AI and Ethics (2025) 5:1943–1954
https://doi.org/10.1007/s43681-024-00577-5

REVIEW

A systematic review of fairness in machine learning

Ricardo Trainotti Rabonato[1] · Lilian Berton[1]

Received: 8 June 2024 / Accepted: 4 September 2024 / Published online: 19 September 2024

Abstract
Fairness in Machine Learning (ML) has emerged as a crucial concern as these models increasingly influence critical decisions in various domains, including healthcare, finance, and criminal justice. The presence of bias in ML systems can lead to unfair and discriminatory outcomes, undermining the reliability and ethical standards of these technologies. As the deployment of ML expands, ensuring that these systems are fair and unbiased is not only a technical challenge but also a moral imperative. Here, a systematic literature review was conducted to explore fairness in machine learning, utilizing the ACM, IEEE, and Springer databases. From an initial retrieval of 975 papers, 30 were included in the review. The results highlight the identification of sensitive attributes, the metrics used to assess bias, and the various databases tested. Additionally, the review categorizes the in-processing and post-processing approaches employed to mitigate bias and examines how studies are managing the trade-off between fairness and accuracy. This comprehensive analysis provides a detailed understanding of the current state of fairness in machine learning and offers insights into effective strategies for bias mitigation.

Keywords Machine learning · Fairness metrics · Sensitive attributes · Bias

1 Introduction

Recent experiments have shown that, just as can happen with human thinking, machines can learn undesirable behaviors. In this context, it is crucial to understand the concept of bias in Artificial Intelligence and Machine Learning (ML). This bias, which often refers to prior information - a necessary prerequisite for intelligent action - can become problematic when it is derived from aspects of human culture known to lead to harmful behaviors [1]. In this work, the term bias will usually be used as a reference to these biases originating from cultural stereotypes, whose actions based on these biases can be prejudiced.

The problem of bias in ML is so relevant that the *World Economic Forum Global Future Council on Human Rights* 2016–18 published a *white paper* on (*How to Prevent Discriminatory Outcomes in Machine Learning*) [2]. From this same perspective, the Association for the Advancement of Artificial Intelligence (AAAI), in its code of ethics and professional conduct says:

> ”The use of information and technology can cause new inequalities, or increase existing ones. Technologies and practices must be as inclusive and accessible as possible, and AI professionals must take steps to avoid creating systems or technologies that deprive or oppress people of their rights. The inability to design inclusion and accessibility may constitute unfair discrimination” [3].

Therefore, there is a need to find ways to identify and minimize the problem of bias. According to [4], to ensure that decisions made by supervised learning methods are less biased, it is necessary to modify one or more of the following aspects: (a) the training data, (b) the learning algorithms, and (c) the decisions that follow. These techniques are, respectively, classified as pre-processing, in-processing, and post-processing approaches.

The problem of bias in ML has been addressed by an area of research called *fairness*, which refers to the search for machine learning models that are impartial and non-discriminatory. The objective is to ensure that the decisions made by algorithms are not influenced by characteristics such as gender, race, belief, etc. If ML is a way of studying institutional decision-making, *fairness* is the moral lens through which we examine these decisions [5].

✉ Lilian Berton
lberton@unifesp.br

[1] Universidade Federal de São Paulo, São José dos Campos, São Paulo, Brazil

To better understand the current landscape of fairness in ML, we conducted a systematic literature review. This review retrieved papers from leading computer science-focused libraries, namely ACM, Springer, and IEEE. From an initial pool of 975 papers, 30 were selected based on relevance and quality criteria. Our review aims to address five specific research questions:

- What techniques are being used to mitigate bias in machine learning?
- What bias assessment metrics are used?
- What databases are being tested?
- What techniques are focused on in-processing and post-processing?
- How are studies dealing with the trade-off between bias mitigation and accuracy?

The motivation behind this systematic literature review stems from the need to consolidate current knowledge and identify gaps in the research on fairness in machine learning. While several previous surveys have explored various aspects of fairness, our review aims to provide a comprehensive examination focusing specifically on bias mitigation techniques, assessment metrics, databases used, and the balance between fairness and accuracy.

Some previous reviews focus on ethical AI for healthcare. In [6] the focus was on issues related to data acquisition, genetic variation, and intra-observer labeling variability. This paper primarily addressed the sources of bias from a data-centric perspective. [7] examined the ethical considerations across the entire machine learning pipeline in health applications, from problem selection to post-deployment issues. This survey highlighted the ethical challenges at each stage of machine learning deployment but did not delve deeply into specific bias mitigation techniques. [8] provided a general overview of bias mitigation strategies within the context of medical imaging. This work was broader in scope and did not focus on the detailed technical approaches for mitigating bias.

Other reviews focus on fairness and ML, the most cited are [9] which analyzed real-world applications that have displayed various forms of bias, identifying their sources and impacts on AI systems. They developed a taxonomy of fairness definitions proposed by machine learning researchers to mitigate these biases. Additionally, they reviewed different AI domains and subdomains, highlighting observed unfair outcomes in state-of-the-art methods and the strategies employed to address them. [10] provides an overview of the key concepts in identifying, measuring, and improving algorithmic fairness in machine learning, particularly in classification tasks. It starts by exploring the causes of algorithmic bias and unfairness, along with common definitions and measures of fairness. The article then reviews fairness-enhancing mechanisms, categorizing them into pre-process, in-process, and post-process methods. [11] organize the approaches into the widely accepted framework of pre-processing, in-processing, and post-processing methods, further subcategorizing them into 11 specific areas. Besides binary classification, the article also discusses fairness in regression, recommender systems, and unsupervised learning. Additionally, it highlights a selection of currently available open-source libraries for implementing these fairness techniques.

In contrast, our review provides a targeted analysis of the techniques used specifically for bias mitigation in machine learning, categorizing them into in-processing and post-processing methods. We excluded pre-processing approaches since this is the most common and mentioned in previous work. Additionally, we examine the metrics used to assess bias, the databases tested, and how studies are managing the trade-off between fairness and accuracy. By focusing on these aspects, our review offers a detailed and practical guide for researchers and practitioners looking to implement fair machine learning systems.

Our systematic literature review contributes to the growing body of knowledge on fairness in machine learning by providing a detailed examination of bias mitigation techniques and their effectiveness. This work aims to bridge the gap between high-level ethical discussions and the technical implementation of fair machine-learning practices.

The remaining of the work is organized as follows. Section 2 presents a brief description of fairness in ML. Section 3 presents the systematic literature review carried out, the search protocol, and the results obtained. Finally, Sect. 4 presents the conclusions and future trends.

2 Bias in machine learning

2.1 Notions of *fairness*

As highlighted by [5], at first glance, discussions about discrimination in the context of machine learning may seem strange if we consider that conceptually the goal of ML applications is to discover how to treat different people differently - i.e. discriminate against them. However, what the field of *fairness* treats as discrimination is not a different treatment in itself, but rather a treatment that systematically imposes a disadvantage on one social group in relation to others. From this perspective, an important concept that *fairness* encompass in ML is that of a *protected attribute*, which is an attribute that divides a population into groups that have parity in terms of benefits received, such as race, gender or religion [12]. They are, therefore, sensitive characteristics of individuals that are ensured to avoid discrimination and guarantee equity.

When these protected attributes are used as features in a model, there is a risk that the model may inadvertently learn and perpetuate societal biases. Even if these variables are not directly included, other correlated variables (known as *proxy variables*) can still lead to biased outcomes. The impact of unfair models can reinforce existing social inequalities and lead to adverse outcomes for marginalized groups. For instance, if a model used for job recruitment is biased against certain racial or gender groups, it can exacerbate employment disparities and perpetuate systemic discrimination [13]. Predictive policing algorithms used to forecast crime hotspots have been criticized for disproportionately targeting minority communities. These algorithms often rely on historical crime data, which can be biased due to over-policing in certain neighborhoods, leading to a feedback loop that reinforces existing biases [14].

In healthcare, there are many examples of bias, such as pulse oximeters, which measure blood oxygen levels, have been shown to be less accurate in patients with darker skin tones [15]. This can lead to underdiagnosis of conditions like hypoxemia in Black patients. Algorithms for detecting skin cancer, have been shown to perform less accurately on images of darker skin [16]. This is often due to training data that predominantly features lighter skin tones. The use of race-based adjustments in estimating kidney function can lead to underestimation of kidney disease severity in Black patients, potentially delaying necessary treatments [17]. These examples illustrate how bias in healthcare can have serious consequences for patient outcomes.

To address the problem of bias in ML, one of the first issues that deserves attention is the need to find a definition for *fairness*. Since algorithms "speak mathematics", the *fairness* research community has attempted to mathematically (or statistically) define aspects of society's fundamentally vague notions of fairness and justice in order to incorporate their ideals into the machine learning [18].

In this sense, there are many attempts to create metrics that allow evaluating and quantifying *fairness* in ML models, among which we list the most frequently found in the works consulted in the literature review. These metrics, as well as the ML model accuracy metrics, are based on the confusion matrix relationships:

Equalized Odds (EOdd) - protected and unprotected groups must have equal rates of true positives and false positives [19].

Equal Opportunity (EO) - protected and non-protected groups must have equal rates of true positives [20].

Demographic parity (DP) - the probability of a positive result must be the same, regardless of whether the person belongs to the protected group or not [21].

Equal Treatment (TE) - achieved when the proportion of false negatives and false positives is the same for both categories of protected groups [22].

Average Odds Difference (AOD) - the average difference between the false positive and true positive rates between disadvantaged and privileged groups [20]. A value 0 means that both groups have the same benefit, while values lower or higher than 0 mean a greater benefit for the privileged or non-privileged group, respectively [20].

Statistical Parity (SP) - states that people from the protected and non-protected groups should have the same probability of being assigned a positive result [19].

2.2 Approaches for mitigating bias

Once a basis has been established to understand the concepts and evaluation possibilities regarding *fairness*, it is necessary to present the approaches that allow us to operationalize this concept in real-world systems. The three types of approaches are presented in [23] as follows:

- **Preprocessing:** proposes changing the sampling distributions of protected variables or performing transformations on the data with the aim of removing discrimination from the training data.
- **In-processing:** proposes to act by incorporating one or more equity metrics into the model's optimization functions, in an attempt to converge on a parameterization that maximizes performance and equity.
- **Post-processing:** proposes to apply transformations to the model output to improve the fairness of the prediction.

Each approach has different advantages and disadvantages [5]. For example, for the use of *toolkits* and ML libraries in which there is no access to the optimization function, pre- or post-processing approaches would be more appropriate. However, sometimes it is not possible to make changes to the training base or model results, leaving the option of working on the notion of *fairness* in the model optimization function, that is, in-processing.

3 Systematic literature review

According to [24], Systematic Literature Review (SLR) refers to a specific research methodology, developed with the aim of collecting and evaluating available evidence relating to a specific topic. In this sense, the works retrieved in this analysis contemplate a broad view of bias reduction in machine learning. [25] points out that the SLR development process has three phases: Planning, Conducting, and Publication of Results.

The planning phase encompasses the stages of defining research objectives and protocol. The conduction phase includes the identification and selection of studies along with the analysis of the data obtained. Finally, the publication phase comprises the description, dissemination, and evaluation of results.

3.1 Search protocol

Following this proposal, in the planning phase, it was defined that the main objective of this SLR would be to identify which approaches are currently being used to attack the problem of bias in machine learning and what results are being obtained in relation to *trade-off* bias mitigation vs. accuracy. In this sense, the following questions were defined to guide the reading of the collected material, in order to achieve the defined objective:

1. What techniques are being used to mitigate bias in machine learning?
2. What bias assessment metrics are used?
3. What databases are being tested?
4. What techniques are focused on in-processing and post-processing?
5. How are studies dealing with the *trade-off* bias mitigation vs. accuracy?

Still at this stage, it was decided that the search would be carried out on the digital libraries: Institute of Electrical and Electronics Engineers (IEEE),[1] Springer[2] and Association for Computing Machinery (ACM),[3] since these cover the majority of production related to research in Computing Science. Publications in these databases are more likely to propose new approaches and methodologies for mitigating algorithmic bias, rather than merely applying machine learning with fairness. The search was limited to the five-year range before 2023, to capture the most up-to-date information and techniques on bias reduction in machine learning. Thus, the search *string* used was:

[Abstract: machine learning] AND [Abstract: bias] AND [[Abstract: mitigate] OR [Abstract: fairness]]

3.2 Primary studies

The searches carried out using the *string* defined in the planning phase returned the following initial results:

- ACM: 187 publications.
- Springer: 775 publications.
- IEEE: 13 publications.

Table 1 List of inclusion and exclusion criteria for the SLR

Inclusion criteria	Exclusion criteria
English Publication	SLR or Survey Publications
Available to read in full and accessible in Brazil via CAFe-CAPES	Summary publications (5 pages or less)
Includes the application of bias reduction techniques/tools in ML	Approach focused on pre-processing
	Publications not related to the topic

According to [25], once the studies have been identified, we need to apply selection and exclusion criteria to identify those that will be incorporated into the SLR. Table 1 brings together the inclusion and exclusion criteria adopted in this review.

The first selection stage was carried out by reading the abstracts of the works that were returned in the initial search, which totaled 975 publications. At this stage, SLR or *Survey* publications, summary publications (with 5 pages or less), publications that are not directly related to the topic of this research, and those that did not include the results of the application of a survey technique were excluded. Then, after reading the remaining works, those whose bias reduction proposal used a pre-processing approaches were discarded, as these approaches are the most common and focus only on data processing, such as reweighting, data transformation, and resampling, that are well-established in the machine learning field.

Reweighting involves adjusting the importance of different samples to correct for imbalances or biases in the training data. Data transformation modifies features to reduce bias, such as through normalization or standardization. Resampling techniques, including oversampling minority classes or undersampling majority classes, aim to balance the dataset.

While these approaches are foundational and frequently used to mitigate bias, they do not introduce novel methodologies. Our review aimed to prioritize more innovative techniques that extend beyond traditional pre-processing and delve into advanced methods that address fairness in machine learning from new angles. By focusing on less conventional techniques and exploring both in-processing and post-processing methods, we aim to provide a more comprehensive and forward-looking perspective on fairness mitigation in machine learning.

Figure 1 shows the number of articles remaining at each selection/exclusion stage, leaving, at the end of the process, 30 articles, which make up the analysis developed in this SLR.

[1] https://www.ieee.org/.

[2] https://www.springer.com/br.

[3] https://www.acm.org/.

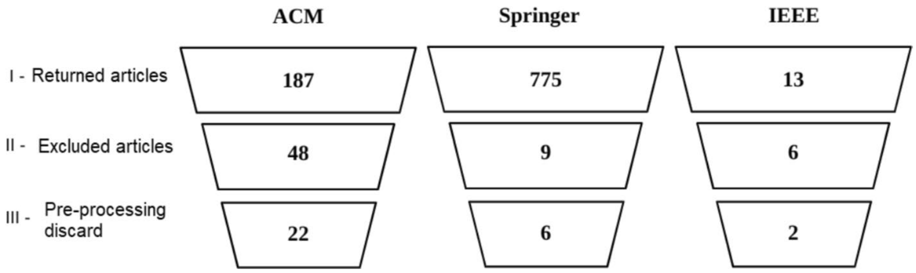

Fig. 1 Articles resulting from the search, selection, and exclusion process for the SLR

Table 2 Protected attributes used in the evaluated works

Attribute	Reference(s)	Frequency
Age	[20, 26–36]	0.40
Race	[19–21, 26–29, 33, 35–43]	0.67
Sex	[19–21, 26–29, 31–35, 35–40, 42–49]	0.83
Functional capacity	[30]	0.03
Education	[35, 42, 50]	0.10
Provenance	[28, 35, 36, 51, 52]	0.17
Occupation	[34]	0.03

3.2.1 Protected attributes

The works analyzed in this review vary in relation to the focus of protected attributes, as shown in Table 2. It is possible to note that gender, race, and age predominate over the others, appearing in 83%, 67%, and 40% of the works, respectively. In general, large datasets often include demographic information related to gender, race, and age, which makes these attributes more accessible for analysis. This availability facilitates the study of biases associated with these factors in machine learning models. Moreover, some countries have laws and regulations addressing discrimination based on gender, race, and age. Fairness in machine learning research often focuses on these attributes to align with legal standards and ethical considerations. Researchers may use these common attributes as a starting point to develop methods and frameworks for fairness, which can then be extended to other attributes or contexts.

While certain sensitive attributes are more commonly used in fairness-aware machine learning models, it is important to consider a broad range of attributes and intersectional identities to ensure that fairness considerations are comprehensive and inclusive. By exploring less explored attributes, researchers and practitioners can develop more robust and equitable machine learning models that address the diverse needs and experiences of individuals and communities.

3.2.2 Metrics

As previously presented, a *fairness* metric represents a quantification of unwanted bias in the data or training models [12]. We can consider that, by evaluating the metrics used in each selected work, we can get an idea about the perspective of equity and justice adopted by those who write the article. We proceed to analyze those that appear in at least two different works, indicating an acceptability of the research community on the topic.

The intention of using ***Equal Oppotunity (EO)*** [19–21, 30, 36, 40, 43, 45–47, 51], which focuses on the true positive rate, is to ensure that the model offers the same chances of positive outcomes (e.g., getting hired, receiving a loan) to individuals as they should have, regardless of belonging to a protected group. In other words, the objective is to eliminate any disparities in the ability to correctly identify positive outcomes between different groups.

Average Odds Difference (AOD) [20, 30, 45] metric considers both true positive rates and false positive rates in different groups. Its use is intended to assess whether there are disparities in the rates of correct identification of positives and false labeling of negatives in different groups. Therefore, the objective is to reduce disparities in error rates between different groups.

Treatment Equality (TE) [27, 30, 35, 41, 44] measures accuracy predictions within each group. With it, the aim is to evaluate whether the algorithm's predictions are equally accurate in different groups, ensuring that its performance is similar for different groups, regardless of belonging to a protected group.

Demographic Parity (DP) [19, 21, 33, 40, 43, 49] focuses on the distribution of positive results by different groups. Therefore, it intends to achieve a proportional representation of different groups in positive results, eliminating disparities in the general distribution of results without favoring or disfavoring any specific group.

Statistical Parity (SP) [19, 30, 31, 36, 39, 50, 51] is a similar metric to DP. It examines whether predictions are consistent with the overall representation of different groups in the population. It intends to ensure that the results of the algorithm are aligned with the demographic composition of the population.

Equalized Odds (EOdd) [19, 26, 28, 33, 43, 46, 50] is a metric that combines EO and AOD and aims to simultaneously achieve equality in the rates of true positives and false positives in different groups. It aims to eliminate disparities in prediction errors and ensure that all groups are treated fairly in terms of positive and negative outcomes.

In addition to those commonly used, some works present new metrics in an attempt to measure *fairness* based on the relationships between predictions and real values of the confusion matrix.

In [37] a concept of discrimination is used which is defined as Demographic Parity, except for explanatory bias[4] $Discrim. = DP - explanatorybias$. [29]'s work only uses the difference in true and false positive rates between groups as a measure. [48] presents a generic framework to measure *fairness* in retrieved results in relation to a protected attribute consisting of three components: the first measures the neutrality of a document's content in relation to the protected attribute; the second is a new metric that provides a normalized score, which characterizes the balance of the classified list content in relation to the protected attribute; the third component aims to separate the effect of document collection on the fairness of retrieval results from the effect of the Information Retrieval model. In [42] *fairness* is measured as the exposure quotient between the protected group and the non-protected group, where values < 1 mean more visibility for the non-protected group, while values > 1 mean more visibility for the protected group. [52] measure *fairness* using *Weighted Proportional Fairness* $PropFair = \sum_{i=1}^{l} w_i log(1 + x_T^i)$, which aims to ensure that recommendations are distributed fairly and equitably among different user segments, where l is the total number of user groups, w_i is the weight assigned to group i and x_T^i is the proportional allocation of group i until time T. The metric used by [34] is reputation disparity, which measures the difference between the means of the reputation distributions of different demographic groups. The disparity is 0 when the average of the reputations is equal and takes on positive or negative values when one group has a higher reputation than the other. In [38] only the accuracy of the classification model is measured to demonstrate an increase with the use of the framework when compared to the original models.

The Weighted Proportional Fairness metric in the context of *FairRec* is a measure used to evaluate the fairness of recommendation allocations between different groups of users based on pre-defined target allocation proportions and weights assigned to each group. The metric aims to ensure that recommendations are distributed fairly and equitably among different user segments, considering the importance of the desired allocation proportions of each group.

3.2.3 Techniques to mitigate bias in machine learning

The techniques used to reduce bias in ML are grouped according to their approach, which may be pre-processing, in-processing, or post-processing. This means that *fairness* conditions can be sought during database manipulation (pre-processing), in the construction of training models (in-processing), or through adjustments of model outputs (post-processing). As already mentioned, in this review, we focused on evaluating in-processing and post-processing approaches.

In-processing techniques

In general terms, **Regularization** is a technique used to prevent the model from overfitting the training data. Its process involves adding restrictions or penalties to the learning algorithm. In the context of *fairness*, restrictions are introduced that seek to optimize for the objectives of EO [45], AOD [20], EOD [20], DP [21, 42]. The model presented in [31] allows users to specify equity constraints, supporting all major constraints, including multiple simultaneous constraints, as well as being model-agnostic (*model-agnostic*). In [43] regularization is used in a decision tree and constraints target DP. In [37] the authors propose an optimization in relation to Demographic Parity, except for explanatory bias, using causal effect estimators *Inverse Probability Weighting* (IPW) and *Doubly Robust* (DR). In [19] the model presented seeks to optimize the hyperparameters of an opaque function while simultaneously satisfying arbitrary *fairness* constraints (defined by the user). The idea of creating adaptive models is also present in [49], whose technique for achieving *fairness* in regression models allows the system developer to flexibly choose the relative importance and thresholds he/she intends to attribute to fairness and accuracy. In [36] a framework is presented for reducing bias in regression models (linear and generalized linear) that applies a penalty *ridge* to reduce the proportion of variance relative to sensitive attributes in relation to the total variance. In [29] regularizers are used for the loss term that is based on an approximation of the distance of the data points to the decision boundary during training. The resulting prediction model seeks to reduce the average distance to the decision boundary between two groups for individuals subject to a negative outcome in each group.

Similarly, an adaptation of the sample weights can be used as a way to impose *fairness* restrictions during training. By assigning different weights to samples based on their attributes or membership in protected groups, training is oriented to give more importance to underrepresented or disadvantaged groups. The model proposed by [27] allows you to adjust the parameters relating to the false positive,

[4] According to [53], for example, attributes such as "number of working hours" and "education level" can, to some extent, explain the wage differences between men and women in a given data set.

false negative, true positive, and true negative rates for which the weight adjustments will optimize. [50] develop a technique based on the AdaBoost framework. But, instead of increasing the weight of the wrongly predicted instances, the proposed technique increases the weight of the biased predicted instances by adjusting the weights correctly. Predicted biased instances are identified using the k-NN based situation testing technique. [32] propose the use of an algorithm *Cooperative Contextual Bandits* that is composed of two *gradient contextual bandits* that uses *fairness* as a reward and tries to maximize it.

Eventually, a bias remover can be added to the training process, which consists of an additional step to the learning algorithm that explicitly reduces the correlation between the protected attribute and the prediction. [41] added a classifier specializing in *African-American English* (AAE) in the offensive language classification task. Thus, for instances predicted to be toxic in the general model and the dialect estimation model (such as AAE), the ensemble will pass it through the specialized classifier.

Another possibility used to achieve better *fairness* indices is the use of **Adversarial** learning, in which an adversarial component (a discriminator) is introduced to identify and counter bias related to the protected attribute. The discriminator takes the predictions from the primary model as input and tries to predict the protected attribute. Its objective is to minimize the prediction accuracy of the protected attribute based on the model results. The adversarial technique proposed by [46] is *model-agnostic*, suitable for any ML model (as long as it is trained using a gradient) and can be used to achieve EO, DP, or EOdd. In [33] the objective of the algorithm is to predict the output Y using a *Gradient Boosting Decision Trees* model while minimizing the ability of an adversarial neural network to predict the sensitive attribute. [51] points out that when reducing adversarial bias if coverage is not obtained for all protected groups, the resulting classifier may still present bias. To mitigate this problem, they propose the addition of a covariance constraint. [48] use an adversarial classifier in an adapted BERT ranker to obtain information retrieval results with less bias. In the study presented by [39] the adversarial technique presents better results for PD when compared to the data balancing method (pre-processing approach).

Post-processing techniques

One of the post-processing techniques found refers to adjusting the decision threshold (***Thresholding***) to achieve a desired level of fairness, balancing the error rates or positive prediction rates between the different groups. [34] employ a technique to adjust the scores of a reputation-based ranking system. The adjusted reputation scores apply the reputations obtained with the same distribution for each group of users under the sensitive attributes (the work proposes the application to reduce bias in multiple attributes).

With the aim of equalizing probabilities (EOdd), or generating equality of opportunities (EO), it is possible to apply techniques to modify model predictions to guarantee equal rates of false positives or false negatives in different groups. [26] propose an algorithm that, given a forest of decision trees for a binary classification task, modifies the prediction of a carefully chosen set of leaves in order to reduce the forest's degree of discrimination (which depends on the number of privileged/non-privileged instances with a favorable forecast). The proposal of [40] is to perform a re-ranking to guarantee a proportional distribution (representative in relation to protected attributes) in an ordered list. To do this, the algorithm can exchange the candidate's position with the previous indexes in the list until reaching the desired *fairness* objectives. [35] developed a *pipeline* that comprises the stages of detection, recognition, and reduction of bias in news titles. The bias reduction module has two stages: *Bias Masking* and *Fairness Infill*. In the first stage (*masking*), the positions of the skewed words are hidden ("masked"). In the second stage (*fairness infill*), the words that were hidden are replaced by others without bias, according to the context of the sentence.

Calibration consists of adjusting the predicted probabilities to match the expected probabilities of the outcome variable in each group. [38] extend the multi-calibration approach, adapting it to the context of binary classification.

[28] shows that a combination of pre-processing (*Disparate Impact Reduction*), in-processing (*Adversarial Debiasing*), and post-processing (*Equalized Odds Post-processing*) approaches is more efficient in reducing bias than when techniques are applied individually.

3.2.4 Databases

To address the issue of fairness and bias reduction in ML models, it is imperative to carefully consider the datasets used in experiments. Data set selection and composition play a key role in determining the effectiveness of bias reduction techniques. Studies can use widely known bases such as, for example, Adult [20, 21, 26, 27, 29, 31–33, 36–38, 43, 46], COMPAS [19, 21, 26, 27, 31–33, 37, 43, 45], German Credit [19, 29, 32, 43, 45] and Bank [27, 31, 33, 36]. Or use specific bases for the analysis intended in the study, such as [51] that assess whether NBA salaries[5] have a bias in relation to the player's origin. Eventually, they may choose not to use real data collected in studies, but to produce experimental data[6] [28, 30, 40, 47–49]. Table 3 brings together

[5] The *National Basketball Association* is the main professional basketball league in North America.

[6] In Table 3, the bases noted with (*) have data that were artificially created for use in the proposed study.

Table 3 Databases used in work

Dataset	Instances	Content	Ref
Adult	48,842	Predictions if income exceeds $50K/year	[19–21, 26, 27, 29, 31–33, 36–38, 43, 46]
MAMI	11,000	Memes about detecting misogyny	[44]
COMPAS	~10,000	Criminal record, prison and detention time	[19, 21, 26, 27, 31–33, 37, 43, 45]
German Credit	1000	Classification as high or low credit risk	[19, 29, 32, 43, 45]
Drug	2051	Substance use risk classification	[45]
LSAC	~27,000	Law Student Information	[21, 31]
NBA	~450	NBA Athlete Statistics from the 2016–2017 Season	[51]
Pokec	~1,600,000	Profile (gender, age, hobbies, interests, education, etc.) of Poket	network users [51]
Bank	45,211	Classification if the customer will contract a banking product (term deposit)	[27, 31, 33, 36]
C & Crime	1994	Combines socioeconomic, law enforcement, and crime data	[37, 50]
MEPS	11,070	Prediction of whether a patient has high utilization (≥10 visits) for medical care	[29]
Recruitment	215	Professional placement of students	[47]
CelebA	202,599	40 image-annotated binary attributes	[38, 38]
Impressions	10,000	Clips of people talking to the camera, labeled with personality traits	[39]
DWMW17	24,802	Tweets noted as hate speech, offensive, or neither	[41]
FDCL18	79,996	Tweets noted as abusive, hate speech, normal or spam	[41]
Golbeck	35,000	Tweets noted as harassing or non-harassing	[41]
WH16	16,849	Tweets containing at least one phrase or word considered hateful by the authors	[41]
W3C experts	9600	48 topics and 200 candidates per topic, labeled as expert or non-expert	[42]
Engineering	–	Academic record of 1st year students at a school at a Chilean university	[42]
Law Students	27,478	Academic record of 1-year students from different Law faculties	[42]
MovieLens	25,000,000	Ratings on 62,000 movies made by 162,000 users	[34, 52]
Health	147,743	Prediction of whether an individual will spend a day in the hospital in the next year	[32]
Credit Default	30,000	Data about users and credit card (non)payment history	[33, 50]
BookCrossing	745,161	Age and location attributes of 53,408 users who rated 263,956 items	[34]
MBIC	1700	Phrases that potentially contain word choice bias	[35]
News	10,000	Crime News Headlines from the Google News API	[49]
IBM HR (*)	1470	Data created by IBM to model factors that lead to worker burnout	[47]
Mobility (*)	–	Data from accelerometer worn on the wrist in staged experiment	[30]
Queries (*)	245	Subset of other databases containing equity-sensitive queries	[48]
LinkedIn (*)	100,000	Mock data for LinkedIn Talent Search ranking	[40]
Applicants (*)	150	Job candidates (with photo) noted regarding profile suitability	[28]

information about the databases used in the works evaluated in this review, including the size of the databases (number of instances), the type of information contained in each one, as well as the articles that used them in the experiments.

3.2.5 *Trade-off* bias mitigation x accuracy

As [20] points out, most bias reduction algorithms harm the performance of the prediction model, in the process of making it fair - which is typically referred to as *trade-off* between *fairness* (or bias reduction) and accuracy. Especially when we consider in-processing approaches, in which objective functions are changed or restrictions are introduced, which can have a negative impact on the accuracy of the model. [50] observe that the method proposed in the work achieves better *fairness* metrics than other tested methods, suggesting its strength in learning fair models. But they also assess that this method achieves a higher prediction error, indicating a *trade-off* between fairness and accuracy, and point out that how to reduce this *trade-off* is an open question. [39] emphasize that, due to *trade-off*, the accuracy of the estimates may deteriorate to improve the fairness of the model. According to the results presented, the adversarial learning approach slightly decreases the model's accuracy, while the data balancing approach damages accuracy less. In an attempt to manage this *trade-off*, it is possible to insert specific parameters into the model for this control. [31] highlight in the model's objective function a hyperparameter λ that regulates the *trade-off* between accuracy and *fairness* metrics. [43] explicitly inserted the fairness and precision requirements as numerical constraints, to make the *trade-off* fairness-accuracy not only predictable but also easy to explain.

Achieving *fairness* objectives without compromising model accuracy is a permanent challenge in the field of AI and Machine Learning. Finding the right balance is essential to ensure that these systems contribute positively to society, minimizing bias and promoting equality.

4 Discussion

Sensitive attributes are often underreported or misrepresented in datasets due to privacy concerns, legal restrictions, or societal stigma. This leads to incomplete data that can skew model outcomes. Besides, individuals often belong to multiple protected groups, and the biases they face may be compounded (e.g., bias against older women of color). Most models consider only one attribute at a time, and may overlook these intersectional biases. Social norms and definitions of fairness can change over time, meaning that what is considered fair today may not be seen as such in the future. Models trained on historical data might perpetuate outdated biases.

Considering the in-processing techniques, ensuring fair representations may be challenging if the underlying data is biased or if the representation is too abstract to capture important nuances. Moreover, it may introduce additional complexity and computational cost. Adversarial debiasing can be complex to implement and tune, requiring balancing between fairness and accuracy through adversarial networks. It may also lead to instability in training and potential trade-offs between fairness and model performance. One of the key challenges with regularization for fairness is balancing the trade-off between fairness and accuracy. Adding fairness constraints typically reduces the model's ability to optimize purely for accuracy. Besides, there is a complexity of incorporating multiple fairness constraints simultaneously, especially when different fairness definitions conflict.

Regarding post-processing techniques, incorporating fairness constraints into the optimization process can lead to a trade-off between fairness and accuracy, potentially reducing the overall performance of the model. It may also add complexity to the optimization problem. Re-calibrating scores to achieve fairness might not address the root causes of bias and can sometimes result in less interpretable models. It also requires careful calibration to avoid over-adjusting predictions.

Balancing fairness and accuracy often involves compromises, where improvements in fairness may lead to reductions in model accuracy. This trade-off can be challenging to manage and may not always align with the specific goals of a given application. Metrics used to evaluate the trade-off between fairness and accuracy might not capture all aspects of model performance or fairness. They can sometimes oversimplify complex trade-offs or fail to reflect the real-world impact of decisions. It's often impossible to fully satisfy all fairness metrics simultaneously. Instead, the choice of which metrics to prioritize depends on the context, the specific application, and the ethical and legal requirements.

5 Conclusions and future trends

When it comes to promoting fairness in machine learning, the most common approaches are categorized into pre-processing, in-processing, and post-processing. Each of these categories has specific techniques that are frequently used. Pre-processing has been widely explored and includes data rebalancing to adjust the distribution of data to ensure that all classes or groups are equally represented. Sensitive attribute removal, resampling (oversampling/undersampling), fair representation learning. In this review, we retrieved works that explored in-processing approaches that include adding regularization terms to the training objective to penalize inequality between groups, and adversarial debiasing. And post-processing approaches wich adjust model predictions to ensure that the true positive rate and false positive rate are equal across groups. Changes the model's decision thresholds to ensure fairness in predictions. Modifies the model output probabilities to meet the desired fairness metrics.

Besides presenting the in-processing and post-processing approaches, we also present the most used fairness metrics, datasets, and the trade-off between fairness and accuracy. Some of the applications that are less explored in the context of fairness and machine learning include:

- Natural Language Processing (NLP): While there has been significant research in NLP, fairness considerations in areas such as sentiment analysis, language translation, and text generation are still relatively underexplored.
- Computer Vision: Although computer vision has seen advancements in areas like object detection and image classification, there is less exploration of fairness issues in tasks such as facial recognition, object tracking, and scene understanding.
- Reinforcement Learning (RL): RL algorithms are commonly used in autonomous systems and gaming applications, but fairness concerns in reward functions, policy

learning, and decision-making processes are not extensively studied.

- Education: Machine learning applications in education, such as personalized learning and student performance prediction, present fairness challenges related to grading, assessment, and resource allocation.
- Environmental Science: Using machine learning to analyze environmental data for climate modeling, biodiversity conservation, and natural disaster prediction raises fairness concerns regarding access to environmental resources and decision-making processes.

Metrics to measure fairness in machine learning are essential for evaluating the performance and impact of algorithms on different demographic groups. Several metrics have been proposed to quantify various aspects of fairness, each addressing different dimensions of bias and discrimination. Some common fairness metrics include EO, AOD, TE, DP, SP, and EOdd. Some new metrics have been proposed too. Many fairness metrics are incompatible with each other, meaning that optimizing one metric may come at the expense of another. For example, optimizing for equalized odds may result in a decrease in overall accuracy or vice versa. This creates a trade-off between fairness and performance, making it difficult to choose the most appropriate metric for a given application. Moreover, the choice of fairness metric depends on the specific context and objectives of the ML application. What may be considered fair in one context may not be appropriate in another. For example, fairness considerations in healthcare may differ from those in criminal justice or hiring practices. Thus, selecting the best fairness metric requires a deep understanding of the domain and the potential implications of different metrics.

Other future trends in the area include incorporating causal inference methods into fairness research to help identify and mitigate systemic biases in machine learning systems, moving beyond correlation-based approaches. These methods will help researchers identify and understand the causal relationships between input features, model decisions, and outcomes, allowing for more effective bias mitigation strategies. Emphasizing user-centered design principles can lead to the development of machine learning systems that prioritize fairness, transparency, and accountability, empowering users to understand and challenge algorithmic decisions. Moreover, increased attention to ethical AI governance frameworks and regulations will shape the future of fairness in machine learning, promoting responsible and accountable deployment of AI systems in various domains.

Acknowledgements We thanks Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grant 21/14725-3 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Data availability This paper is a review, no dataset was used or generated.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017). https://doi.org/10.1126/science.aal4230
2. Forum, W.E.: How to prevent discriminatory outcomes in machine learning (2018). https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf
3. INTELLIGENCE, A.F.T.A.O.A.: Code Of Professional Ethics And Conduct (2019). https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php
4. Calmon, F.P., Wei, D., Ramamurthy, K.N., Varshney, K.R.: Optimized data pre-processing for discrimination prevention (2017)
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning: limitations and opportunities. fairmlbook.org, ??? (2019). http://www.fairmlbook.org
6. Chen, R.J., Wang, J.J., Williamson, D.F., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng. **7**(6), 719–742 (2023)
7. Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., Ghassemi, M.: Ethical machine learning in healthcare. Ann. Rev. Biomed. Data Sci. **4**, 123–144 (2021)
8. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nat. Commun. **13**(1), 4581 (2022)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 1–35 (2021)
10. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Comput. Surv. (CSUR) **55**(3), 1–44 (2022)
11. Caton, S., Haas, C.: Fairness in machine learning: A survey. ACM Comput. Surv. **56**(7), 1–38 (2024)
12. Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias (2018)
13. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women. In: Ethics of Data and Analytics, pp. 296–299. Auerbach Publications, ??? (2022)
14. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016)
15. Sjoding, M.W., Dickson, R.P., Iwashyna, T.J., Gay, S.E., Valley, T.S.: Racial bias in pulse oximetry measurement. N. Engl. J. Med. **383**(25), 2477–2478 (2020)
16. Adamson, A.S., Smith, A.: Machine learning and health care disparities in dermatology. JAMA Dermatol. **154**(11), 1247–1248 (2018)
17. Diao, J.A., Wu, G.J., Taylor, H.A., Tucker, J.K., Powe, N.R., Kohane, I.S., Manrai, A.K.: Clinical implications of removing

race from estimates of kidney function. JAMA **325**(2), 184–186 (2021)

18. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68. ACM, Atlanta GA USA (2019). https://doi.org/10.1145/3287560.3287598 . Accessed 14 July 2023
19. Perrone, V., Donini, M., Zafar, M.B., Schmucker, R., Kenthapadi, K., Archambeau, C.: Fair bayesian optimization. In: Proceedings of the 2021 AAAI/ACM Conference on AI, ethics, and society, pp. 854–863. ACM, Virtual Event USA (2021). https://doi.org/10.1145/3461702.3462629 . Accessed 21 November 2022
20. Chakraborty, J., Majumder, S., Yu, Z., Menzies, T.: Fairway: a way to build fair ML software. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 654–665. ACM, Virtual Event USA (2020). https://doi.org/10.1145/3368089.3409697 . Accessed 21 November 2022
21. Zhao, T., Dai, E., Shu, K., Wang, S.: Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In: Proceedings of the Fifteenth ACM International Conference on Web Search And Data Mining, pp. 1433–1442. ACM, Virtual Event AZ USA (2022). https://doi.org/10.1145/3488560.3498493 . Accessed 21 November 2022
22. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art (2017)
23. Caton, S., Haas, C.: Fairness in machine learning: a survey (2020)
24. Biolchini, J., Mian, P.G., Natali, A.C.C., Travassos, G.H.: Systematic Review in Software Engineering. Technical Report ES **679**(05), 45 (2005)
25. Felizardo, K.R., Nakagawa, E.Y., Fabbri, S.C.P.F., Ferrari, F.C.: Revisão Sistemática da Literatura em Engenharia de Software: Teoria e Prática. Elsevier, Rio de Janeiro (2017)
26. Abebe, S.A., Lucchese, C., Orlando, S.: EiFFFeL: enforcing fairness in forests by flipping leaves. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, pp. 429–436. ACM, Virtual Event (2022). https://doi.org/10.1145/3477314.3507319 . Accessed 21 November 2022
27. Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In: Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18, pp. 853–862. ACM Press, Lyon, France (2018). https://doi.org/10.1145/3178876.3186133 . Accessed 21 November 2022
28. G. Harris, C.: Mitigating Cognitive Biases in Machine Learning Algorithms for Decision Making. In: Companion Proceedings of the Web Conference 2020, pp. 775–781. ACM, Taipei Taiwan (2020). https://doi.org/10.1145/3366424.3383562 . Accessed 21 November 2022
29. Sharma, S., Gee, A.H., Paydarfar, D., Ghosh, J.: FaiR-N: Fair and Robust Neural Networks for Structured Data. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, And Society, pp. 946–955. ACM, Virtual Event USA (2021). https://doi.org/10.1145/3461702.3462559 . Accessed 21 November 2022
30. Alam, M.A.U.: AI-Fairness Towards Activity Recognition of Older Adults. In: MobiQuitous 2020 - 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. 108–117. ACM, Darmstadt Germany (2020). https://doi.org/10.1145/3448891.3448943 . Accessed 21 November 2022
31. Zhang, H., Chu, X., Asudeh, A., Navathe, S.B.: OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning. In: Proceedings of the 2021 International Conference on Management Of Data, pp. 2076–2088. ACM, Virtual Event China (2021). https://doi.org/10.1145/3448016.3452787 . Accessed 21 November 2022
32. Hu, Q., Rangwala, H.: Metric-Free Individual Fairness with Cooperative Contextual Bandits. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 182–191. IEEE, Sorrento, Italy (2020). https://doi.org/10.1109/ICDM50108.2020.00027 . https://ieeexplore.ieee.org/document/9338312/ Accessed 21 November 2022
33. Grari, V., Ruf, B., Lamprier, S., Detyniecki, M.: Achieving Fairness with Decision Trees: An Adversarial Approach. Data Sci. Eng. **5**(2), 99–110 (2020). https://doi.org/10.1007/s41019-020-00124-2. (**21 November 2022**)
34. Ramos, G., Boratto, L., Marras, M.: Robust reputation independence in ranking systems for multiple sensitive attributes. Mach. Learn. **111**(10), 3769–3796 (2022). https://doi.org/10.1007/s10994-022-06173-0. (**Accessed 21 November 2022**)
35. Raza, S., Reji, D.J., Ding, C.: Dbias: detecting biases and ensuring fairness in news articles. Int. J. Data Sci. Anal. (2022). https://doi.org/10.1007/s41060-022-00359-4. (**Accessed 21 November 2022**)
36. Scutari, M., Panero, F., Proissl, M.: Achieving fairness with a simple ridge penalty. Stat. Comput. **32**(5), 77 (2022). https://doi.org/10.1007/s11222-022-10143-w. (**Accessed 21 November 2022**)
37. Ogura, H., Takeda, A.: Convex Fairness Constrained Model Using Causal Effect Estimators. In: Companion Proceedings of the Web Conference 2020, pp. 723–732. ACM, Taipei Taiwan (2020). https://doi.org/10.1145/3366424.3383556. Accessed 21 November 2022
38. Kim, M.P., Ghorbani, A., Zou, J.: Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, And Society, pp. 247–254. ACM, Honolulu HI USA (2019). https://doi.org/10.1145/3306618.3314287 . Accessed 21 November 2022
39. Yan, S., Huang, D., Soleymani, M.: Mitigating Biases in Multimodal Personality Assessment. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 361–369. ACM, Virtual Event Netherlands (2020). https://doi.org/10.1145/3382507.3418889 . Accessed 21 November 2022
40. Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2221–2231. ACM, Anchorage AK USA (2019). https://doi.org/10.1145/3292500.3330691 . Accessed 21 November 2022
41. Halevy, M., Harris, C., Bruckman, A., Yang, D., Howard, A.: Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. In: Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–11. ACM, – NY USA (2021). https://doi.org/10.1145/3465416.3483299 . Accessed 21 November 2022
42. Zehlike, M., Castillo, C.: Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In: Proceedings of The Web Conference 2020, pp. 2849–2855. ACM, Taipei Taiwan (2020). https://doi.org/10.1145/3366424.3380048 . Accessed 21 November 2022
43. Wang, J., Li, Y., Wang, C.: Synthesizing Fair Decision Trees via Iterative Constraint Solving. In: Shoham, S., Vizel, Y. (eds.) Computer Aided Verification vol. 13372, pp. 364–385. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-13188-2_18 . Series Title: Lecture Notes in Computer Science. Accessed 21 November 2022
44. Nozza, D., Volpetti, C., Fersini, E.: Unintended Bias in Misogyny Detection. In: IEEE/WIC/ACM International Conference on Web Intelligence, pp. 149–155. ACM, Thessaloniki Greece (2019).https://doi.org/10.1145/3350546.3352512 . Accessed 21 November 2022

45. Wu, Z., He, J.: Fairness-aware Model-agnostic Positive and Unlabeled Learning. In: 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1698–1708. ACM, Seoul Republic of Korea (2022). https://doi.org/10.1145/3531146.3533225. Accessed 21 November 2022
46. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating Unwanted Biases with Adversarial Learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, And Society, pp. 335–340. ACM, New Orleans LA USA (2018). https://doi.org/10.1145/3278721.3278779 . Accessed 21 November 2022
47. Liu, D., Shafi, Z., Fleisher, W., Eliassi-Rad, T., Alfeld, S.: RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, And Society, pp. 745–755. ACM, Virtual Event USA (2021). https://doi.org/10.1145/3461702.3462618 . Accessed 21 November 2022
48. Rekabsaz, N., Kopeinik, S., Schedl, M.: Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 306–316. ACM, Virtual Event Canada (2021). https://doi.org/10.1145/3404835.3462949 . Accessed 21 November 2022
49. Almuzaini, A.A., Singh, V.K.: Balancing Fairness and Accuracy in Sentiment Detection using Multiple Black Box Models. In: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics In Multimedia, pp. 13–19. ACM, Seattle WA USA (2020). https://doi.org/10.1145/3422841.3423536 . Accessed 21 November 2022
50. Bhaskaruni, D., Hu, H., Lan, C.: Improving Prediction Fairness via Model Ensemble. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1810–1814. IEEE, Portland, OR, USA (2019). https://doi.org/10.1109/ICTAI.2019.00273 . https://ieeexplore.ieee.org/document/8995403/ Accessed 21 November 2022
51. Dai, E., Wang, S.: Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 680–688. ACM, Virtual Event Israel (2021). https://doi.org/10.1145/3437963.3441752 . Accessed 21 November 2022
52. Liu, W., Liu, F., Tang, R., Liao, B., Chen, G., Heng, P.A.: Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning. In: Lauw, H.W., Wong, R.C.-W., Ntoulas, A., Lim, E.-P., Ng, S.-K., Pan, S.J. (eds.) Advances in Knowledge Discovery and Data Mining vol. 12084, pp. 155–167. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-47426-3_13 . Series Title: Lecture Notes in Computer Science. Accessed 21 November 2022
53. Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: 2013 IEEE 13th International Conference on Data Mining, pp. 71–80 (2013).https://doi.org/10.1109/ICDM.2013.114

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.