Fairness in Machine Learning: A Survey

SIMON CATON, University College Dublin, Ireland
CHRISTIAN HAAS, University of Nebraska at Omaha, USA and Vienna University of Economics and Business (WU), Austria

When Machine Learning technologies are used in contexts that affect citizens, companies as well as researchers need to be confident that there will not be any 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 that aim to increase the fairness of Machine Learning. It organizes 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, and unsupervised learning is also provided along with a selection of currently available open source libraries. The article concludes by summarizing open challenges articulated as five dilemmas for fairness research.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Social and professional topics** → *User characteristics*; • **General and reference** → *Surveys and overviews;*

Additional Key Words and Phrases: Fairness, accountability, transparency, machine learning

ACM Reference format:
Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7, Article 166 (April 2024), 38 pages.
https://doi.org/10.1145/3616865

1 INTRODUCTION

Machine Learning (ML) technologies solve challenging problems that often have high social impact, such as examining re-offence rates (e.g., References [11, 25, 27, 44, 100, 224]), automating chat and (tech) support, and screening job applications (see References [241, 285]). 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, the idea of fair ML gained significant attention in the 2010s. However,

[1]https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
[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-assistants/

Authors' addresses: S. Caton, UCD School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland; email: simon.caton@ucd.ie; C. Haas, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria; email: christianhaas@unomaha.edu.
Author's current address: University of Nebraska at Omaha, 6001 Dodge Street, PKI 173A, Omaha, NE 68182, USA.

0360-0300/2024/04-ART166 $15.00
https://doi.org/10.1145/3616865

from a historical perspective modern approaches often build on prior definitions, concepts, and considerations 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 [147]. For example, discrimination in hiring decisions has been examined since the 1960s [133]. 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 contexts [72, 73] and discrimination in employment contexts [133]. The Civil Rights act spurred 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.

It is noteworthy that much of the early literature into fairness considered regression and correlation-based scenarios (standardized test scores; see, e.g., Reference [81]), compared to the recent emphasis on classification (binary decisions). However, the general notions transfer to (binary) classification settings and thus define essential concepts such as protected/demographic variables (e.g., References [72, 73]), notions of group vs. individual fairness (e.g., References [257, 278]), impossibility conditions between fairness conditions [81], and metric-based fairness quantification (e.g., true positive rates [74]).

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 Reference [147] notes, 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.

The importance of fairness and related ethical principles, however, is recognized as key to improving the trustworthiness of ML (and AI in general) [75, 198]. However, fairness is a complex topic that has to try and balance many different contexts, multi-faceted sociocultural concepts (beyond race, gender, age, etc.), aspects of equality and diversity, and ethical principles such as whether it is even appropriate to use ML at all. 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 [33, 42], inappropriate uses of data leading to (un)conscious bias(es) [41, 42], data veracity and quality [320], data relativity and context shifts [42, 122, 249], and subjectivity filters [41]. All of this results in an array of challenges that can be overwhelming, and current ML libraries often do not (yet) accommodate means to ascertain social accountability. Note that data are not the only source of bias and discrimination; here, we refer to Reference [213] for a more general discussion.

We recognize that becoming familiar with approaches to fairness in ML can feel insurmountable. Thus, this survey aims to provide a concise overview of key topics in the fairness literature. We seek to achieve this by: (1) Providing an entry-level introduction to the area fairness in ML (Section 2); (2) Summarizing the current approaches to measure fairness in ML within a standardized notation framework discussing the various tradeoffs of each approach as well as their overarching objectives (Section 3); (3) Defining 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); we have structured this discussion specifically to highlight high-level fairness objectives, a brief synopsis, and a commentary on key benefits and challenges in the use of approaches; and (4) Outlining key "dilemmas" in fairness research to act as a means to help identify meaningful future research endeavors to improve the accessibility of the domain and maximize future

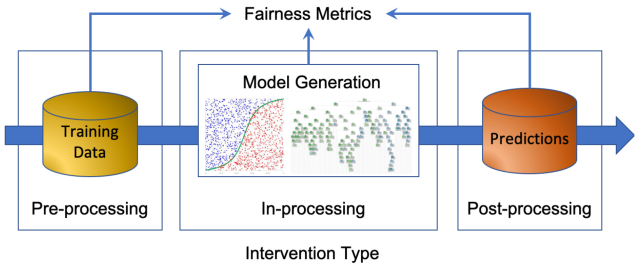

Fig. 1. High-level illustration of fairness intervention types in ML.

research impact (Section 6). We also highlight currently available toolkits for fair ML in an online appendix.

This article builds on other similarly themed ones that have focused on the history of fairness in ML [147], discrimination analysis [247], key choices and assumptions [216], a cross-domain consideration of bias detection, fairness management, and explainability management [226], different types of bias [213, 275], and a number of methods to mitigate bias [213]. In this article, we assume a working knowledge of applied ML, i.e., familiarity with the basic structure of data mining methodologies such as Reference [97] and how to apply and evaluate "standard" ML methods. We focus on fairness in classification tasks and closely related domains such as regression models and rating predictions. While bias and fairness in Natural Language Processing is an important topic, we refer to overview articles in this domain [20, 261, 274] and likewise for reinforcement learning [112].

2 FAIRNESS IN MACHINE LEARNING: KEY METHODOLOGICAL COMPONENTS

The fairness literature focuses on either the technical aspects of bias and fairness in ML or theorizes on the social, legal, and ethical aspects of ML discrimination [124]. Technical approaches are typically applied prior to modeling (pre-processing), at the point of modeling (in-processing), or after modeling (post-processing), i.e., they emphasize intervention [33].

In this article, 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 1. While not all approaches will fit into this framework, it is easy to understand and acts as one dimension in an approach taxonomy. For visual simplicity, note that we omit standard practices to prepare and sample data for ML methods and their evaluation that are common within methodologies like **KDD (Knowledge Discovery in Databases)** [97]. While this is common in fairness research, we discuss the need to include these methodologies in fairness pipelines in Section 6.5.

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 (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 that are socioculturally precarious for the application of ML. Examples are gender, ethnicity, age, their synonyms, and essentially any other feature of the data that involves or concerns people [16].

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 References [25, 129, 130, 191, 192, 209, 268, 280, 306] 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. References [2, 106] both present approaches specifically looking at the 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. Reference [66] 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[4]) [54, 88, 89, 89, 196, 202, 231, 247, 286, 308].

Understanding "related" variables in a general sense is well studied, especially in the privacy and data archiving literature, where overlooked variable relationships can enable the deanonymization of published data (see Reference [318]). The fairness literature, however, often overlooks these effects on fairness, although the relationship between discrimination and privacy was noted in Reference [89]. In particular, sensitive data disclosure is a long-standing challenge in protecting anonymity when data is published and/or analyzed [6, 111, 195]. Key approaches (e.g., References [194, 205, 276]) seek to protect specific individuals and groups from being identifiable, i.e., minimize disclosure risk. Yet, these approaches can still struggle to handle multiple sensitive attributes at once [194]. While there has been success in anonymizing datasets, approaches still often require a list of features to protect. For explicit identifiers (name, gender, etc.), lists exist. Yet, 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 [111] or seeks to discover them on a case-by-case basis (e.g., References [150, 217]) towards the idea of privacy preserving data mining [6]. Reference [134] also discusses the notion of proxy groups: "similar" data instances 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., Reference [301]) or specific notions of fairness (we introduce these approaches in Section 4.2). Reference [251] provides a brief overview of different theoretical applications to algorithmic fairness, with Reference [118] 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.

To illustrate why handling proxy variables is an important consideration for fairness, consider this example: A dataset has been stripped of all *a priori* known sensitive variables (an intervention we refer to as blinding in Section 4.1) yet the proxy university faculty is left intact. This proxy hints at the subject the individual studied (which may indicate their gender) and the level of their education (itself a sensitive variable). Thus, even though sensitive variables were removed, some amount of information through the proxy variable is retained, which may erroneously give the impression that the intervention has produced a fair ML model. Of course, this example is contrived for illustrative purposes, but it seeks to highlight that proxies are not straightforward for inexperienced practitioners. We would encourage review of Reference [38, p. 1014] for further examples.

2.2 Metrics

Underpinning intervention-based approaches are an ever-increasing array of fairness measures seeking to quantify fairness. The implication of "measurement" is, however, precarious, as it

[4]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 [155].

Table 1. Example Proxy Relationships Based on Findings from References [25, 38, 106, 137, 210, 251, 259, 260, 296, 304]

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

implies a straightforward process [16]. 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 [27, 69, 177]. 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 tradeoffs in a given scenario [135].

2.3 Technical Fairness Interventions: Pre-, In-, and Post-processing Approaches

All technical fairness interventions operate at specific locations in the ML pipeline, i.e., before, during, or after model training. Here, we give a high-level introduction to each of these approach types, discuss their potential benefits and limitations, and conclude with a brief contrasting summary.

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 [165]. The main idea here is to train a model on a "repaired" dataset. 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 [88].

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 that 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.

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 the entire ML pipeline is not exposed.

It can often be quite difficult to ascertain which type of approach will benefit a given scenario. 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

Table 2. Overview of Suggested Fairness Metrics for Binary Classification

	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 3. Notation for Binary Classification

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

model itself. Yet, modification of the data and/or model output may have legal implications [17] and can mean models are less interpretable [192, 202], 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 myriad definitions and metrics (e.g., References [17, 27, 69, 139, 178, 298, 307]) to mathematically represent bias, fairness, and/or discrimination; but they lack consistency in naming conventions [76] and notation. More so, there are many different interpretations of what it means to be "fair." Several publications provide a (limited) overview of multiple fairness metrics and definitions, e.g., References [108, 188, 238, 245, 275, 288]. We extend these by including additional perspectives for types of biases and a larger set of metrics and definitions.

Although the literature has defined myriad notions to quantify fairness, each measures and emphasizes different aspects of "fairness." Many are difficult/impossible to combine [69, 177], but ultimately, we must keep in mind (as noted in Reference [68]) there is no universal means to measure fairness and 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 the main categories of metrics with Table 3 introducing key notation.

3.1 Abstract Fairness Criteria

Most quantitative definitions and measures of fairness are centered around three fundamental aspects of a (binary) classifier[5]: 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

[5]However, we note that extensions to multi-class classification are an area of active research (see, e.g., Reference [35]).

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 [16]:

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, then 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 Reference [16] points out, Sufficiency is closely related to some of the calibration-based metrics. Reference [16] also discusses 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 Section 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 (Section 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 [77, 98, 168, 313]; 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: Like statistical parity, disparate impact looks at the probability of being positively classified. In contrast to parity, it considers the ratio between unprivileged and privileged groups. Its origins are in legal fairness considerations for selection procedures that sometimes use an 80% rule to define if a process has disparate impact (ratio smaller than 0.8) or not [98].

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

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 [235].

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 consider additional aspects such as **True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR)**, and **False Negative Rate (FNR)**. The advantage of this is that they can include underlying differences between groups who would otherwise not be included in the parity-based approaches. This is related to the Separation criterion (Section 3.1).

Equal Opportunity: As parity and disparate impact do not consider potential differences in the groups being compared, References [139, 239] consider additional metrics utilizing the FPR and TPR between groups. Such that equal opportunity promotes that the 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 toTPR, 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. Reference [27] adjusts this concept by looking at relative accuracy rates across different groups. If two groups have the same accuracy, then they are considered equal based on their accuracy.

$$\frac{TP_{g_i} + TN_{g_i}}{TP_{g_i} + TN_{g_i} + FP_{g_i} + FN_{g_i}} = \frac{TP_{g_j} + TN_{g_j}}{TP_{g_j} + TN_{g_j} + FP_{g_j} + FN_{g_j}} \tag{8}$$

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.

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

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 [158]: 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, Reference [30] provides an additional metric that specifies equal opportunity on a specific attribute a out 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. Related to the Sufficiency criterion, calibration-based metrics take the predicted probability, or score, into account, differentiating them from metrics above that use predicted and actual values.

Test fairness/calibration/matching conditional frequencies [69, 139]: 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 [178]: 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** [178]: 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 [86] extends the balance concept from Reference [178] 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 that 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. Reference [182] proposes the concept of counterfactual fairness that builds on causal fairness models and is related to both individual and group fairness concepts. Reference [270] proposes a generalized entropy index that 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 changing an individual's sensitive variable, while holding all other variables that are not causally dependent on the sensitive variable constant, does not change the prediction (distribution).

Generalized Entropy Index: Reference [270] 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 $\mu = \frac{\sum_i b_i}{n}$:

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

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

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

3.4 Commentary on Fairness Metrics

The literature is at odds with respect to prioritizing individual or group fairness. Reference [270] notes that many approaches to group fairness tackle only between-group issues, as a consequence worsening within-group fairness. Consequently, users must decide on where to place emphasis, but be mindful of the tradeoff between any fairness measure and model accuracy [26, 54, 77, 89, 139, 319]. With a reliance on expressing fairness and bias mathematically, References [76, 128] argue that these definitions often do not map to normative social, economic, or legal understandings of fairness. This is corroborated by Reference [266], which notes the over-emphasis of disparate treatment and References [3, 54, 270], which 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 variables (as introduced in Section 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 decisions (hiring vs. not hiring; offering a loan vs. not offering a loan, etc.); (2) quantifying fairness on a binary dependent variable is mathematically more convenient; addressing multi-class problems would add terms to the fairness quantity.

In this section, we discuss the main approaches for tackling fairness in the binary classification case: providing a concise fairness objective statement along with a brief synopsis and commentary for each approach. We arrange approaches of fairness intervention into a visual taxonomy according to the location in the ML framework (Figure 1), i.e., pre-processing: Figure 2, in-processing: Figure 3, and post-processing: Figure 4. We note an abundance of pre- and in-processing vs. post-processing methods and that many researchers increasingly develop intervention strategies that belong to multiple stages. However, we are yet to find a strategy using pre-, in-, and post-processing methods combined. We also note that we do not comment on the advantages of specific approaches over others (this is a very context specific consideration), instead, we outline challenges researchers must navigate. While many papers compare specific subsets of the approaches discussed in this section, the literature has an urgent need for a structured meta review of approaches to fairness. In the absence of such a review, offering advice on approach selection is unrealistic.

4.1 Blinding → Pre-processing

Fairness Objective: Make a classifier "immune" to one or more sensitive variables [177].

Synopsis: A classifier is, for example, race blind if there is no observable outcome differentiation based on race. Reference [139] sought to train a race blind classifier (among others) in that each of the four race groups have the same threshold value (see Section 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** as distinct from **omission of** sensitive variables. Whereas omission refers to not including the sensitive variables as input for the prediction models, immunity also considers the indirect effect that sensitive variables can have on other (input) variables of a prediction model. For instance, sensitive variables often are correlated with other variables in the data, and approaches focusing on immunity aim to prevent these indirect effects from resulting in discrimination measured through the sensitive variable.

Commentary: Omission has been shown to decrease model accuracy [61, 136, 253] and increase discrimination [54, 89, 165]. Both omission and immunity overlook relationships with proxy variables (as discussed in Section 2.1, we note Reference [291] as an exception here, which omits

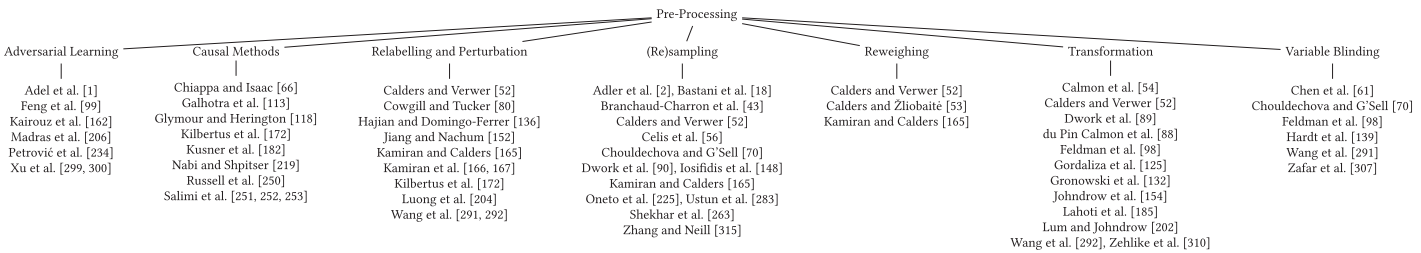

Fig. 2. Pre-processing methods.

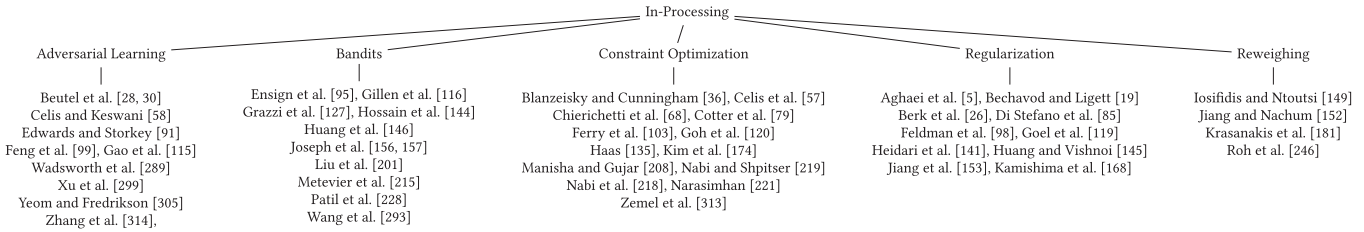

Fig. 3. In-processing methods.

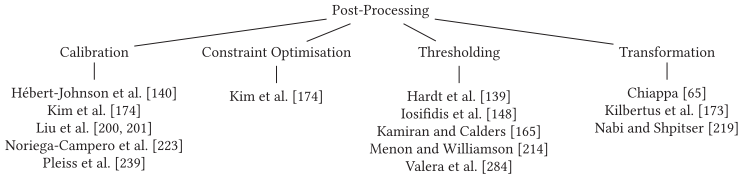

Fig. 4. Post-processing methods.

proxies), which can result in increasing bias and discrimination [177] or indirectly concealing discrimination [88]. It also ignores that discrimination may not be one variable in isolation, but rather the result of several joint characteristics [231]. Yet, determining which combination(s) of variables to blind is non-trivial, and the risk of omitted variable bias should not be downplayed [71, 160].

Approaches in subgroup analysis (see Section 4.3) have used statistical techniques to determine when variable immunity does not adversely affect fairness (e.g., Reference [70]). Similarly, researchers still use omission in their evaluation methodologies to compare to earlier works and act as a form of fairness baseline. Omission can also be used in specific parts of the fairness methodology, for example, Reference [98] temporarily omits sensitive variables prior to transforming (see Section 4.4) the training data. Blinding (or partial blinding) has also been used as a fairness audit mechanism [2, 82, 143]. Specifically, such approaches explore how partially blinding features (sensitive or otherwise) affect model performance. This is similar to the idea of causal models (Section 4.2) and can help identify problematic sensitive or proxy variables with black-box-like analysis of an ML model.

4.2 Causal Methods → Pre-processing

Fairness Objective: Identify potentially useful relationships between sensitive and non-sensitive variables to provide insights for fairness-related methodological decisions.

Synopsis: Approaches using causal methods recognize that the training data 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 [66, 113, 118, 172, 182, 219, 250]. Thus, causal methods are specifically well suited to identifying (1) proxies of sensitive variables as discussed in Section 2.1; (2) which subgroups are most (un)fairly treated and differentiate the types of bias exhibited [118]; and (3) provide transparency with respect to how (classification) decisions were made [143]. Researchers have also leveraged causal dependencies to "repair" training data [251–253] using dependency information to insert, modify, and remove training samples to satisfy fairness-specific constraints and conditional independence properties of the training data. Initial results with data repair methods have shown to result in "debiased" classifiers that are robust to unseen test data, yet this process requires significant computational resources.

Commentary: Causal methods can provide visual descriptions of (un)fairness in the dataset (see References [143, 173, 182, 183]). **Directed acyclic graphs (DAGs)** are a common means to represent conditional independence assumptions between variables [155]. This means they can have high utility in studies of fairness in industry settings. The main challenge for causal models is the requirement for background information and context regarding the causal model that may not be accessible [253]. They have also been criticized for not well examining how they would be applied in practice [253].

4.3 Sampling and Subgroup Analysis → Pre-processing

Fairness Objective: Sampling methods have two primary objectives: (1) to create samples for the training of robust algorithms (e.g., References [9, 43, 56, 90, 148, 263, 283, 307]), i.e., "correct" training data and eliminate biases [148]; 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., References [2, 70, 315]).

Synopsis: Within the sampling approaches, the application of decoupled classifiers and multitask learning has emerged (see References [9, 52, 90, 225, 283, 307]). Here, the training data are decoupled, i.e., split, into subgroups (or combinations) of one or more sensitive variables (e.g., [old, males]). These groupings can also be learned in a pre-processing step: multitask learning. Reference [134] also learns proxy groupings. Thus, such approaches seek to make the most

accurate models for each subgroup (decoupled classifiers) or consider the observation of different subgroups (multitask learning). Reference [148] extends this approach, creating an ensemble of ensembles to operate on the protected groups.

Subgroup analysis can also be a useful exercise in model evaluation [2, 70, 315] 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 Section 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 [2, 70]. See Reference [212] for an example set of statistical tests to indicate the likelihood of fair decisions. Probabilistic verification (e.g., Reference [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., Reference [10]), these approaches present fairness as a dichotomous outcome: a model is fair, or it is not, 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.

Commentary: Reference [56] articulates the challenge of sampling for fairness as the following question: How are samples selected from a (large) dataset that is both diverse in features and fair to sensitive attributes? Without care, sampling can propagate biases within the training data [90, 224], as ensuring diversity in the data used to train the model makes no guarantees of producing fair(er) models [224]. As such, approaches that seek to create fair training samples include notions of fairness in the sampling strategy. Reference [165] proposes 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. Reference [290] proposes an iterative human-in-the-loop resampling approach where users can evaluate potential fairness issues of a model through interactive visualizations. Subsequent bias mitigation options can then iteratively reduce the observed bias.

A key challenge for sampling and subgroup analysis is to ensure that sufficient data are available for each subgroup. Otherwise, this method of sampling can negatively affect performance and fairness, as shown by Reference [9]. Outliers can also be problematic [46, 222]. Similarly, and especially for decoupled classifiers, is the challenge of selecting groups: Some can be rarer than others [283] and as such, a balance is needed to ensure groups are as atomic as possible but robust against gerrymandering [140, 170]. 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., References [170, 298]) and clustering (e.g., References [70, 148]), as (good) clusters well approximate underlying data distributions and subgroups. Reference [148] 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.

4.4 Transformation → Pre-processing & Post-processing

Fairness Objective: Learn or generate new "fair" representations of the data (e.g., a mapping or projection function) that still preserves the fidelity of the ML task [98].

Synopsis: Current transformation approaches operate mainly on numeric data, which is a significant limitation [98], yet approaches for other scenarios such as image classifications have been proposed [132]. There are different perspectives to transforming the training data: operating on

the dependent variable (e.g., Reference [88]), operating on numeric non-sensitive variables (e.g., References [52, 98, 202]), mapping individuals to an input space that is independent of specific protected subgroupings (e.g., References [54, 89, 125, 154, 185, 310, 313]), transforming the distribution of model predictions in accordance to specific fairness objectives (e.g., References [153, 292]), and combining representation learning and data augmentation using generative models [229]. 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 Section 4.5).

To illustrate transformation, Reference [98] discusses transforming the distribution of SAT scores towards the median to "degender" the distribution into one that retains only the rank order for individuals, i.e., they remove information about protected variables from a set of covariates. Transformation approaches often seek to retain rank orders within transformed variables to preserve predictive ability. References [52, 202] 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., Reference [88]) that seek to prevent mapping "high" values to "low" values and vice versa.

Although largely a pre-processing method, transformation can also be applied within a post-processing phase. References [65, 173, 219] transform the output of a classifier in accordance to the identification of unfair causal pathways, either by averaging [219], constraining the conditional distribution of the decision variable [173], or through counterfactual correction [65]. As an approach, this is similar to the idea of calibration (see Section 4.10) and thresholding (see Section 4.11) .

Commentary: There are a number of challenges 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 [88, 98, 202] through data loss [125]. To balance this tradeoff, approaches often partially repair by transforming the data towards some target distribution, but not in its entirety (e.g., References [98, 125]). (2) Understanding the relationship(s) between sensitive and potential proxy variables is hard [98], thus causal methods (Section 4.2) may be useful precursor to transformation techniques. (3) The selection of "fair" target distributions is not straightforward [88, 125, 319]. (4) Finding an "optimal" transformation under high dimensionality can be computationally expensive, even under assumptions of convexity [88]. (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 [192, 202], which may be at odds with data protection legislation. (7) There are no guarantees that the transformed data have "repaired" discriminatory latent relationships with proxy variables [54].

4.5 Relabelling and Perturbation → Pre-processing

Fairness Objective: Modify the training data such that underprivileged and privileged instances are treated similarly and/or explore the effects of such modifications on model fairness.

Synopsis: Relabelling and perturbation are a specific subset of transformation approaches: They either flip or modify the dependent variable (relabelling; e.g., References [52, 80, 165–168, 204]) or otherwise change the distribution of one or more variables in the training data directly (perturbation; e.g., References [136, 152, 291]). Referred to as data-massaging by References [165, 313], relabelling involves the modification of the labels of training data instances so 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. For example, Reference [204] relabels the dependent variable (flips 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 used in counterfactual studies (e.g., References [142, 161, 291]) that investigate if flipping the (in)dependent variable(s) affect the classification outcome.

Perturbation is similar to "repairing" some aspect(s) of the data to improve fairness. Applications in perturbation-based data repair [98, 125, 136, 152, 154, 172] have shown that accuracy is not significantly affected. Often, perturbation-based approaches are applied as a pre-processing step to prepare for an in-processing approach; often reweighing (e.g., Reference [152]; see Section 4.6) and/or regularization/optimization (e.g., References [283, 291]; see Section 4.7). It has been proposed as a mechanism to detect proxy variables, influential variables [292], and counterfactual distributions [291].

Commentary: While 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. As for transformation, modification of the data via relabelling and perturbation is not always legally permissible [17], and changes to the data should be minimized [136, 204]. Reference [181] also notes 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.

Closely related to perturbation approaches is the use of sensitivity analysis (see Reference [254]) to explore how various aspects of the feature vector affect a given outcome. This is a relatively under-addressed area in the fairness literature, perhaps due to the introduction of SHAP values[6] [203], which has seen an increase in attention from fairness researchers (e.g., Reference [126]). Yet, sensitivity analysis (and explainable fairness in general) has been well motivated (although perhaps indirectly): Reference [124] called for a better understanding of bias stemming from uncertainty, and Reference [139] stressed that assessment of data reliability is needed. The application of sensitivity analysis in ML is often to measure the stability of a model [242]. While a number of approaches exist to determine model stability [39, 45, 171, 184, 243, 262], 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 principles of sensitivity analysis. Reference [80] proposed the perturbation of feature vectors to measure the effect on model performance of specific interventions. References [114, 211] investigated visual mechanisms to better display "issues" with data to users, yet these approaches generally lack support for novice users [24]. References [159, 161] used sensitivity analysis to evaluate sensitive variables and their relationship(s) with classification outcomes, indicating that sensitivity analysis (while not a method to improve fairness and thus omitted from Figures 2–4) can help to better understand uncertainty with respect to fairness.

4.6 Reweighing → Pre-processing & In-processing

Fairness Objective: Change the "impact" of instances (observations) on the prediction model during training to promote "fair(er)" handling of sensitive variables and/or underprivileged groups.

Synopsis: Unlike transformation, relabelling, and perturbation approaches that alter (certain instances of) the data, reweighing assigns weights to instances of the training data while leaving

[6]SHAP stands for SHapley Additive exPlanations, which are based on the game-theoretic concept of Shapley values. The goal of this approach is to measure the importance of features in a prediction model for specific predictions [203].

the data themselves unchanged. Weights can be introduced for multiple purposes: (1) to indicate a frequency count for an instance type (e.g., Reference [52]), (2) to place lower/higher importance on "sensitive" training samples (e.g., References [53, 152, 165]), or (3) to improve classifier stability (e.g., Reference [181]). Reweighing as an approach straddles the boundary between pre-processing and in-processing. For example, Reference [165] seeks to assign weights that take into consideration the likelihood of an instance with a specific class and sensitive value pairing (pre-processing). Whereas, Reference [181] first builds an unweighted classifier, learns the weights of samples, then retrains their classifier using these weights (in-processing). A similar approach is taken by Reference [152], which identifies sensitive training instances (pre-processing) but then learns weights for these instances (in-processing) to optimize for the chosen fairness metric. In gradient descent-based models, several in-processing approaches have been suggested to learn instance weights such that fairness can be improved [149, 207, 227, 246].

Commentary: With appropriate sampling (see Section 4.3), reweighing can maintain high(er) accuracy when compared to relabelling and blinding (omission) approaches [165]. However, as References [117, 181] note, classifier stability and robustness can be an issue. Thus, ML researchers need to carefully consider how reweighing approaches are applied and evaluate model stability. Reweighing also subtly changes the data composition, making the process less transparent [192, 202].

4.7 Regularization and Constraint Optimisation → In-processing

Fairness Objective: Extend the classifier's loss function such that it penalizes "unfair" outcomes.

Synopsis: Classically, regularization penalizes the complexity of the ML model to inhibit overfitting. Applied to fairness, regularization means adding penalty terms to penalize the classifier for discriminatory practices [168]. Thus, it is not hypothesis- (or model-) driven, but data-driven [19] and based upon the notion(s) of fairness considered. When extending the classifier's (convex) loss function with fairness terms, researchers typically seek to balance fairness and accuracy (e.g., References [5, 26, 36, 52, 57, 98, 119, 141, 153, 168, 208, 314]). Some notable exceptions emphasize: (1) empirical risk subject to fairness constraints or welfare conditions (e.g., References [87, 142]), (2) TPR/FPR of protected groups (e.g., Reference [19]), (3) stability of fairness (e.g., Reference [145]), or (4) counterfactual terms (e.g., Reference [85]).

In-processing (constraint) optimization approaches (e.g., References [3, 57, 68, 79, 120, 135, 174, 208, 218, 219, 221, 307, 313]) 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. Reference [218] 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., References [3, 120, 221]). Similarly, a multi-fairness metric approach has been proposed by Reference [174], where adaptions to stochastic gradient descent optimize weighted fairness constraints as an in-processing or post-processing (when a pre-trained classifier is used) scenario. References [79, 120, 221] summarize a number of additional constraint types as precision or budget constraints to address the accuracy-fairness tradeoff (often expressed as utility or risk functions, e.g., References [77, 135]); 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 References [109, 145].

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

Commentary: References [109, 145] 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). Reference [145] argues that stability of fairness can be addressed through regularization and presents corresponding empirical evidence through extensions of References [168, 307]. Aside from this, References [117, 313] note that regularization as a mechanism is fairly generic and can lead to a lack of model robustness and generalizability. In light of this, Reference [103] suggests a robustness framework for fairness that ensures fairness over different training samples. With the goal of better fairness generalization, the approach can also be used to quantify the robustness of fairness. This idea of "robustness of fairness" is often not considered by many works and would warrant better inclusion in future fairness studies.

Key challenges for regularization approaches are: (1) they are often non-convex in nature or achieve convexity at the cost of probabilistic interpretation [119]; (2) not all fairness measures are equally affected by the strength of regularization parameters [26, 85]; and (3) different regularization terms and penalties have diverse results on different datasets, i.e., this choice can have qualitative effects on the tradeoff between accuracy and fairness [26]. For constraint optimization, it can be difficult to balance conflicting constraints, leading to more difficult or unstable training [79].

4.8 Adversarial Learning → Pre-processing & In-processing

Fairness Objective: Tutor a classification model to be "fairer" by providing in-training feedback or modifying the training data to promote immunity to one or more sensitive variables.

Synopsis: In adversarial learning the objective is for an adversary to try and determine whether a model training algorithm is robust enough. The framework of Reference [123] helped popularize the approach through the process of detecting falsified data samples [58]. 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 [58]. 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., References [28, 30, 58, 91, 289, 299, 305, 314]). This is often formulated as a multi-constraint optimization problem considering many of the constraints, as discussed in Section 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., References [1, 99, 162, 206, 234, 299, 300]) with similar objectives to transformation, as discussed in Section 4.4, yet often moving towards a notion of "censoring" the training data with similar objectives to variable blinding, as discussed in Section 4.1. Work has also started applying the notions of causal and counterfactual fairness to adversarial learning (e.g., Reference [299]). 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. Additionally, in a Deep Neural Network setting, Reference [115] shows that different tradeoffs between accuracy and fairness can be achieved by selective dropout of neurons.

Commentary: An advantage of adversarial approaches is that they can consider multiple fairness constraints [289], often treating the model as a black box [206]. However, adversarial approaches have been reported to often lack stability, which can make them hard to train reliably [30, 99] and also specifically in some transfer learning scenarios [206] 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., Reference [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. Reference [256] illustrates this by

using a GAN with fairness constraints to produce "unbiased" image datasets, and Reference [299] have evidenced similar results for structured data.

4.9 Bandits → In-processing

Fairness Objective: Instrument fair online decision making with little or no training data.

Synopsis: As a reinforcement learning framework, bandits are motivated on the need for decisions to be made in an online manner [156] (although not always, e.g., Reference [215]), 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 [116]. Approaches that use bandits (e.g., References [95, 116, 156, 157, 201, 228, 293]) often do so on the basis of Reference [89]'s individual fairness, i.e., that similar individuals should be treated similarly. This allows researchers to frame the fairness problem as a stochastic multi-armed bandit, assigning either individuals or groups of "similar" individuals to arms, and represent fairness as regret [156, 201]. A key difference between the applications of bandits for fairness and bandits in general is that the learning algorithm is not trying to identify the "best" action to take, but rather generate a fair distribution over the actions that can be taken [144]. The main notions of fairness applied in the application of bandits are meritocratic fairness [127, 156, 157] (group agnostic), subjective fairness [201] (emphasizes fairness in each time period t), and counterfactual fairness (e.g., Reference [146]).

Commentary: The application of bandits to ML fairness has proliferated recently. It is important to highlight that the difference between the application of bandits to fair classification and fair recommendation (Section 5.2) is often hard to distinguish. However, the application of bandits has facilitated the exploration of multiple fairness considerations. Reference [215] allows for user-specific fairness preferences to be expressed based on multiple fairness definitions. We will come back to more recommendation-focused work in Section 5.2 and also refer the interested reader to Reference [112]'s survey of fairness in reinforcement learning. We do note, however, that this area of fairness research operates quite often in simulation and not using "standard" datasets (Reference [215] is a notable exception) and thus there is a need to produce or identify meaningful datasets for these approaches to be evaluated with and to compare to other approaches discussed throughout this article.

4.10 Calibration → Post-processing

Fairness Objective: To adjust the probability outputs of a model such that the portion of predicted positive outcomes matches that of positive examples across (or within) all (sub)groups in the dataset.

Synopsis: Calibration is the process of ensuring that the proportion of positive predictions is equal to the proportion of positive examples [83]. In the context of fairness, this should also hold for all subgroups (protected or otherwise) in the data [61, 239, 315]. Calibration is particularly useful when the output is not a direct decision but used to inform human judgment when assessing risks (e.g., awarding a loan) [223]. As such, a calibrated model does not inhibit biases of decision makers but ensures that risk estimates for various (protected) subgroups carry the same meaning [239].

Commentary: Calibrating an ML model for multiple protected groups and/or using multiple fairness criteria at once has been shown to be impossible [69, 140, 174, 178, 200, 201, 223, 239]. Reference [239] even notes that the goals of low error and calibration are competing objectives for a model. This occurs as calibration has limited flexibility [152]. Reference [298] 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 References [76, 139, 178, 223, 239] 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. Reference [223] also notes that this approach is Pareto sub-optimal, and instead propose a cost-based approach to balance calibration and error parity. Reference [200] suggests that calibration is sufficient as a fairness criterion if the model is unconstrained. Reference [140] instead seeks 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 [200]. References [170, 175] undertake a similar approach under different settings. Reference [201] has proposed a bandit-based approach to calibration.

4.11 Thresholding → Post-processing

Fairness Objective: Consider fairness metrics in the setting of thresholds for predicted scores (or decision boundaries in general) produced by an ML model.

Synopsis: Thresholding is a post-processing approach that is motivated on the basis that discriminatory decisions are often made close to decision-making boundaries because of a decision maker's bias [167] and that humans apply threshold rules when making decisions [176]. 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 [167]. 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 [139]. The underlying idea here is to incentivize good performance (in terms of both fairness and accuracy) across all classes and groups. Thresholding can claim compelling notions of equity, however, only when the threshold is correctly chosen [76].

Commentary: The main challenge for thresholding approaches is to find a tolerance level for unfairness in the calculation of threshold values. Computing threshold value(s) can be undertaken by hand to enable specific user preferences in the fairness accuracy tradeoff or with other statistical methods. Threshold values are often computed with respect to accuracy, but cases of class imbalance would invalidate this [148]. Computing by hand can introduce new biases without training [286], as fairness is typically not monotonic, thus assigning a threshold value may be quite arbitrary [270]. Reference [214] estimates the thresholds for each protected group using logistic regression, then uses a fairness frontier to illustrate disalignment between threshold values. Reference [167] uses an ensemble to identify instances in an uncertainty region to assist in setting a threshold value. Reference [105] proposes a method to shift decision boundaries using a form of post-processing regularization. Reference [284] uses posterior sampling to maximize a fairness utility measure. Reference [148] learns a threshold value after training an ensemble of decoupled ensembles (see Section 4.3) such that the discrepancy between protected and non-protected groups falls below a user-specified threshold value.

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.

[8] We note that there is a fine line between thresholding and calibration approaches and that they often overlap. We distinguish them as their objectives are subtly different: looking to ensure distributional qualities in the proportion of positive predictions (calibration) vs. moving a decision boundary (or threshold value) to maximize some fairness metric.

5.1 Fair Regression

Fair regression aims to minimize a loss function $l(Y, \hat{Y})$, which measures the difference between actual and predicted values, while also taking fairness aspects into account. The general formulation is similar to (binary) classification, with the difference that Y and $\hat{Y}$ are continuous rather than binary or categorical, and thus adapts the principles defined in Section 3. For example, parity-based metrics aim to make the loss function equal for different groups [4]. With respect to defining fairness metrics or measurements, Reference [27] suggests several metrics that can be used for general regression models. Reference [4] defines regression variants of statistical parity and bounded-group-loss metrics, the latter providing a customizable maximum allowable loss to address the tradeoff between fairness and loss (performance). Reference [51] considers 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. Reference [232] suggests a fairness degree as a measure of regression fairness alongside a search-based fairness testing strategy.

The effect of bias on parameter estimates and coefficients in multiple linear regression is discussed by Reference [155], which also suggests a post-processing approach to make parameter estimates impartial with respect to a sensitive attribute. Reference [180] includes fairness perspectives in non-convex optimization for (linear) regression using the coefficient of determination 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. Reference [233] proposes 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. Reference [168] suggests a regularization approach that can be applied to general prediction algorithms. Reference [110] defines the concept of μ-neutrality that measures if probabilistic models are neutral with respect to specific variables and shows that this definition is equivalent to statistical parity. Reference [26] proposes a family of regularization approaches that work with a variety of group and individual fairness metrics. Through a regularization weight, the proposed method is able to calculate and evaluate the efficient frontier of achievable accuracy-fairness tradeoffs. Reference [107] considers 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 and Ranking

Considerations of fairness have been actively studied in the context of rankings and recommender systems. For rankings in general, References [32, 302, 309] define different types of fairness notions such as group-based fairness in top-k ranking [302, 309], an individual fairness measure in rankings following concepts similar to Reference [139] and the equality of opportunity in binary classification [32], and unfairness of rankings over time through a dynamic measure called amortized fairness [32]. For a general overview and taxonomy of fairness concepts, we refer to the surveys for fairness in information retrieval and recommender systems [92, 294] and fairness in ranking [311, 312].

For recommender systems in particular, Reference [190] argues 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 new metrics and algorithms [12, 29, 59, 272, 303, 317]. 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 [303], fairness goals for recommender systems as overcoming algorithmic bias and making neutral recommendations independent of group membership (e.g., based on gender or age) [317], recommendation calibration, i.e., the proportional representation of items in recommendations [272], 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 [59]. Further approaches such as tensor-based recommendations have been proposed that take statistical parity into account [317] 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 [235].

An aspect that distinguishes fairness considerations in recommender systems is that fairness can be seen as multi-sided concept that can be relevant for both users (who consume the recommendations) and items. References [48, 49] 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. Reference [94] presents an empirical analysis of P-fairness for several collaborative filtering algorithms. Similarly, Reference [316] aims to find an optimal tradeoff between the utilities of multiple stakeholders. Other works considering fairness from either the consumer or provider side include the analysis of different recommendation strategies for a variety of (fairness) metrics [151], subset-based evaluation metrics to measure the utility of recommendations for different groups (e.g., based on demographics) [93], and a general framework to optimize utility metrics under fairness of exposure constraints [264, 265]. Several authors have also proposed bias mitigation strategies. This includes a pre-processing approach to make recommendations independent of a specific attribute (recommendation independence) [169], adding specifically designed "antidote" data to the input instead of input data manipulation to improve the social desirability of recommendations [244], a re-ranking algorithm that considers both consumer and provider sides while improving overall recommendation quality [220], and a post-processing algorithm to improve user-based fairness via calibrated recommendations [272].

5.3 Unsupervised Methods

Currently, unsupervised methods fall into three distinct areas: (1) fair clustering (e.g., References [7, 8, 15, 22, 23, 62, 67, 179, 248, 258, 269]); (2) investigating the presence and detection of discrimination in association rule mining (e.g., References [136, 230, 231]); and (3) transfer learning (e.g., References [78, 90]).

Fair clustering started with the initial work of Reference [67], which introduced the idea of micro-cluster fairlet decomposition as a pre-processing stage applied prior to standard centroid-based methods such as k-means and k-medians. Many clustering approaches have operated on Reference [98]'s disparate impact introducing this as cluster balance, where balance pertains to uniformity of distribution over k clusters of belonging to some protected group. Reference [67] uses color to represent belonging to the protected group or not. When multiple protected groups are in place, this means optimizing for both the number of clusters and the number as well as spread of colors. This is undertaken by Reference [22], which extends the work of Reference [67] 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 [258] and approximate fairlet decomposition [15]. Fair clustering has also seen applications in the discovery of potentially protected groups (e.g., References [21, 68]) and as a pre-processing method that augments the original data to achieve fairness [64]. For a more detailed overview of fairness in clustering, we refer to Reference [63].

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, 240, 273]. Here, transfer learning acts as an unsupervised domain adaption technique to account for such covariate shifts [90, 121, 281]. 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., Reference [78]) or forms of joint loss functions (e.g., Reference [90]).

6 CONCLUDING REMARKS AND KEY FUTURE WORK: THE FAIRNESS DILEMMAS

In this article, we have provided an introduction to the domain of fairness in ML research. This encompasses a general introduction (Section 2), different measures of fairness for ML (Section 3), and methods to mitigate bias and unfairness in binary classification problems (Section 4) as well as beyond binary classification (Section 5). We also list some key open-source tools in the online appendix 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 more general challenges for the domain as a set of five dilemmas for future research (the ordering is coincidental): **Dilemma-1:** Balancing the tradeoff between fairness and model performance (Section 6.1); **Dilemma-2:** Quantitative notions of fairness permit model optimization, yet cannot balance different notions of fairness (Section 6.2); **Dilemma-3:** Tensions between fairness, situational, ethical, and sociocultural context and policy (Section 6.3); **Dilemma-4:** Recent advances to the state-of-the-art have increased the skills gap inhibiting "on-the-street" and industry uptake (Section 6.4); and **Dilemma-5:** The challenge of both advancing the state-of-the-art and addressing real-world data contexts (Section 6.5).

6.1 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 [37, 306]. 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, 54, 77, 89, 135, 139, 319]. As noted by Reference [98], however, a reduction in accuracy may in fact be the desired result if it was discrimination that raised accuracy in the first place. Also, even prior to recognizing this tradeoff, 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 [84, 267]. 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 References [15, 258]). 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.

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

On top of the performance tradeoff, there is no consensus in literature whether individual or group fairness should be prioritized. Sadly, we cannot combine both [69, 177]. Reference [270] also notes that often approaches to group fairness tackle between-group issues, worsening within-group issues through this choice. To further complicate things, References [76, 128] argue that with a reliance on expressing fairness mathematically, these definitions often do not map to normative social, economic, or legal understandings of fairness. This is corroborated by Reference [266], which notes 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 differing views of fairness and bias. References [131, 236, 237, 271] study the differing views of people in this regard and observe that this is not a trivial challenge to address; e.g., Reference [237] notes that women have differing views in the inclusion/exclusion of gender as a protected variable to men. Reference [147] notes 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. Reference [164] has 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 References [47, 61, 140], 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 categorize specific differences, ideological tradeoffs, 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 while the tools discussed in the online appendix go some way to facilitate this, there is a need for more general toolkits and methodologies for comparing fairness approaches. We commendably note a number of comparative studies, i.e., References [109, 113, 282], but these only scratch the surface.

6.3 Dilemma-3: Tensions with Context and Policy

The literature typically hints toward "optimizing" fairness without transparency of the root(s) of (un)fairness [192] rarely extending beyond "(un)fair" [80, 270] typically to mirror current legal thought [98]. 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 [50, 193, 266, 286, 287, 297, 297, 306] 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, 76, 77]. This suggests a disconnect between organizational realities and current research, which undermines advancements [197, 287]. Other suggestions have been for ML researchers and practitioners to better understand the limitations of human decision making [245].

It is easy to criticize, however, the underlying challenge is a lack of realistic data. Currently, the literature relies unilaterally on convenience datasets (enabling comparative studies), often from the UCI repository [13] or similar with limited industry context and engagement [193, 286, 287]. References [163, 172, 186, 187, 298] 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 [186]. 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 Reference [240]. 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 [111, 268], and policy. Reference [287] 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.

6.4 Dilemma-4: Democratization of ML vs. the Fairness Skills Gap

Today, ML technologies are more accessible than ever. This has occurred through a combination of a surge in third-level courses and the wide availability of point-and-click tools such as WEKA [138], RapidMiner,[9] and SPSS Modeler.[10] Alternatively, Cloud-based solutions such as Google's Cloud AutoML [34], Uber AI's Ludwig,[11] and Baidu's EZDL[12] remove the need to even run models locally. The no-/low-code ML movement (efforts to make building ML models more accessible to non-experts, requiring no or little coding skills) is arguably enabling more companies to adopt ML technologies. In addition, there is a growing trend in the use of **Automated Machine Learning (AutoML)** [104, 279] to train ML models. AutoML abstracts much of the core methodological expertise (e.g., KDD [97] and CRISP-DM [60]) by automated feature extraction and training multiple models, often combining them into an ensemble of models that maximizes a set of performance measures. This positively democratizes ML, as it means lower barriers of use: "push button operationalization" [14] with online marketplaces[13] and APIs/services like GPT-4, DALL-E, and so on.

Lowering the entry barrier to ML through democratization means 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 [192, 287]. This was our motivation in writing this survey. However, the fairness domain has relatively few open-source tools available for practitioners (see online appendix), and there is little accommodation for varying levels of technical proficiency, which undermines current advancement [50, 124, 266, 286, 287]. There is a tension between educational programs (as called for in Reference [50]) and the degree of proficiency needed to apply methods and methodologies for fair ML. References [76, 286] 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, Reference [255] calls 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 is the democratization of fairness in ML. This means a shift in terms of scientific reporting, open-source frameworks, and multi-stage (i.e., where one ML model is downstream to another) decision-making processes [40, 124, 277]; the latter is significantly underaddressed in the literature.

[9] https://rapidminer.com
[10] https://www.ibm.com/ie-en/products/spss-modeler
[11] https://uber.github.io/ludwig/
[12] https://ai.baidu.com/ezdl/
[13] E.g.: Amazon's https://aws.amazon.com/marketplace/solutions/machinelearning/ and Microsoft's https://gallery.azure.ai ML Marketplaces.

6.5 Dilemma-5: Scientific Advancement vs. the Reality of Data

There is one common theme throughout the literature: the assumption that the data used for the evaluation of new mitigation strategies have already been suitably prepared. It is not present in all papers, of course, yet a significant portion of publications relies on a small set of common datasets that are either already prepared or do not require significant data preparation (see, e.g., Reference [189]).

As data preparation is fundamental to ML, the fairness literature is slowly exhibiting research that considers forms of data preparation that are not per se fairness interventions (thus, excluded from Figure 2). One such example is treating missing data. Many approaches discussed throughout this article assume (often implicitly) that data are both complete and clean. Realistically, this will never be the case. Specifically for missing data, Reference [102] provides a comprehensive discussion on the considerations of missing data on fairness, Reference [55] illustrates how different methods of treating missing data have different effects on fairness, and Reference [295] discusses selection biases (and their impact on fairness) for categorical data in the presence of missing data. There is worryingly little literature to discuss: We cannot comprehend intervention robustness under real-world data challenges.

Investigating the impacts of real-world data issues would slow down the development of new methods, require new forms of benchmarking and comparative studies akin to Reference [101]. We would expect many of these papers to be "unpopular," as they may illustrate that approaches we believed to be "good" are not under different contextual assumptions. Similar to Dilemma-3, we would advocate that new(er) approaches are encouraged to evaluate their work in the presence of more realistic scenarios or use more common forms of data preparation to provide a more holistic evaluation.

6.6 Concluding Remarks and Moving the Literature Forwards

In this article, we focus primarily on introducing the main areas of research into making ML fairer, with the goal of acting as a means to on-board researchers and practitioners new to the area or generally interested in it. It should be clear that there has been a tremendous amount of research into trying to improve the fairness of ML models. The literature almost unilaterally focuses on supervised learning with an overwhelming emphasis on binary classification [26]: Diversification is needed. However, this diversification should occur at the level of addressing unsolved research challenges, many of which we have highlighted in the discussion of the dilemmas above. We also note that there is a current trend in the literature to create new fairness measures as well as interventions. Instead, we would argue that the development of methodologies to identify "good" solutions be emphasized. Linked to this, and 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 does not help practitioners. Tools potentially based on causal methods (Section 4.2) are needed to assist in the identification of protected variables and groups as well as their proxies.

More realistic datasets are needed: Reference [242] argues 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, 76], rendering the model as intrinsically discriminatory as a random model [90, 239]. There is also the risk of redlining, where although the sensitive attribute is "handled" sufficiently, correlated variables are still present [54, 89, 231, 247, 286, 308], amplifying instead of reducing unfairness [89].

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 **General Data Protection Regulation (GDPR)**'s right to an explanation and can introduce

new subjectivity biases [286]. They also assume sufficient knowledge of the data and make assumptions over its veracity [76]. Uptake of in-processing approaches requires better integration with standard ML libraries to overcome porting challenges. Reference [298] noted that generally post-processing methods have suboptimal accuracy compared to other "equally fair" classifiers, with Reference [3] noting that often test-time access to protected attributes is needed, which may not be legally permissible and have other undesirable effects [61]. Linked to this discussion is the effect that specific forms of data preparation have on approaches. We have highlighted missing data (as an exercise in data cleaning), yet it is generally unclear how robust suggested approaches are to different forms of data preparation.

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 [61]. 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 References [61, 96, 199, 298] note, it may often be better to fix the underlying data sample (e.g., collect more data, which better represent minority or protected groups and delay the modeling phase) than try to fix a discriminatory ML model.

REFERENCES

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

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

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

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

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

[6] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 439–450.

[7] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. 2020. Fair correlation clustering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. PMLR, 4195–4205.

[8] Sara Ahmadian and Maryam Negahbani. 2022. Improved approximation for fair correlation clustering. *arXiv preprint arXiv:2206.05050* (2022).

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

[10] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V. Nori. 2017. FairSquare: Probabilistic verification of program fairness. *Proc. ACM Program. Lang.* 1, OOPSLA (2017), 1–30.

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

[12] Ashwathy Ashokan and Christian Haas. 2021. Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.* 58, 5 (2021), 102646.

[13] Arthur Asuncion and David Newman. 2007. UCI Machine Learning Repository. https://archive.ics.uci.edu/

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

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

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

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

[18] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. 2019. Probabilistic verification of fairness properties via concentration. *Proc. ACM Program. Lang.* 3, OOPSLA (2019), 1–27.

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

[20] Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Ling.* 6 (2018), 587–604.

[21] Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 289–298.
[22] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. 2019. Fair algorithms for clustering. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 4954–4965.
[23] Ioana O. Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. 2018. On the cost of essentially fair clusterings. *arXiv:1811.10319* (2018).
[24] Bettina Berendt and Sören Preibusch. 2014. Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artif. Intell. Law* 22, 2 (2014), 175–209.
[25] Richard Berk. 2019. Accuracy and fairness for juvenile justice risk assessments. *J. Empir. Legal Stud.* 16, 1 (2019), 175–194.
[26] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).
[27] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociolog. Meth. Res.* 50, 1 (2018), 0049124118782533.
[28] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).
[29] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'19)*. ACM, 2212–2220.
[30] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H. Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.
[31] Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *J. Mach. Learn. Res.* 10, Sep. (2009), 2137–2155.
[32] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. Association for Computing Machinery, New York, NY, 405–414.
[33] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 149–159. Retrieved from http://arxiv.org/abs/1712.03586
[34] Ekaba Bisong. 2019. Google AutoML: Cloud vision. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Springer, 581–598.
[35] Cody Blakeney, Gentry Atkinson, Nathaniel Huish, Yan Yan, Vangelis Metris, and Ziliang Zong. 2021. Measure twice, cut once: Quantifying bias and fairness in deep neural networks. *arXiv preprint arXiv:2110.04397* (2021).
[36] William Blanzeisky and Pádraig Cunningham. 2022. Using Pareto simulated annealing to address algorithmic bias in machine learning. *Knowl. Eng. Rev.* 37 (2022).
[37] Paula Boddington. 2017. *Towards a Code of Ethics for Artificial Intelligence*. Springer.
[38] Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick, and Jintong Tang. 2017. The law and policy of people analytics. *U. Colo. L. Rev.* 88 (2017), 961.
[39] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *J. Mach. Learn. Res.* 2, Mar. (2002), 499–526.
[40] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391* (2017).
[41] Danah Boyd and Kate Crawford. 2011. Six provocations for big data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, Vol. 21. Oxford Internet Institute Oxford, UK.
[42] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf., Commun. Soc.* 15, 5 (2012), 662–679.
[43] Frédéric Branchaud-Charron, Parmida Atighehchian, Pau Rodríguez, Grace Abuhamad, and Alexandre Lacoste. 2021. Can active learning preemptively mitigate fairness issues? *arXiv preprint arXiv:2104.06879* (2021).
[44] Tim Brennan and William L. Oliver. 2013. Emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. *Criminol. Pub. Polic.* 12 (2013), 551.
[45] Bénédicte Briand, Gilles R. Ducharme, Vanessa Parache, and Catherine Mercat-Rommens. 2009. A similarity measure to assess the stability of classification trees. *Comput. Stat. Data Anal.* 53, 4 (2009), 1208–1217.
[46] J. Paul Brooks. 2011. Support vector machines with the ramp loss and the hard margin loss. *Oper. Res.* 59, 2 (2011), 467–479.
[47] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. 77–91.
[48] Robin Burke. 2017. Multisided Fairness for Recommendation. arXiv:1707.00093

[49] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 202–214.
[50] Jenna Burrell. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data Soc.* 3, 1 (2016).
[51] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *Proceedings of the IEEE 13th International Conference on Data Mining*. 71–80.
[52] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21, 2 (Sep. 2010), 277–292.
[53] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society*. Springer, Berlin, 43–57.
[54] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 3992–4001.
[55] Simon Caton, Saiteja Malisetty, and Christian Haas. 2022. Impact of imputation strategies on fairness in machine learning. *J. Artif. Intell. Res.* 74 (2022), 1011–1035.
[56] L. Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2016. How to be fair and diverse? *arXiv preprint arXiv:1610.07183* (2016).
[57] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328.
[58] L. Elisa Celis and Vijay Keswani. 2019. Improved adversarial learning for fair classification. *arXiv:1901.10443* (2019).
[59] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced Top-K recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*. ACM Press, New York, NY, 129–138.
[60] R. Wirth and J. Hipp. 2000. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, Vol. 1, 29–39.
[61] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *Adv. Neural Inf. Process. Syst.* 31 (2018).
[62] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally fair clustering. In *Proceedings of the International Conference on Machine Learning*. 1032–1041.
[63] Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. 2021. An overview of fairness in clustering. *IEEE Access* 9 (2021).
[64] Anshuman Chhabra, Adish Singla, and Prasant Mohapatra. 2022. Fair clustering using antidote data. In *Proceedings of the Algorithmic Fairness through the Lens of Causality and Robustness Workshop*. PMLR, 19–39.
[65] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.
[66] Silvia Chiappa and William S. Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *Proceedings of the IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.
[67] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 5029–5037.
[68] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvtiskii. 2019. Matroids, matchings, and fairness. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. 2212–2220.
[69] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163. arXiv:stat.AP/1610.07524
[70] Alexandra Chouldechova and Max G'Sell. 2017. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046* (2017).
[71] Kevin A. Clarke. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Manag. Peace Sci.* 22, 4 (2005), 341–352.
[72] T. Cleary. 1966. Test bias: Validity of the scholastic aptitude test for Negro and White students in integrated colleges. *ETS Res. Bull. Series* 1966 (1966). DOI: https://doi.org/10.1002/j.2333-8504.1966.tb00529.x
[73] T. Anne Cleary. 1968. Test bias: Prediction of grades of negro and white students in integrated colleges. *J. Educ. Measur.* 5, 2 (1968), 115–124. Retrieved from http://www.jstor.org/stable/1434406
[74] Nancy S. Cole. 1973. Bias in selection. *J. Educ. Measur.* 10, 4 (1973), 237–255.
[75] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics Guidelines for Trustworthy AI*. Publications Office. DOI: https://doi.org/doi/10.2759/177365
[76] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

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

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

[79] Andrew Cotter, Heinrich Jiang, Maya R. Gupta, Serena Lutong Wang, Taman Narayan, Seungil You, and Karthik Sridharan. 2019. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.* 20, 172 (2019), 1–59.

[80] Bo Cowgill and Catherine Tucker. 2017. Algorithmic bias: A counterfactual perspective. *NSF Trustwor. Algor.* (2017).

[81] Richard B. Darlington. 1971. Another look at "Cultural Fairness." *J. Educ. Measur.* 8, 2 (1971), 71–82.

[82] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'16)*. IEEE, 598–617.

[83] A. Philip Dawid. 1982. The well-calibrated Bayesian. *J. Am. Stat. Assoc.* 77, 379 (1982), 605–610.

[84] Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* 7, Jan. (2006), 1–30.

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

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

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

[88] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2018. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE J. Select. Topics Sig. Process.* 12, 5 (2018), 1106–1119.

[89] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

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

[91] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[92] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in information access systems. *Found. Trends Inf. Retriev.* 16 (2022).

[93] Micheal D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 172–186.

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

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

[96] Danielle Ensign, Frielder Sorelle, Neville Scott, Scheidegger Carlos, and Venkatasubramanian Suresh. 2018. Runaway feedback loops in predictive policing. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 160–171.

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

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

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

[100] Andrew Guthrie Ferguson. 2015. Big data and predictive reasonable suspicion. *University of Pennsylvania Law Review* (2015), 327–410.
[101] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* 15, 1 (2014), 3133–3181.
[102] Martínez-Plumed Fernando, Ferri Cèsar, Nieves David, and Hernández-Orallo José. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *Int. J. Intell. Syst.* 36, 7 (2021).
[103] Julien Ferry, Ulrich Aivodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2022. Improving fairness generalization through a sample-robust optimization method. *Mach. Learn.* 12, 6 (2022), 1–62.
[104] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2962–2970.
[105] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 144–152.
[106] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. 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).
[107] Jack Fitzsimons, AbdulRahman Al Ali, Michael Osborne, and Stephen Roberts. 2019. A general framework for fair regression. *Entropy* 21, 8 (2019), 741.
[108] Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. 2022. An ontology for fairness metrics. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 265–275.
[109] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.
[110] Kazuto Fukuchi, Jun Sakuma, and Toshihiro Kamishima. 2013. Prediction with model-based neutrality. In *Machine Learning and Knowledge Discovery in Databases*, Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný (Eds.). IEEE, 499–514.
[111] Benjamin CM Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* 42, 4 (2010), 1–53.
[112] Pratik Gajane, Akrati Saxena, Maryam Tavakol, George Fletcher, and Mykola Pechenizkiy. 2022. Survey on fair reinforcement learning: Theory and practice. *arXiv preprint arXiv:2205.10032* (2022).
[113] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: Testing software for discrimination. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*. 498–510.
[114] Bo Gao. 2015. *Exploratory Visualization Design towards Online Social Network Privacy and Data Literacy*. Ph.D. Dissertation. KU Leuven.
[115] Xuanqi Gao, Juan Zhai, Shiqing Ma, Chao Shen, Yufei Chen, and Qian Wang. 2022. FairNeuron: Improving deep neural network fairness with adversary games on selective neurons. *arXiv preprint arXiv:2204.02567* (2022).
[116] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. 2018. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2600–2609.
[117] Amir Globerson and Sam Roweis. 2006. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 353–360.
[118] Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 269–278.
[119] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
[120] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P. Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2415–2423.
[121] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2066–2073.
[122] Sandra Gonzélez-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Soc. Netw.* 38 (2014), 16–27.
[123] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2672–2680.
[124] Bryce W. Goodman. 2016. Economic models of (algorithmic) discrimination. In *Proceedings of the 29th Conference on Neural Information Processing Systems*, Vol. 6.

[125] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. 2019. Obtaining fairness using optimal transport theory. In *Proceedings of the International Conference on Machine Learning*. 2357–2365.
[126] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 1905–1916.
[127] Riccardo Grazzi, Arya Akhavan, John IF Falk, Leonardo Cella, and Massimiliano Pontil. 2022. Group meritocratic fairness in linear contextual bandits. *arXiv preprint arXiv:2206.03150* (2022).
[128] Ben Green. 2018. "Fair" risk assessments: A precarious approach for criminal justice reform. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*.
[129] Ben Green. 2020. 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. DOI: https://doi.org/10.1145/3351095.3372869
[130] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*'20)*.
[131] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the World Wide Web Conference*. 903–912.
[132] Adam Gronowski, William Paul, Fady Alajaji, Bahman Gharesifard, and Philippe Burlina. 2022. Achieving utility, fairness, and compactness via tunable information bottleneck measures. *arXiv preprint arXiv:2206.10043* (2022).
[133] Robert M. Guion. 1966. Employment tests and discriminatory hiring. *Industr. Relat.: J. Econ. Soc.* 5, 2 (1966), 20–37.
[134] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. 2018. Proxy fairness. *arXiv:1806.11212* (2018).
[135] Christian Haas. 2019. The price of fairness—A framework to explore trade-offs in algorithmic fairness. In *Proceedings of the International Conference on Information Systems (ICIS'19)*.
[136] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.* 25, 7 (2012), 1445–1459.
[137] Margeret Hall and Simon Caton. 2017. Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. *PloS One* 12, 9 (2017), e0184417.
[138] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explor. Newslett.* 11, 1 (2009), 10–18.
[139] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 3315–3323.
[140] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2017. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513* (2017).
[141] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Adv. Neural Inf. Process. Syst.* 31 (2018).
[142] Hoda Heidari and Andreas Krause. 2018. Preventing disparate treatment in sequential decision making. In *Proceedings of the International Joint Conferences on Artificial Intelligence*. 2248–2254.
[143] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: Exploring classifiers by randomization. *Data Mining Knowl. Discov.* 28, 5–6 (2014), 1503–1529.
[144] Safwan Hossain, Evi Micha, and Nisarg Shah. 2021. Fair algorithms for multi-agent multi-armed bandits. *Adv. Neural Inf. Process. Syst.* 34 (2021), 24005–24017.
[145] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and fair classification. In *Proceedings of the International Conference on Machine Learning*. 2879–2890.
[146] Wen Huang, Kevin Labille, Xintao Wu, Dongwon Lee, and Neil Heffernan. 2022. Achieving user-side fairness in contextual bandits. *Hum.-centr. Intell. Syst.* 2, 3 (2022), 1–14.
[147] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (Un)Fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*. ACM, 49–58.
[148] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2020. FAE: A fairness-aware ensemble framework. *arXiv preprint arXiv:2002.00695* (2020).
[149] Vasileios Iosifidis and Eirini Ntoutsi. 2019. AdaFair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 781–790.
[150] Yasser Jafer, Stan Matwin, and Marina Sokolova. 2014. Privacy-aware filter-based feature selection. In *Proceedings of the IEEE International Conference on Big Data (Big Data'14)*. IEEE, 1–5.
[151] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Model. User-adapt. Interact.* 25, 5 (Dec. 2015), 427–491.
[152] Heinrich Jiang and Ofir Nachum. 2019. Identifying and correcting label bias in machine learning. Retrieved from https://arxiv.org/pdf/1901.04966.pdf

[153] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2019. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059* (2019).
[154] James E. Johndrow and Kristian Lum. 2019. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Ann. Appl. Stat.* 13, 1 (2019), 189–220.
[155] Kory D. Johnson, Dean P. Foster, and Robert A. Stine. 2016. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528* (2016).
[156] Matthew Joseph, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 325–333.
[157] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2018. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 158–163.
[158] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2020. *Fair Prediction with Endogenous Behavior*. Technical Report. arXiv. org.
[159] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. Simple rules for complex decisions. *Available at SSRN 2919024* (2017).
[160] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. 2018. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651* (2018).
[161] Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. 2018. Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868* (2018).
[162] Peter Kairouz, Jiachun Liao, Chong Huang, and Lalitha Sankar. 2019. Censored and fair universal representations using generative adversarial models. *arXiv* (2019), arXiv–1910. https://doi.org/10.48550/arXiv.1910.00411
[163] Nathan Kallus. 2018. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 8895–8906.
[164] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).
[165] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33, 1 (Oct. 2012), 1–33.
[166] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE, 869–874.
[167] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *Proceedings of the IEEE 12th International Conference on Data Mining*. IEEE, 924–929.
[168] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Springer, Berlin, 35–50.
[169] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 187–201.
[170] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the International Conference on Machine Learning* . 2564–2572.
[171] Michael Kearns and Dana Ron. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Comput.* 11, 6 (1999), 1427–1453.
[172] Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2019. Fair decisions despite imperfect predictions. *arXiv preprint arXiv:1902.02979* (2019).
[173] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. arXiv:stat.ML/1706.02744
[174] Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 4842–4852.
[175] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.
[176] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *Quart. J. Econ.* 133, 1 (2018), 237–293.
[177] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 108. 22–27.
[178] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Innov. Theor. Comput. Sci.* (2017). https://drops.dagstuhl.de/opus/volltexte/2017/8156/
[179] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2020. A notion of individual fairness for clustering. *arXiv preprint arXiv:2006.04960* (2020).
[180] Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. 2018. Nonconvex optimization for regression with fairness constraints. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2737–2746.

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

[182] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. 4069–4079.

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

[184] Samuel Kutin and Partha Niyogi. 2002. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. 275–282.

[185] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning individually fair data representations for algorithmic decision making. In *Proceedings of the IEEE 35th International Conference on Data Engineering (ICDE'19)*. 1334–1345.

[186] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. 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*. 275–284.

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

[188] Luis Lämmermann, Patrick Richter, Amelie Zwickel, and Moritz Markgraf. 2022. AI fairness at subgroup level—A structured literature review. In *Proceedings of the 30th European Conference on Information Systems (ECIS'22)*. 147.

[189] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* 12, 3 (2022), e1452.

[190] Eric L. Lee, Jing-Kai Lou, Wei-Ming Chen, Yen-Chi Chen, Shou-De Lin, Yen-Sheng Chiang, and Kuan-Ta Chen. 2014. Fairness-aware loan recommendation for microfinance services. In *Proceedings of the International Conference on Social Computing (SocialCom'14)*. 1–4.

[191] Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *J. Inf., Commun. Ethics Soc.* 6, 3 (2018).

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

[193] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. 2017. 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*. Springer, 3–24.

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

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

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

[197] Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 30.

[198] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2021. Trustworthy AI: A computational perspective. *arXiv preprint arXiv:2107.06641* (2021).

[199] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3150–3158.

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

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

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

[203] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30 (2017).

[204] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 502–510.

[205] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1 (Mar. 2007), 3–es.

[206] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning*. 3381–3390.

[207] Gaurav Maheshwari and Michaël Perrot. 2022. FairGrad: Fairness aware gradient descent. *arXiv preprint arXiv:2206.10923* (2022).
[208] P. Manisha and Sujit Gujar. 2018. A neural network framework for fair classifier. *arXiv:1811.00247* (2018).
[209] Olivera Marjanovic, Dubravka Cecez-Kecmanovic, and Richard Vidgen. 2018. Algorithmic pollution: Understanding and responding to negative consequences of algorithmic decision-making. In *Proceedings of the Working Conference on Information Systems and Organizations*. Springer, 31–47.
[210] Douglas S. Massey and Nancy A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Harvard University Press.
[211] Bruce McKeown and Dan B. Thomas. 2013. Q methodology. *Quantit. Applic. Soc. Sci.* 66 (2013).
[212] Douglas S. McNair. 2018. Preventing disparities: Bayesian and frequentist methods for assessing fairness in machine-learning decision-support models. *New Insights Bayes. Infer.* IntechOpen.
[213] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[214] Aditya Krishna Menon and Robert C. Williamson. 2017. The cost of fairness in classification. *arXiv:1705.09055* (2017).
[215] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S. Thomas. 2019. Offline contextual bandits with high probability fairness guarantees. *Adv. Neural Inf. Process. Syst.* 32 (2019).
[216] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).
[217] Rajeev Motwani and Ying Xu. 2007. Efficient algorithms for masking and finding quasi-identifiers. In *Proceedings of the Conference on Very Large Data Bases (VLDB'07)*. 83–93.
[218] Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. 2019. Learning optimal fair policies. *Proc. Mach. Learn. Res.* 97 (2019), 4674.
[219] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.
[220] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. 2022. CPFair: Personalized consumer and producer fairness re-ranking for recommender systems. *arXiv preprint arXiv:2204.08085* (2022).
[221] Harikrishna Narasimhan. 2018. Learning with complex loss functions and constraints. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 1646–1654.
[222] Tan Nguyen and Scott Sanner. 2013. Algorithms for direct 0–1 loss optimization in binary classification. In *Proceedings of the International Conference on Machine Learning*. 1085–1093.
[223] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. 2019. Active fairness in algorithmic decision making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 77–83.
[224] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.
[225] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. 2019. Taking advantage of multitask learning for fair classification. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 227–237.
[226] Kalia Orphanou, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Fausto Giunchiglia, Veronika Bogina, Avital Shulner Tal, Alan Hartman, and Tsvi Kuflik. 2021. Mitigating bias in algorithmic systems—A fish-eye view. *ACM Comput. Surv.* 55, 5 (2021).
[227] Mustafa Safa Ozdayi, Murat Kantarcioglu, and Rishabh Iyer. 2021. BiFair: Training fair models with bilevel optimization. *arXiv preprint arXiv:2106.04757* (2021).
[228] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. 2021. Achieving fairness in the stochastic multi-armed bandit problem. *J. Mach. Learn. Res.* 22 (2021), 174–1.
[229] William Paul, Armin Hadzic, Neil Joshi, Fady Alajaji, and Philippe Burlina. 2022. TARA: Training and representation alteration for AI fairness and domain generalization. *Neural Comput.* 34, 3 (2022), 716–753.
[230] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2012. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. 126–131.
[231] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 560–568.
[232] Anjana Perera, Aldeida Aleti, Chakkrit Tantithamthavorn, Jirayus Jiarpakdee, Burak Turhan, Lisa Kuhn, and Katie Walker. 2022. Search-based fairness testing for regression-based machine learning systems. *Empir. Softw. Eng.* 27, 3 (2022), 1–36.
[233] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. 2017. Fair kernel learning. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 339–355.

[234] Andrija Petrović, Mladen Nikolić, Sandro Radovanović, Boris Delibašić, and Miloš Jovanović. 2022. FAIR: Fair adversarial instance re-weighting. *Neurocomputing* 476 (2022), 14–37.
[235] Alexander Peysakhovich and Christian Kroer. 2019. Fair division without disparate impact. *arXiv:1906.02775* (2019).
[236] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv:1712.09124* (2017).
[237] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. *arXiv:1712.09124* (2017).
[238] Pitoura Evaggelia, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *ACM SIGMOD Rec.* 46, 4 (2018), 16–21.
[239] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 5680–5689.
[240] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. *Dataset Shift in Machine Learning*. The MIT Press.
[241] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 469–481.
[242] Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. 2016. Information-theoretic analysis of stability and bias of learning algorithms. In *Proceedings of the IEEE Information Theory Workshop (ITW'16)*. IEEE, 26–30.
[243] Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. 2005. Stability results in learning theory. *Anal. Applic.* 3, 4 (2005), 397–417.
[244] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting fire with fire. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM'19)*. 231–239.
[245] Alexander S. Rich and Todd M. Gureckis. 2019. Lessons for artificial intelligence from the study of natural stupidity. *Nat. Mach. Intell.* 1, 4 (2019), 174–180.
[246] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2021. FairBatch: Batch selection for model fairness. In *9th International Conference on Learning Representations*.
[247] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.* 29, 5 (2014), 582–638.
[248] Clemens Rösner and Melanie Schmidt. 2018. Privacy preserving clustering with constraints. *arXiv:1802.02497* (2018).
[249] Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. 2013. Representation and communication: Challenges in interpreting large social media datasets. In *Proceedings of the ACM Computer Supported Cooperative Work (CSCW'13)*. 357–362.
[250] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 6414–6423.
[251] Babak Salimi, Bill Howe, and Dan Suciu. 2019. Data management for causal algorithmic fairness. *Data Eng.* 42, 3 (2019), 24.
[252] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283* (2019).
[253] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the International Conference on Management of Data*. 793–810.
[254] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. 2004. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Wiley, New York.
[255] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. 2019. Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.* 19, 4 (2019), 1–26.
[256] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM J. Res. Devel.* 63, 4/5 (2019), 3–11.
[257] Richard L. Sawyer, Nancy S. Cole, and James W. L. Cole. 1976. Utilities and the issue of fairness in a decision theoretic model for selection. *J. Educ. Measur.* 13, 1 (1976), 59–76.
[258] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. 2019. Fair coresets and streaming algorithms for fair k-means. In *Proceedings of the International Workshop on Approximation and Online Algorithms*. Springer, 232–251.
[259] Hansen Andrew Schwartz, Johannes C. Eichstaedt, Lukasz Dziurzynski, Margaret L. Kern, Eduardo Blanco, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Toward personality insights from language exploration in social media. In *Proceedings of the AAAI Spring Symposium Series*.
[260] Andrew D. Selbst. 2017. Disparate impact in big data policing. *Ga. L. Rev.* 52 (2017), 109.
[261] Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078* (2019).

[262] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2010. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* 11, Oct. (2010), 2635–2670.
[263] Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. 2021. Adaptive sampling for minimax fair classification. *Adv. Neural Inf. Process. Syst.* 34 (2021), 24535–24544.
[264] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18)*. 2219–2228.
[265] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems 32*. 5427–5437.
[266] Michael Skirpan and Micha Gorelick. 2017. The authority of "Fair" in machine learning. *arXiv:1706.09976* (2017).
[267] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence*, Abdul Sattar and Byeong-ho Kang (Eds.). Springer Berlin, 1015–1021.
[268] Ana Sokolovska and Ljupco Kocarev. 2018. Integrating technical and legal concepts of privacy. *IEEE Access* 6 (2018), 26543–26557.
[269] Hanyu Song, Peizhao Li, and Hongfu Liu. 2021. Deep clustering based fair outlier detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1481–1489.
[270] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18)*. 2239–2248.
[271] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. 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*. 2459–2468.
[272] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys'18)*. Association for Computing Machinery, New York, NY, 154–162.
[273] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* 8, May (2007), 985–1005.
[274] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).
[275] Harini Suresh and John V. Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).
[276] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *Int. J. Uncertain., Fuzz. Knowl.-based Syst.* 10, 5 (2002), 557–570.
[277] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (2019), 999–1004.
[278] Robert L. Thorndike. 1971. Concepts of culture-fairness. *J. Educ. Measur.* 8, 2 (1971), 63–70.
[279] Thornton, Chris, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 847–855.
[280] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law*. 83–92.
[281] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 242–264.
[282] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P'17)*. 401–416.
[283] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *Proceedings of the International Conference on Machine Learning*. 6373–6382.
[284] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. 2018. Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems 31*. 1769–1778.
[285] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2020. Hiring algorithms: An ethnography of fairness in practice. In *40th International Conference on Information Systems (ICIS'19)*. Association for Information Systems, 1–9.
[286] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data Soc.* 4, 2 (2017).

[287] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the Chi Conference on Human Factors in Computing Systems*. 1–14.

[288] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the IEEE/ACM International Workshop on Software Fairness (FairWare'18)*. 1–7.

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

[290] Hao Wang, Snehasis Mukhopadhyay, Yunyu Xiao, and Shiaofen Fang. 2021. An interactive approach to bias mitigation in machine learning. In *Proceedings of the IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC'21)*. IEEE, 199–205.

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

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

[293] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. 2021. Fairness of exposure in stochastic bandits. In *Proceedings of the International Conference on Machine Learning*. PMLR, 10686–10696.

[294] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.

[295] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *Int. J. Data Sci. Anal.* 12, 2 (2021), 1–19.

[296] Lauren Weber and Elizabeth Dwoskin. 2014. Are workplace personality tests fair. *Wall Street J.* 29 (2014).

[297] Pak-Hang Wong. 2019. Democratizing algorithmic fairness. *Philos. Technol.* 33, 2 (2019), 1–20.

[298] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Proceedings of the Conference on Learning Theory*. 1920–1953.

[299] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving causal fairness through generative adversarial networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.

[300] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware generative adversarial networks. In *Proceedings of the IEEE International Conference on Big Data (Big Data'18)*. IEEE, 570–575.

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

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

[303] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems 30*. 2921–2930.

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

[305] Samuel Yeom and Matt Fredrikson. 2020. Individual fairness revisited: Transferring techniques from adversarial robustness. *arXiv preprint arXiv:2002.07738* (2020).

[306] Karen Yeung. 2018. Algorithmic regulation: A critical interrogation. *Regul. Govern.* 12, 4 (2018), 505–523.

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

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

[309] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair Top-k ranking algorithm. In *Proceedings of the ACM on Conference on Information and Knowledge Management (CIKM'17)*. Association for Computing Machinery, New York, NY, 1569–1578.

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

[311] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part I: Score-based ranking. *ACM Comput. Surv.* 55, 6 (2022).

[312] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Comput. Surv.* 55, 6 (2022).

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

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

[315] Zhe Zhang and Daniel B. Neill. 2016. Identifying significant predictive bias in classifiers. *arXiv:1611.08292* (2016).
[316] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. 2018. Fairness in reciprocal recommendations: A speed-dating study. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. 29–34.
[317] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1153–1162.
[318] Michael Zimmer. 2010. "But the data is already public": On the ethics of research in Facebook. *Ethics Inf. Technol.* 12, 4 (2010), 313–325.
[319] Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. arXiv:1505.05723
[320] Andrej Zwitter. 2014. Big data ethics. *Big Data Soc.* 1, 2 (2014), 1–6.

Received 14 October 2022; revised 16 May 2023; accepted 3 August 2023