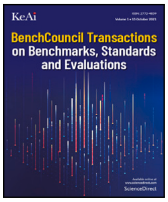
Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/



Review article

Algorithmic fairness in social context

Yunyou Huang [a], Wenjing Liu [a], Wanling Gao [b], Xiangjiang Lu [a], Xiaoshuang Liang [a], Zhengxin Yang [b], Hongxiao Li [b], Li Ma [a], Suqin Tang [a,*]

[a] *Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, No. 15 Yucai Road, Qixing District, Guilin 541004, Guangxi, China*

[b] *Research Center for Advanced Computer Systems, Institute of Computing Technology, Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Haidian District, 100190, Beijing, China*

ARTICLE INFO

Keywords:
Social fairness
Bias
Discrimination
Fairness measure
Fairness datasets
Fairness algorithms

ABSTRACT

Algorithmic fairness research is currently receiving significant attention, aiming to ensure that algorithms do not discriminate between different groups or individuals with similar characteristics. However, with the popularization of algorithms in all aspects of society, algorithms have changed from mere instruments to social infrastructure. For instance, facial recognition algorithms are widely used to provide user verification services and have become an indispensable part of many social infrastructures like transportation, health care, etc. As an instrument, an algorithm needs to pay attention to the fairness of its behavior. However, as a social infrastructure, it needs to pay even more attention to its impact on social fairness. Otherwise, it may exacerbate existing inequities or create new ones. For example, if an algorithm treats all passengers equally and eliminates special seats for pregnant women in the interest of fairness, it will increase the risk of pregnant women taking public transport and indirectly damage their right to fair travel. Therefore, algorithms have the responsibility to ensure social fairness, not just within their operations. It is now time to expand the concept of algorithmic fairness beyond mere behavioral equity, assessing algorithms in a broader societal context, and examining whether they uphold and promote social fairness. This article analyzes the current status and challenges of algorithmic fairness from three key perspectives: fairness definition, fairness dataset, and fairness algorithm. Furthermore, the potential directions and strategies to promote the fairness of the algorithm are proposed.

1. Introduction

Currently, the fairness of algorithms has drawn a lot of attention in many fields, such as recidivism prediction [1], item recommendation [2], and outcome prediction [3] et al. Numerous studies have demonstrated the prevalence of unfairness in decision-making algorithms and algorithm-based systems [4–10]. In response, researchers have been actively working towards eliminating algorithmic unfairness through the development of fairness measures [11–13], the creation of fairness datasets [14–16], and the proposal of fair algorithms [17–19], among other approaches.

Research on algorithmic fairness can be categorized according to two different principles: whether to consider the long-term impact and whether to consider non-technical factors [22,23]. Table 1 shows the explanation of the terminology used in this paper. Based on the first principle, algorithmic fairness can be divided into static fairness and dynamic fairness [17,23–27]. For example, when a loan application algorithm tackles discrimination in selection rates between races, it is classified as static fairness research, but if it also takes into account the long-term effects (such as credit score change) of its decisions on the underlying population, it is categorized as dynamic fairness research [28]. According to the second principle, algorithmic fairness can be classified as technical fairness, social fairness, and sociotechnical fairness [29–36]. For example, when a loan application research optimizes mathematical rate-related fairness measures (such as equalized odds), it is classified as technical fairness research, when it pursues regulation of non-algorithmic factors (for example, making a norm to uphold the developer, user, and executor of algorithms [37]), it is classified as social fairness research, and when it addresses discrimination against different races from both technical and non-technical perspectives, it is classified as sociotechnical fairness research [22].

The central idea behind algorithmic fairness in current literature is to minimize discrimination by algorithms or systems that use algorithms, both against different groups and against individuals who are similar to each other. However, according to the definition of infrastructure — the myriad structures that underpin modern society, the algorithm has become an important social infrastructure [38]. Thus,

* Corresponding author.
E-mail address: sqtang@mailbox.gxnu.edu.cn (S. Tang).

https://doi.org/10.1016/j.tbench.2023.100137
Received 24 July 2023; Received in revised form 23 August 2023; Accepted 24 August 2023
Available online 1 September 2023

Table 1
Terminology and Explanation of Terms.

Term	Definition
Fairness [20]	“Fairness means the absence of any biases based on an individual’s inherent or acquiredcharacteristics that are irrelevant within the specific decision context.”
Static Fairness [21]	Without considering changes in the environment, only the current state is taken into account.Usually, “static fairness provides a one-time fair solution based on optimizing fairness constraints.”
Dynamic Fairness [21]	It is an ongoing process that requires considering environmental changes,learning, and adapting to those changes to maintain fairness in decision-making.
Social Fairness [22]	Society maintains fairness by continuously striving to balance various forces.
Technical Fairness [22]	Efforts are made to utilize fairness metrics and other approaches to measure biasesin algorithms, seeking technological means to mitigate algorithmic discrimination against different subgroups or individuals.
Sociotechnical Fairness [22]	“The outcomes of a system are influenced by the interplay betweentechnical structure and social structure, as well as the interplay between instrumental values and humanistic values.”
Process Fairness	Emphasizing the fairness in the process of decision-making or allocation,without being concerned about the actual outcomes.
Outcome Fairness	Focus on whether the actual outcome is fair.
Group Fairness [4]	A certain group should receive equal treatment as privileged groups or the overall population.
Individual Fairness [4]	“Similar individuals should be treated similarly.”

algorithmic fairness research should not only strive to be fair but also bear the responsibility of creating a fair society, otherwise may lead to seemingly fair algorithms causing societal unfairness or creating new forms of unfairness. For example, face recognition provides audit services for all railways and aviation in China, supports the normal operation of railways and aviation, and becomes an important social infrastructure. For fairness of algorithmic behavior itself, even if the failure rate of facial recognition for individuals with facial impairments is the same as that of normal individuals, due to their heightened psychological sensitivity, they may avoid using facial recognition technology out of fear of recognition failure. This, to some extent, harms the interests of this group. For social fairness algorithmic, individuals with facial impairments should be treated specially (for example, adjust the algorithm recognition threshold) to protect them from non-technical discrimination in public due to facial defects.

In this paper, we first review and analyze existing fairness definitions (problem instantiation), fairness datasets (problem instantiation), and fairness algorithms (solution instantiation) to summarize the progress of algorithm fairness [39,40]. Then extend fairness from the algorithm level to the social level across the entire life cycle of the algorithm. As shown in Fig. 1, for the problem definition, fairness in a social environment demands not only fair behavior from algorithms and systems as societal infrastructure but also their contribution to promoting social fairness. Hence, we emphasize that the assessment of fairness extends beyond the behavior of algorithms or algorithm-based systems and includes fairness within society. For the problem instantiation, in addition to subjects, the instantiation of fairness problems should also involve algorithm developers, users, and executors. Further, in order to reflect the real fairness of society, the instances of the problem must also maintain the consistency of characteristics with the real society. For the solution instantiation, the design of the algorithm not only needs to consider optimizing the fairness metric but also needs to be able to detect the degree of social fairness and then adjust the behavior of the algorithm. Additionally, we highlight the need to restructure the algorithmic fairness benchmark in light of the new algorithmic fairness techniques above to advance algorithmic fairness research.

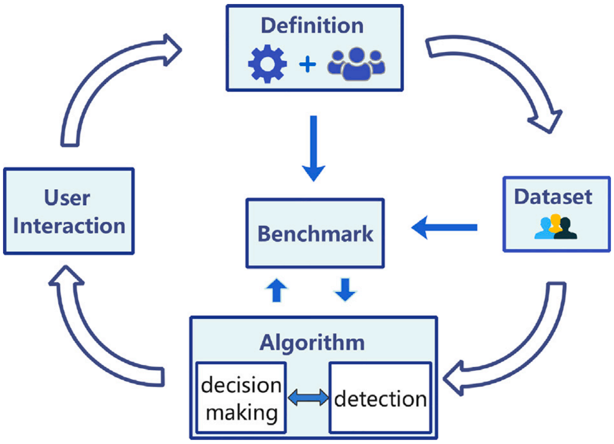


Fig. 1. The extension of algorithm fairness.

The paper is structured as follows. Section 2 reviews the definition of algorithmic fairness. Section 3 reviews data used in recent algorithmic fairness research. Section 4 reviews the fair algorithms. Section 5 extends the algorithmic fairness in the social context. Section 6 draws a concluding remark.

2. Fairness definitions: Problem definition

The concept of fairness has been widely debated in moral and political discussions, but it lacks a consistent definition [41]. With the increasing integration of AI in various domains, ethical and moral concerns have arisen, leading scientists to explore ways of incorporating fairness into algorithmic systems [42]. Currently, fairness in algorithms is defined as the overall performance of an algorithm or algorithm-based system in treating individuals or groups, as assessed by fairness metrics. In the subsequent sections, we discuss the definition and metrics related to algorithmic fairness.

2.1. Fairness definition

The presence of diverse preferences and perspectives across different cultural backgrounds makes it challenging to establish a universal definition that applies to all individuals [43]. Broadly, fairness means that there are no biases towards an individual's inherent or acquired characteristics during the decision-making process [44]. Fairness can be divided into two categories: process fairness and outcome fairness [45], depending on whether the focus is on the fairness of the decision-making process itself or its resulting outcomes [46]. Ensuring fairness throughout the entire decision-making process is particularly challenging due to factors such as the black-box nature of deep learning algorithms. So current research primarily concentrates on outcome fairness. In terms of outcome fairness, it can be further divided into group fairness and individual fairness based on the goals of algorithmic fairness [45,46]. Distributive individual fairness holds that outcomes should be fair at the individual level, while group fairness holds that outcomes should be fair across groups [47]. Although fairness can be classified based on objectives, different researchers have different views on what the outcome of fairness should be, which we call fairness concepts [45]. The most influential concepts include Consistent Fairness and Calibrated Fairness. Consistent Fairness holds that similar individuals or diverse groups with similarities should obtain similar outcomes, while Calibrated Fairness requires that an individual's (or group's) outcome value should be proportional to their merit [45]. The above definitions are all based on the behavior of the algorithm itself without considering the existence of social unfairness. In Section 5.1, we will expand the definition of fairness: social fairness not only requires maintaining fairness in algorithmic behavior but also considers eliminating social unfairness and promoting social equity.

2.2. Fairness metrics

The definition of algorithmic fairness is intertwined with its measurement metrics. The fairness is determined by the values of fairness metrics, and the design of fairness metrics relies on the definition of fairness. In Table 2, we have provided a list of commonly used fairness metrics, categorized into individual fairness and group fairness. Individual fairness lacks a simple and executable definition, making it often difficult to achieve a consensus. On the other hand, group fairness, due to its simplicity and quantifiability, is widely utilized in fairness research. All these metrics can be useful in bias mitigation tasks when dealing with protected attributes, where A represents the protected attribute. The true label is denoted as Y, and the predicted label is denoted as $\hat{Y}$, where 0 represents negative outcomes and 1 represents positive outcomes [4]. Probabilities are represented as P.

However, it is important to note that discussions surrounding algorithmic fairness extend beyond technical metrics. The social objectives of deploying a model, the group of individuals affected by the model's decisions, and the available decision space for decision-makers to interact with the model's predictions must also be considered [48]. Different stakeholders have varying objectives, and the selection of fairness metrics must consider various application scenarios.

In summary, while fairness metrics can serve as useful tools for mitigating task biases, it is crucial to adopt a holistic approach to algorithmic fairness by considering the social context and the needs of all relevant stakeholders.

3. Fairness datasets: Problem instantiation

A dataset, which is an instance of a problem or task, is a fundamental component for the development of data-driven machine learning algorithms as it reflects the essential characteristics of a problem or task. In recent years, many datasets have been utilized for algorithmic fairness research [43,55]. In addition, due to the long-term impact of fairness, several simulators have also been developed to address the limitations of static datasets in fairness research [23]. In the following subsections, we discuss the efforts related to datasets and simulators in the context of algorithmic fairness.

3.1. Dataset

In Table 3, we have compiled a list of 10 fairness datasets and described their protective attributes and other characteristics. However, some of the datasets used in equity research may exhibit biases against protected attributes such as gender, race, and age. These biases can have adverse effects on vulnerable groups. For example, the UCI Adult dataset includes three protected attributes — gender, age, and race. However, an analysis of the dataset shows that high-income men outnumber women in almost all relationship statuses, and there is also some racial bias present [43]. Although this dataset is commonly used for categorical tasks such as predicting income levels, the $50k threshold is set inappropriately, leading to biases against Blacks and women [43].

In practice, addressing these biases may involve expanding the dataset, changing the labels of some data points, or weighting the protected attributes. For example, Retiring Adult, a reconstructed UCI Adult dataset, has been created using real data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) [15]. This dataset encompasses income as a continuous variable, enabling a more realistic prediction of whether an individual earns over $50k annually. Moreover, the dataset includes forecasting tasks for various applications such as income, employment, health, transportation, and housing.

Furthermore, Ilias et al. [16] proposed a benchmark set of legal texts covering multiple regions and languages. They adopted a competency-centered equity approach with the goal of ensuring that each group had sufficient resources to achieve similar performance levels. Ultimately, this approach is centered on the important factors of how individuals are treated in the legal process, making it an equitable approach.

Although researchers have made many improvements to fairness datasets, fairness datasets are typically static and difficult to support the development of dynamic fairness algorithms. Moreover, currently collected fairness datasets do not consider how to maintain social fairness characteristics, making it challenging to support the development of algorithms that are oriented towards social equity.

3.2. Simulator

In the pursuit of achieving dataset fairness, simulators have been proposed as a valuable tool, particularly for long-term and dynamic scenarios. D'Amour et al. [23] have developed an open-source software framework called ml-evenness-gym, an extension of OpenAI's Gym, to examine the long-term impact of the existing fair decision-making system. The framework employs a Markov decision process (MDP) where an agent chooses an action at each step to influence the state of the environment. The environment presents an observation to the agent, which is then used to determine the next action. This iterative process continues until the environment reaches an end state. Similarly, Xueru et al. [56] adopted a partially observed Markov decision process (POMDP) framework to model sequential decisions in different situations. They consider a discrete-time sequential decision process applicable to a particular population, where the effects of decisions made in each time step are reflected in the population characteristics

Table 2
Common fairness metrics.

Type	Name	Mathematical expression	Meaning
Group Fairness	Statistical Parity [44]	$P(\hat{Y}\|A=0)$ $= P(\hat{Y}\|A=1)$	The unprotected groupand the protected group have an equal proportion of favorable outcomes.
	Equalized Odds [13]	$P(\hat{Y}\|A=0,Y=y)$ $= P(\hat{Y}\|A=1,Y=y)$	Individuals from different group should have an equal chance of being correctly classified as positive (true positive) and incorrectly classified as positive(false positive).
	Equal Opportunity [13]	$P(\hat{Y}=1\|A=0,Y=1)= P(\hat{Y}=1\|A=1,Y=1)$	The true positive rateis equal across protected and unprotected groups.
	Treatment Equality [49,50]	$\frac{P(\hat{Y}=1\|A=1,Y=0)}{P(\hat{Y}=0\|A=1,Y=1)} = \frac{P(\hat{Y}=1\|A=0,Y=0)}{P(\hat{Y}=0\|A=0,Y=1)}$	The ratio of false positiverate to false negative rate is the same between different populations.
	Test Equality [51]	$P(Y=1\|A=0,\hat{Y})$ $= P(Y=1\|A=1,\hat{Y})$	The probability ofindividuals in both the protected and unprotected groups belonging to the positive class is equal.
Individual Fairness	Fairness ThroughUnawareness [44,52,53]	\	An algorithm is considered fairas long as it does not explicitly use any protected attribute A in the decision-making process.
	Fairness ThroughAwareness [44,54]	\	For a given task-specificsimilarity measure(inverse distance), any two similar individuals should receivesimilar outcomes.
	Counterfactual Fairness [44,53]	$P(\hat{Y}_{A\longleftarrow a}(U)=y\|X=x,A=a)=$ $P(\hat{Y}_{A\longleftarrow a'}(U)=y\|X=x,A=a)$	"If a decision remains consistenttowards an individual in both the actual world and a counterfactual world where the individual belongs to a different demographic group, then that decision is considered fair towards the individual."

Table 3
Overview of real-world datasets for fairness.

Dataset name		#Instances (cleaned)	Class	Domain	Protected attributes	Collection location
Law School [58]		20, 798	Binary classification	Education	Male, race	USA
UCI adult dataset [59]		45, 222		Finance	Sex, race, age	
Diabetes [60]		45, 715		Healthcare	Gender	
Dutch Census [61]		60, 420		Social	Sex	The Netherlands
Diversity in faces dataset [62]		1, 000, 000	Face recognition	Facial images	–	–
Credit Card Clients [63]		30, 000	Binary classification	Finance	Sex, marriage, education	Taiwan, China
Bank Marketing [64]		45, 211			Age, marital	Portugal
COMPAS Recid [65]		6, 172		Criminology	Race, sex	USA
COMPAS Viol Recid [65]		4, 020				
Retiring Adult: 2018 PUMS [15]	ACSIncome	1, 599, 229		Finance	Sex, race, age	
	ACSPublicCoverage	1, 127, 446		Healthcare		
	ACSMobility	620, 937		Housing		
	ACSEmployment	2, 320, 013		Employment		
	ACSTravelTime	1, 428, 642		Transportation		

in subsequent time steps. This work successfully addresses the limitations of using limited long-term dynamic datasets. In addition, some simulation methods utilize data augmentation techniques to address discrepancies in data sets. For instance, Iosifidis et al. [57] have used oversampling and SMOTE to generate pseudo-instances in minority communities.

Simulators have been widely used in the research of dynamic fairness algorithms to compensate for the limitations of fairness datasets, which cannot adequately support dynamic development. However, current simulators cannot provide personalized simulations for participants, leading to lower accuracy. Additionally, existing simulators do not consider the interaction between the simulated algorithm system and the participants, making it difficult to support research on algorithms focused on social fairness.

4. Fairness algorithms: Solution instantiation

In recent years, a multitude of algorithms have emerged with the aim of reducing bias and discrimination in the behavior of algorithms and systems. These innovative approaches have been introduced to address this pressing issue and ensure fairer outcomes [45,82]. Table 4 summarizes the pre-process, in-process, and post-process mechanisms for algorithmic fairness. Pre-process mechanisms, while applicable to any classification algorithm, may compromise interpretability. In-process mechanisms effectively address accuracy and fairness in the objective function, but are closely tied to the algorithm. Conversely, post-process mechanisms can be used with any classification algorithm but often yield inferior results due to their delayed application.

4.1. Pre-process

Pre-processing mechanisms play a crucial role in preparing data for machine learning algorithms. Their purpose is to minimize or eradicate bias and unfairness within the data. These methods are employed prior to feeding the data into the algorithms, ensuring that the subsequent analysis and modeling are based on a more equitable and unbiased foundation. Typically, these methods encompass techniques that focus on manipulating the distribution of protected variables within the sample or applying specific transformations to the data. The goal is to ensure that the input data remains impartial and unbiased, thereby

Table 4
Methods for fairness.

Paper	Stage	Scheme	Datasets	Evaluation Measure
[66]	Pre-process	Transformation	Adult	Discrimination=0.11, AUC=0.78
[67]	Pre-process	Reweighing	Adult	p%-rule=100%, Accuracy=82%
[68]	Pre-process	Causal Methods	NYCSF	FACE=0.273
[69]	Pre-process	Adversarial learning	Adult	EMD=0.001, Avg.Score=0.239
[70]	Pre-process	Adversarial learning	Adult	Risk Difference=0.0411, Balanced Error Rate (BER)=0.3862
[71]	In-process	Regularization	Adult, Crime and Communities, COMPAS, Default,Law School, Sentencing	–
[72]	In-process	Regularization	COMPAS	Benefits=0.97, Accuracy=68%
[73]	In-process	Adversarial learning	Adult	–
[74]	In-process	Adversarial learning	Adult	FPR=0.0248, FNR=0.4492
[75]	In-process	constraint optimization	Bank Marketing	Accuracy=87%, p%-rule=45%
[76]	In-process	constraint optimization	Use data from 3 real conferences	Paper Score(PS)=1.65, The assigned papers per reviewer(RA)=0.63
[13]	Post-process	Threshold	FICO	Profit=99.3%
[77]	Post-process	Threshold	COMPAS	–
[78]	Post-process	Transformation	Adult	PSE<3.7, Accuracy=73.8%
[79]	Post-process	Transformation	COMPAS	NDE=(0.95,1.05), Accuracy=67.8%
[80]	Post-process	Calibration	–	–
[81]	Post-process	Calibration	Racial Faces in the Wild	Accuracy=90.58%

enabling machine learning algorithms to generate fair and equitable decisions [83].

Du et al. [66] introduced a convex optimization approach to learn data transformations that aim to control group discrimination, limit distortion in individual data samples, and preserve utility. Krasanakis et al. [67] proposed a novel approach called CULEP for mitigating bias in binary classifiers. It uses an iterative reweighting process to recognize sources of bias and diminish their impact without affecting features or labels. The approach encapsulates both fairness- and classifier-related information and allows for a more precise stochastic analysis. Khademi et al. [68] made significant contributions by introducing two novel definitions of group causal relations from a causal perspective. These definitions were developed using causal methods and were designed to effectively quantify group fairness.

Recently, adversarial learning techniques have been utilized by researchers to generate fair samples. Feng et al. [69] proposed a framework for learning a latent representation of attributes through adversarial learning, preprocessing the data, and preserving useful information while preventing useless information as much as possible. Wu et al. [70] introduced a unified framework called FairGAN for generating data that meets various fairness requirements while having good utility.

4.2. In-process

The main concept of the in-process approach is to incorporate fairness considerations into the model optimization process during machine learning training, with the aim of addressing issues of unfairness resulting from dataset bias [83]. This method involves integrating fairness metrics into the model's objective function, allowing for the simultaneous optimization of performance and fairness during training.

Regularization is a common technique used in the in-process stage of fairness mechanisms. Berk et al. [71] introduced a flexible regularizer incorporating individual and group penalty mechanisms into the framework. Heidari et al. [72] addressed the fairness problem in welfare measures by enhancing the penalty for the fair benefit function. Adversarial learning methods are also rapidly developing. Celis et al. [73] employed an adversarial learning paradigm to design fair classifiers by introducing fairness objectives to enhance model performance. Similarly, Zhang et al. [74] utilized adversarial learning to mitigate bias, which is flexible and applicable to various definitions of fairness.

Constrained optimization is also a popular method. Zafar et al. [75] designed a classifier to maximize accuracy while adhering to fairness constraints to ensure that algorithmic decisions do not have unfair effects on certain sensitive attribute groups. Kobren et al. [76] proposed a novel formulation for the paper matching problem. The proposed algorithm, FAIRIR, simultaneously optimizes the global objective, obeys local fairness constraints, and satisfies lower and upper bounds on reviewer loads to ensure more balanced allocation.

4.3. Post-process

The post-processing mechanisms discussed in this passage are applied to machine learning models during the prediction and evaluation stages. These mechanisms aim to adjust the model's output to enhance fairness.

Threshold adjustment and transformation are commonly used methods for improving fairness in machine learning models. Hardt et al. [13] and Corbett et al. [77] have implemented different decision thresholds for various groups to enhance equal opportunities. Chiappa et al. [78] proposed a path-specific approach to address fairness issues. The approach corrects individual decisions by removing unfair information caused by sensitive attributes while preserving the remaining fair information along a specific path. The method is demonstrated using linear models and graphical causal models. To address the problem of fair statistical inference based on results, Nabi et al. [79] formulated the existence of discrimination as the presence of specific path effects (PSE). This refers to a path of effect determined by mediation analysis and can assist in understanding the mechanisms and reasons for discrimination.

Compared to threshold adjustment and transformation, calibration of prediction results is a method that can adjust the bias of predictions to make them closer to the true values. Hebert et al. [80] proposed a multi-calibration approach that considers the predictions of multiple calibration predictors to reduce bias and ensure fairness and accuracy. Salvador et al. [81] proposed a fair calibration method due to the recognition bias of facial recognition technology towards minority groups. This method improves the model's accuracy and generates fair calibration probabilities, thereby reducing the unfair treatment of minority groups.

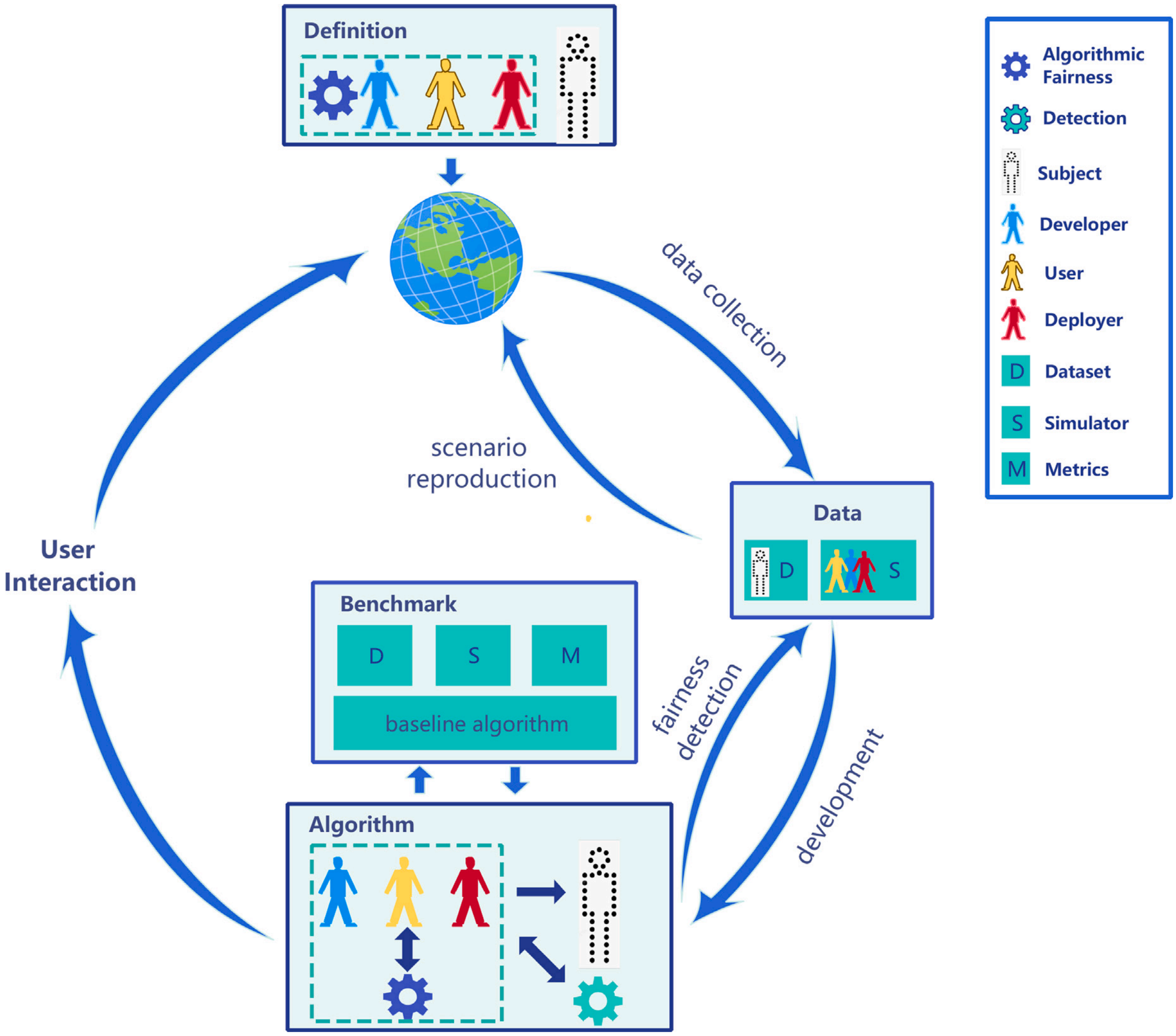


Fig. 2. The lifecycle of fairness in social content. We have expanded three stages of the development of fairness algorithms. Firstly, in the definition stage of fairness algorithms, besides including the definition of algorithmic behavior itself, it is essential to consider relevant social factors during the development process, such as developers, users, and executors. Additionally, the dynamic changes of subjects should also be taken into account. Secondly, in the instantiation stage of fairness problems, besides collecting data related to the fairness definition of the algorithm, it is necessary to model and develop simulators for each individual to simulate the dynamic interaction between the algorithm and the subjects. At this stage, the characteristics of the dataset should align with those of the real world, enabling the dataset to simulate the real world. Lastly, in the development stage of fairness algorithms, we emphasize that fairness algorithms should be capable of detecting and perceiving the level of social fairness and be able to dynamically adjust their behavior accordingly. Furthermore, to promote the development of algorithmic fairness, we need to reconstruct the current fairness benchmarks based on the aforementioned changes.

5. Fairness in social context

As social infrastructure, algorithms are not only responsible for their own behavioral fairness but also for alleviating unfairness and maintaining fairness in society. Fig. 2 presents the fairness lifecycle in the social context and identifies where algorithmic fairness can contribute. In the following subsections, current gaps in algorithmic fairness across different stages of the algorithm lifecycle are identified, and recommendations are provided for ensuring fairness in the social context.

5.1. Fairness definitions: Problem definition

The current approach to algorithmic fairness only considers the behavior of algorithms or algorithm-based systems towards people. However, fairness is a social attribute, and the impact of algorithm or system behavior on social fairness should be considered. Otherwise, a fair algorithm may risk undermining social fairness. To illustrate this, we can take the example of the MRI-PET-based diagnostic model proposed by Janghel et al. [84].

Under the current definition of algorithmic fairness, the algorithm's fairness is defined by the algorithm's prediction accuracy, sensitivity, and specificity in different groups based on gender, age, ethnicity, and disease. However, this definition has a significant flaw: the developer and user of the diagnostic model require all patients to undergo costly MRI and PET examinations, which is unfair to most patients, especially those who are healthy.

Current research has expanded the definition of algorithmic fairness to address this shortcoming by considering social perspectives. For example, Huang et al. [85] proposed a model that can tailor diagnostic strategies to patient-specific conditions. Algorithmic fairness can be defined as the prediction accuracy, sensitivity, and specificity of the system, composed of developers, users, executors, and algorithms, in different groups based on gender, age, ethnicity, income, and disease, ultimately promoting social fairness. While this approach considers humans in the loop, it also focuses only on the fairness of the algorithm-based system behavior itself.

A fair algorithm that does not take into account social inequalities can perpetuate or exacerbate social inequities. For instance, due to

Table 5
Differences between in the social context and current research in Fairness Definitions, Fairness Datasets, Fairness Algorithms.

	In The Social Context	Current Researchs
Fairness Defintions	The behavior exhibited jointly bythe algorithm and its associated social factors is equitable, and it contributes to the enhancement of societal fairness.	The behavior of the algorithm itself is fair.
Fairness Datasets	In addition to the data concerningthe target objects,the dataset also incorporates information involving developers,users, deployers,and their interactions with the system, enabling the data to recreate real-world scenarios.	Collecting data of the algorithm'sor system's target objects.
Fairness Algorithms	With the involvement of developers, users,deployers, and other individuals, this falls under the "human in the loop" mode. Moreover, it allows for the dynamic adjustment and evaluation of the algorithm's fairness.	Focusing solely on the fairnessof the algorithm itself.

the uneven distribution of medical resources, the diagnosis of irreversible Alzheimer's disease presents considerable inequity. In low-income areas, there is even a lack of specialized outpatient clinics for Alzheimer's disease, leading to a significant number of undiagnosed patients and missed early intervention. A fair Alzheimer's disease diagnosis algorithm alone will not alleviate this inequity; we need to consider improving the diagnosis accuracy of low-income groups and reducing the resource requirements of diagnosis strategies in similar situations of high-income groups.

Therefore, the definition of algorithmic fairness in specific tasks needs to focus on promoting social fairness by ensuring the prediction accuracy, sensitivity, and specificity of a system composed of algorithms, developers, users, and executors in different groups, and ultimately slowing down social discrimination in the short and long term.

5.2. Fairness datasets: Problem instantiation

Developing data-driven solutions requires using datasets to instantiate the problem and developing algorithms on the dataset. Researchers in different fields have collected datasets that can represent fairness problems in their respective fields to develop fairness algorithms. Researchers have also proposed various simulators to supplement the current static dataset to restore the dynamic and long-term nature of the fairness problem.

However, the behavior of an algorithm or algorithm-based system is not solely determined by the algorithm itself. Still, it should also include the behavior of developers, users, and executors, collectively called the "human-in-the-loop". Fairness datasets do not collect data on these social roles and cannot fully represent fairness problems in the real world. Collecting data on these social roles to create appropriate case examples for the problem will become a new direction for future research on algorithmic fairness datasets.

Furthermore, algorithms, as infrastructure, should be responsible for promoting social fairness. Unlike traditional fair datasets with loose inclusion criteria during data collection, future fair datasets must truly reflect the degree of social fairness, thereby supporting research and development to promote social equity algorithms. Maintaining fairness in real-world characteristics during data collection will become an urgent problem to be addressed. In order to ensure that datasets remain consistent with the real world, it is necessary to consider stratifying participants during data collection to select more representative individuals. Additionally, a dynamic updating mechanism for the dataset needs to be established to ensure that its characteristics continuously align with the real world. Furthermore, there is a need to develop Fairness Metrics Tool to measure the level of fairness in both the real world and the dataset. These Fairness Metrics Tool will guide the updating process of the dataset.

5.3. Fairness algorithms: Solution instantiation

While fairness algorithms have made progress in addressing algorithmic bias, they have primarily focused on the technical aspects of algorithmic fairness. However, it is important to consider the role of humans in the algorithm-based system, particularly in human-in-the-loop scenarios. The interaction between human behavior and algorithm behavior is complex and requires a more comprehensive approach. This can be achieved by modeling human roles and optimizing fairness alongside the algorithm, which has become a crucial research direction in algorithmic fairness.

In addition to being fair, algorithms should also improve social fairness. Social fairness is not static, and continued protection of vulnerable groups can inadvertently create new injustices. To achieve social fairness, algorithms must detect social fairness and dynamically adjust decision-making behaviors based on the degree of social fairness. This approach will ultimately improve social fairness while maintaining it over time. In this social context, social fairness detection and fairness dynamic game modeling will become crucial extensions of algorithmic fairness research. In order to promote social fairness, future research should focus on enhancing the current fairness algorithms by incorporating features such as dynamism, interactivity, and detectability.

5.4. Fairness benchmark

Fairness benchmarks have garnered considerable attention as a driver of algorithmic innovation. Currently, existing datasets containing sensitive information are often used as benchmark datasets for algorithmic fairness. To evaluate the long-term fairness of the algorithm, researchers have combined simulators and benchmark datasets as the evaluation benchmark of the algorithm. However, current simulators lack data on human responses to decisions of algorithms, and the accuracy of the simulation is difficult to guarantee. In the future, system–human interaction and long-term human behavior data will play an essential role in fairness benchmark research.

As shown in the Table 5, for algorithmic fairness benchmarks in the social context, the algorithmic fairness problem definition, algorithmic fairness dataset construction (problem instantiation), and fairness algorithm baseline (solution instantiation) are different from current algorithmic fairness benchmarks. Based on the above chapters on fairness definition, fairness instantiation, and fairness algorithm design, it is necessary to redesign the current fairness benchmark to promote the innovation of algorithmic fairness research in the social context. The newly introduced benchmark should be capable of concretely formulating the problem, instantiating fairness issues, adhere to the new concepts mentioned in Sections 5.1, 5.2, and 5.3, and offering a standardized and quantifiable evaluation approach.

6. Conclusion

In summary, to promote algorithmic fairness in the social context, it is important to consider the interaction between humans and algorithms and to incorporate subject-based definitions of social fairness into algorithm design. Additionally, collecting and simulating interaction data between humans and algorithms or systems, as well as addressing real-world characteristics and maintenance of social fairness are crucial. Finally, a dynamic fairness algorithm that combines subject-system interaction modeling and fairness detection, as well as benchmark refactoring in the social context, are important research directions. By addressing these challenges, we can make progress towards creating algorithms that promote social fairness and contribute to a more fair society.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Project of Guangxi Science and Technology, China (No. GuiKeAD20297004 to Y.H) and the National Natural Science Foundation of China (Grant No. 61967002 to S.T.).

References

[1] W. Dieterich, C. Mendoza, T. Brennan, COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Vol. 7, Northpointe Inc, 2016, p. 1, 7.4.

[2] Y. Wu, J. Cao, G. Xu, FASTER: A dynamic fairness-assurance strategy for session-based recommender systems, ACM Trans. Inf. Syst. (2023).

[3] H. Estiri, Z.H. Strasser, S. Rashidian, J.G. Klann, K.B. Wagholikar, T.H. McCoy Jr., S.N. Murphy, An objective framework for evaluating unrecognized bias in medical AI models predicting COVID-19 outcomes, J. Am. Med. Inf. Assoc. 29 (8) (2022) 1334–1341.

[4] Z. Chen, J.M. Zhang, F. Sarro, M. Harman, A comprehensive empirical study of bias mitigation methods for machine learning classifiers, ACM Trans. Softw. Eng. Methodol. (2023).

[5] A. Aggarwal, P. Lohia, S. Nagar, K. Dey, D. Saha, Black box fairness testing of machine learning models, in: Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2019, pp. 625–635.

[6] S. Biswas, H. Rajan, Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline, in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 981–993.

[7] J. Chakraborty, S. Majumder, T. Menzies, Bias in machine learning software: Why? how? what to do? in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 429–440.

[8] M. Hort, J.M. Zhang, F. Sarro, M. Harman, Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods, in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 994–1006.

[9] S. Udeshi, P. Arora, S. Chattopadhyay, Automated directed fairness testing, in: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, 2018, pp. 98–108.

[10] P. Zhang, J. Wang, J. Sun, G. Dong, X. Wang, X. Wang, J.S. Dong, T. Dai, White-box fairness testing through adversarial sampling, in: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, 2020, pp. 949–960.

[11] A. Wang, O. Russakovsky, Directional bias amplification, in: International Conference on Machine Learning, PMLR, 2021, pp. 10882–10893.

[12] Y. Hirota, Y. Nakashima, N. Garcia, Quantifying societal bias amplification in image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13450–13459.

[13] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.

[14] A. Fabris, S. Messina, G. Silvello, G.A. Susto, Algorithmic fairness datasets: the story so far, Data Min. Knowl. Discov. 36 (6) (2022) 2074–2152.

[15] F. Ding, M. Hardt, J. Miller, L. Schmidt, Retiring adult: New datasets for fair machine learning, in: Advances in Neural Information Processing Systems, Vol. 34, 2021, pp. 6478–6490.

[16] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S.F. Schwemer, A. Sø gaard, FairLex: A multilingual benchmark for evaluating fairness in legal text processing, 2022, arXiv preprint arXiv:2203.07228.

[17] Y. Hu, L. Zhang, Achieving long-term fairness in sequential decision making, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 9, 2022, pp. 9549–9557.

[18] T. Hashimoto, M. Srivastava, H. Namkoong, P. Liang, Fairness without demographics in repeated loss minimization, in: International Conference on Machine Learning, PMLR, 2018, pp. 1929–1938.

[19] Q. Zhang, J. Liu, Z. Zhang, J. Wen, B. Mao, X. Yao, Mitigating unfairness via evolutionary multi-objective ensemble learning, IEEE Trans. Evol. Comput. (2022).

[20] N.A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D.C. Parkes, Y. Liu, How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 99–106.

[21] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, Y. Zhang, Towards Long-Term Fairness in Recommendation, Association for Computing Machinery, New York, NY, USA, 2021.

[22] M. Dolata, S. Feuerriegel, G. Schwabe, A sociotechnical view of algorithmic fairness, Inf. Syst. J. 32 (4) (2022) 754–818.

[23] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, Y. Halpern, Fairness is not static: deeper understanding of long term fairness via simulation studies, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 525–534.

[24] V. Guardieiro, M.M. Raimundo, J. Poco, Enforcing fairness using ensemble of diverse Pareto-optimal models, Data Min. Knowl. Discov. (2023) 1–29.

[25] C. Makri, A. Karakasidis, E. Pitoura, Towards a more accurate and fair SVM-based record linkage, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4691–4699.

[26] S. Liu, Y. Ge, S. Xu, Y. Zhang, A. Marian, Fairness-aware federated matrix factorization, in: Proceedings of the 16th ACM Conference on Recommender Systems, 2022, pp. 168–178.

[27] A. Weber, B. Metevier, Y. Brun, P.S. Thomas, B.C. da Silva, Enforcing delayed-impact fairness guarantees, 2022, arXiv preprint arXiv:2208.11744.

[28] L.T. Liu, S. Dean, E. Rolf, M. Simchowitz, M. Hardt, Delayed impact of fair machine learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 3150–3158.

[29] S. Ahmadian, A. Epasto, M. Knittel, R. Kumar, M. Mahdian, B. Moseley, P. Pham, S. Vassilvitskii, Y. Wang, Fair hierarchical clustering, Adv. Neural Inf. Process. Syst. 33 (2020) 21050–21060.

[30] J. Cho, G. Hwang, C. Suh, A fair classifier using kernel density estimation, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 15088–15099.

[31] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P.W. Michalak, S. Asoodeh, F.P. Calmon, Beyond adult and compas: Fairness in multi-class prediction, 2022, arXiv preprint arXiv:2206.07801.

[32] liobait, Indr, Measuring discrimination in algorithmic decision making, Data Min. Knowl. Discov. 31 (4) (2017) 1060–1089.

[33] S. Yao, B. Huang, Beyond parity: Fairness objectives for collaborative filtering, 2017.

[34] E.S. Jo, T. Gebru, Lessons from archives: Strategies for collecting sociocultural data in machine learning, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 306–316.

[35] C. Kuhlman, L. Jackson, R. Chunara, No computation without representation: Avoiding data and algorithm biases through diversity, 2020, arXiv preprint arXiv:2002.11836.

[36] D. Saha, C. Schumann, D. Mcelfresh, J. Dickerson, M. Mazurek, M. Tschantz, Measuring non-expert comprehension of machine learning fairness metrics, in: International Conference on Machine Learning, PMLR, 2020, pp. 8377–8387.

[37] S. Mohamed, M.-T. Png, W. Isaac, Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence, Philos. Technol. 33 (2020) 659–684.

[38] S. Thacker, D. Adshead, M. Fay, S. Hallegatte, M. Harvey, H. Meller, N. O'Regan, J. Rozenberg, G. Watkins, J.W. Hall, Infrastructure for sustainable development, Nat. Sustain. 2 (4) (2019) 324–331.

[39] J. Zhan, A BenchCouncil view on benchmarking emerging and future computing, in: BenchCouncil Transactions on Benchmarks, Standards and Evaluations, Elsevier, 2022, 100064.

[40] J. Zhan, Three laws of technology rise or fall, in: BenchCouncil Transactions on Benchmarks, Standards and Evaluations, Elsevier, 2022, 100034.

[41] B. Goldman, R. Cropanzano, "Justice" and "fairness" are not the same thing, J. Organ. Behav. 36 (2) (2015) 313–318.

[42] J. Susskind, Future Politics: Living Together in a World Transformed By Tech, Oxford University Press, 2018.

[43] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 12 (3) (2022) e1452.

[44] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.

[45] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Trans. Inf. Syst. 41 (3) (2023) 1–43.

[46] M.K. Lee, A. Jain, H.J. Cha, S. Ojha, D. Kusbit, Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation, Proc. ACM Hum.-Comput. Interact. 3 (CSCW) (2019) 1–26.

[47] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, et al., Towards long-term fairness in recommendation, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 445–453.

[48] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, Algorithmic fairness: Choices, assumptions, and definitions, Annu. Rev. Stat. Appl. 8 (2021) 141–163.

[49] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in criminal justice risk assessments: The state of the art, Sociol. Methods Res. 50 (1) (2021) 3–44.

[50] Y. Li, H. Chen, S. Xu, Y. Ge, J. Tan, S. Liu, Y. Zhang, Fairness in recommendation: A survey, 2022, arXiv preprint arXiv:2205.13619.

[51] C. Simoiu, S. Corbett-Davies, S. Goel, The problem of infra-marginality in outcome tests for discrimination, 2017.

[52] N. Grgic-Hlaca, M.B. Zafar, K.P. Gummadi, A. Weller, The case for process fairness in learning: Feature selection for fair decision making, in: NIPS Symposium on Machine Learning and the Law, Vol. 1, No. 2, Barcelona, Spain, 2016, p. 11.

[53] M.J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[54] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012, pp. 214–226.

[55] M. Zilka, B. Butcher, A. Weller, A survey and datasheet repository of publicly available US criminal justice datasets, Adv. Neural Inf. Process. Syst. 35 (2022) 28008–28022.

[56] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, C. Zhang, How do fair decisions fare in long-term qualification? Adv. Neural Inf. Process. Syst. 33 (2020) 18457–18469.

[57] V. Iosifidis, E. Ntoutsi, Dealing with bias via data augmentation in supervised learning scenarios, Jo Bates Paul D. Clough Robert Jäschke 24 (2018) 11.

[58] L.F. Wightman, LSAC National Longitudinal Bar Passage Study, LSAC Research Report Series, ERIC, 1998.

[59] A. Asuncion, D. Newman, UCI Machine Learning Repository, Irvine, CA, USA, 2007.

[60] B. Strack, J.P. DeShazo, C. Gennings, J.L. Olmo, S. Ventura, K.J. Cios, J.N. Clore, Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records, BioMed Res. Int. 2014 (2014).

[61] P. Van der Laan, The 2001 census in the Netherlands, in: Conference the Census of Population, 2000.

[62] M. Merler, N. Ratha, R.S. Feris, J.R. Smith, Diversity in faces, 2019, arXiv preprint arXiv:1901.10436.

[63] I.-C. Yeh, C.-h. Lien, The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Syst. Appl. 36 (2) (2009) 2473–2480.

[64] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decis. Support Syst. 62 (2014) 22–31.

[65] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: Ethics of Data and Analytics, Auerbach Publications, 2016, pp. 254–264.

[66] F. du Pin Calmon, D. Wei, B. Vinzamuri, K.N. Ramamurthy, K.R. Varshney, Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis, IEEE J. Sel. Top. Sign. Proces. 12 (5) (2018) 1106–1119.

[67] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, Adaptive sensitive reweighting to mitigate bias in fairness-aware classification, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 853–862.

[68] A. Khademi, S. Lee, D. Foley, V. Honavar, Fairness in algorithmic decision making: An excursion through the lens of causality, in: The World Wide Web Conference, 2019, pp. 2907–2914.

[69] R. Feng, Y. Yang, Y. Lyu, C. Tan, Y. Sun, C. Wang, Learning fair representations via an adversarial framework, 2019, arXiv preprint arXiv:1904.13341.

[70] X. Wu, D. Xu, S. Yuan, L. Zhang, Fair data generation and machine learning through generative adversarial networks, in: Generative Adversarial Learning: Architectures and Applications, Springer, 2022, pp. 31–55.

[71] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, A. Roth, A convex framework for fair regression, 2017, arXiv preprint arXiv:1706.02409.

[72] H. Heidari, C. Ferrari, K. Gummadi, A. Krause, Fairness behind a veil of ignorance: A welfare analysis for automated decision making, Adv. Neural Inf. Process. Syst. 31 (2018).

[73] L.E. Celis, V. Keswani, Improved adversarial learning for fair classification, 2019, arXiv preprint arXiv:1901.10443.

[74] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[75] M.B. Zafar, I. Valera, M.G. Rogriguez, K.P. Gummadi, Fairness constraints: Mechanisms for fair classification, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 962–970.

[76] A. Kobren, B. Saha, A. McCallum, Paper matching with local fairness constraints, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 1247–1257.

[77] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq, Algorithmic decision making and the cost of fairness, in: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2017, pp. 797–806.

[78] S. Chiappa, Path-specific counterfactual fairness, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 7801–7808.

[79] R. Nabi, I. Shpitser, Fair inference on outcomes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, No. 1, 2018.

[80] U. Hébert-Johnson, M. Kim, O. Reingold, G. Rothblum, Multicalibration: Calibration for the (computationally-identifiable) masses, in: International Conference on Machine Learning, PMLR, 2018, pp. 1939–1948.

[81] T. Salvador, S. Cairns, V. Voleti, N. Marshall, A. Oberman, Faircal: Fairness calibration for face verification, 2021, arXiv preprint arXiv:2106.03761.

[82] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (3) (2022) 1–44.

[83] S. Caton, C. Haas, Fairness in machine learning: A survey, 2020, arXiv preprint arXiv:2010.04053.

[84] R. Janghel, Y. Rathore, Deep convolution neural network based system for early diagnosis of Alzheimer's disease, IRBM 42 (4) (2021) 258–267.

[85] Y. Huang, N. Wang, S. Tang, L. Ma, T. Hao, Z. Jiang, F. Zhang, G. Kang, X. Miao, X. Guan, et al., OpenClinicalAI: Enabling AI to diagnose diseases in real-world clinical settings, 2021, arXiv preprint arXiv:2109.04004.