit Gujar. A neural network framework for fair classifier. *arXiv:1811.00247*, 2018.

[196] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. Algorithmic pollution: Understanding and responding to negative consequences of algorithmic decision-making. In *Working Conference on Information Systems and Organizations*, pages 31–47. Springer, 2018.

[197] Annette Markham and Others. Ethical decision-making and Internet research: Version 2.0. *Association of Internet Researchers*, 2012.

[198] Douglas S Massey and Nancy A Denton. *American apartheid: Segregation and the making of the underclass*. Harvard University Press, 1993.

[199] Bruce McKeown and Dan B Thomas. Q methodology. *Quantitative applications in the social sciences*, 66, 2013.

[200] Douglas S McNair. Preventing disparities: Bayesian and frequentist methods for assessing fairness in machine-learning decision-support models. *New Insights into Bayesian Inference*, page 71, 2018.

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

[202] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in classification. *arXiv:1705.09055*, 2017.

[203] Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, 3(1):1–14, 2016.

[204] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 230–239, 2019.

[205] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[206] Rajeev Motwani and Ying Xu. Efficient algorithms for masking and finding quasi-identifiers. In *Proceedings of the Conference on Very Large Data Bases (VLDB)*, pages 83–93, 2007.

[207] C Munoz, M Smith, and D Patil. Big data: a report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President*, 2016.

[208] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[209] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.

[210] Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654, 2018.

[211] Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

[212] Alejandro Noriega-Campero, Michiel A Bakker, Bernardo Garcia-Bulle, and Alex'Sandy' Pentland. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 77–83, 2019.

[213] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[214] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–237, 2019.

[215] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018.

[216] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 126–131, 2012.

[217] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

[218] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 339–355. Springer, 2017.

[219] Alexander Peysakhovich and Christian Kroer. Fair division without disparate impact. *arXiv:1906.02775*, 2019.

[220] Emma Pierson. Demographics and discussion influence views on algorithmic fairness. *arXiv:1712.09124*, 2017.

[221] Emma Pierson. Gender differences in beliefs about algorithmic fairness. *arXiv:1712.09124*, 2017.

[222] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. On measuring bias in online information. *ACM SIGMOD Record*, 46(4):16–21, 2018.

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

[224] J Podesta and Others. Big data: Seizing opportunities, preserving values. *Executive Office of the President*, 2014.

[225] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[226] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481, 2020.

[227] Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.

[228] Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

[229] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Fighting Fire with Fire. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*, pages 231–239, 2019.

[230] Alexander S Rich and Todd M Gureckis. Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, 1(4):174–180, 2019.

[231] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638, 2014.

[232] Clemens Rösner and Melanie Schmidt. Privacy preserving clustering with constraints. *arXiv:1802.02497*, 2018.

[233] Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. Representation and communication: challenges in interpreting large social media datasets. In *ACM Computer Supported Cooperative Work (CSCW)*, pages 357–362, 2013.

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

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

[236] Babak Salimi, Bill Howe, and Dan Suciu. Data management for causal algorithmic fairness. *Data Engineering*, page 24, 2019.

[237] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

[238] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[239] Andrea Saltelli and Others. Sensitivity analysis in practice: a guide to assessing scientific models. *Chichester, England*, 2004.

[240] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, 19(4):1–26, 2019.

[241] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

[242] Richard L Sawyer, Nancy S Cole, and James W L Cole. Utilities and the Issue of Fairness in a Decision Theoretic Model for Selection. *Journal of Educational Measurement*, 13(1):59–76, 1976. ISSN 00220655, 17453984.

[243] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[244] Hansen Andrew Schwartz, Johannes C Eichstaedt, Lukasz Dziurzynski, Margaret L Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. Toward personality insights from language exploration in social media. In *2013 AAAI Spring Symposium Series*, 2013.

[245] Andrew D Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[246] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

[247] Ashudeep Singh and Thorsten Joachims. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2219–2228, 2018.

[248] Ashudeep Singh and Thorsten Joachims. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems 32*, pages 5427–5437. 2019.

[249] Michael Skirpan and Micha Gorelick. The Authority of "Fair" in Machine Learning. *arXiv:1706.09976*, 2017.

[250] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[251] Ana Sokolovska and Ljupco Kocarev. Integrating technical and legal concepts of privacy. *IEEE Access*, 6: 26543–26557, 2018.

[252] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A Unified Approach to Quantifying Algorithmic Unfairness. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '18*, pages 2239–2248, 2018. ISBN 9781450355520.

[253] Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2459–2468, 2019.

[254] Harald Steck. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 154–162, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016.

[255] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

[256] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[257] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[258] Rachael Tatman. Gender and dialect bias in Youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, 2017.

[259] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

[260] Robert L Thorndike. Concepts of Culture-Fairness. *Journal of Educational Measurement*, 8(2):63–70, 1971. ISSN 00220655, 17453984.

[261] Chris Thornton and Others. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *ACM SIGKDD Int.conf. on Knowledge Discovery and Data mining*, pages 847–855, 2013.

[262] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in catalonia. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 83–92, 2019.

[263] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[264] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–416. IEEE, 2017.

[265] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382, 2019.

[266] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems 31*, pages 1769–1778. 2018.

[267] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. Hiring algorithms: An ethnography of fairness in practice. 2019.

[268] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 2017.

[269] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[270] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7, 2018. ISBN 9781450357463.

[271] Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

[272] Hao Wang, Berk Ustun, Flavio P Calmon, and SEAS Harvard. Avoiding disparate impact with counterfactual distributions. In *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*, 2018.

[273] Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627, 2019.

[274] Lauren Weber and Elizabeth Dwoskin. Are workplace personality tests fair. *Wall Street Journal*, 29, 2014.

[275] Pak-Hang Wong. Democratizing algorithmic fairness. *Philosophy & Technology*, pages 1–20, 2019.

[276] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

[277] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.

[278] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[279] Yan Yan, Wanjun Wang, Xiaohong Hao, and Lianxiu Zhang. Finding quasi-identifiers for k-anonymity model by the set of cut-vertex. *Engineering Letters*, 26(1), 2018.

[280] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6, 2017. ISBN 9781450352826.

[281] Sirui Yao and Bert Huang. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems 30*, pages 2921–2930, 2017.

[282] Tal Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373, 2010.

[283] Karen Yeung. Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4):505–523, 2018.

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

[285] Tal Zarsky. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132, 2016.

[286] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 1569–1578, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185.

[287] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. Matching code and law: Achieving algorithmic fairness with optimal transport. *arXiv:1712.07924*, 2019.

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

[289] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

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

[291] Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv:1611.08292*, 2016.

[292] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. Fairness in reciprocal recommendations: A speed-dating study. In *Adjunct publication of the 26th conference on user modeling, adaptation and personalization*, pages 29–34, 2018.

[293] Alina Zhiltsova, Simon Caton, and Catherine Mulwa. Mitigation of unintended biases against non-native english texts in sentiment analysis. In *27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*, 2019.

[294] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1153–1162, 2018.

[295] Michael Zimmer. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology*, 12(4):313–325, 2010.

[296] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. may 2015.

[297] Andrej Zwitter. Big data ethics. *Big Data & Society*, 1(2):1–6, 2014.