SCIENCE CHINA
Information Sciences

• REVIEW •

September 2024, Vol. 67, Iss. 9, 191201:1–191201:21
https://doi.org/10.1007/s11432-023-4060-x

Fairness in machine learning: definition, testing, debugging, and application

Xuanqi GAO[1], Chao SHEN[1*], Weipeng JIANG[1], Chenhao LIN[1], Qian LI[1], Qian WANG[2], Qi LI[3] & Xiaohong GUAN[1]

[1]*Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049*, China;*
[2]*School of Cyber Science and Engineering, Wuhan University, Wuhan* 430072*, China;*
[3]*Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing* 100084*, China*

Received 14 June 2023/Revised 17 September 2023/Accepted 10 January 2024/Published online 15 August 2024

Abstract In recent years, artificial intelligence technology has been widely used in many fields, such as computer vision, natural language processing and autonomous driving. Machine learning algorithms, as the core technique of AI, have significantly facilitated people's lives. However, underlying fairness issues in machine learning systems can pose risks to individual fairness and social security. Studying fairness definitions, sources of problems, and testing and debugging methods of fairness can help ensure the fairness of machine learning systems and promote the wide application of artificial intelligence technology in various fields. This paper introduces relevant definitions of machine learning fairness and analyzes the sources of fairness problems. Besides, it provides guidance on fairness testing and debugging methods and summarizes popular datasets. This paper also discusses the technical advancements in machine learning fairness and highlights future challenges in this area.

Keywords artificial intelligence security, machine learning security, machine learning fairness, model testing, model debugging

1 Introduction

In recent years, artificial intelligence technology represented by machine learning (ML) algorithms has been constantly developing and innovating, and has empowered various aspects of people's daily life. It has achieved remarkable success in various tasks such as computer vision, natural language processing (NLP), speech recognition and intelligent medical care.

However, the development and application of various artificial intelligence technologies have led to an increasing social impact, and their fairness risks are gradually being exposed. In 2016, ProPublica reported on an ML fairness issue that resulted in discrimination in crime prediction. Despite having similar backgrounds and experiences, some prisoners with only racial differences received significantly different risk levels in the ML-driven crime prediction system. Gender-neutral pronouns translated into English by the Google Translate API showed strong male default bias [1]. Common commercial facial classification systems such as Face++ and Microsoft's facial recognition API have also been found to exhibit gender and racial differences, with classification error rates for dark-skinned women several times higher than for light-skinned men in their prediction results [2].

The bias towards fairness in existing ML systems has been brought to the forefront, emphasizing the need for in-depth research on the fairness of ML models. Despite the variety of fairness definitions, testing methods, and repair approaches proposed in existing research, a complete technical path has not yet been established due to their diverse and focused forms. Furthermore, the limited understanding of black-box characteristics and interpretable mechanisms of ML models in existing research has hampered the effectiveness of testing and repair methods. To address these issues, there is an urgent need to comprehensively summarize, analyze, and discuss the existing research in the field of fairness of ML

* Corresponding author (email: chaoshen@mail.xjtu.edu.cn)

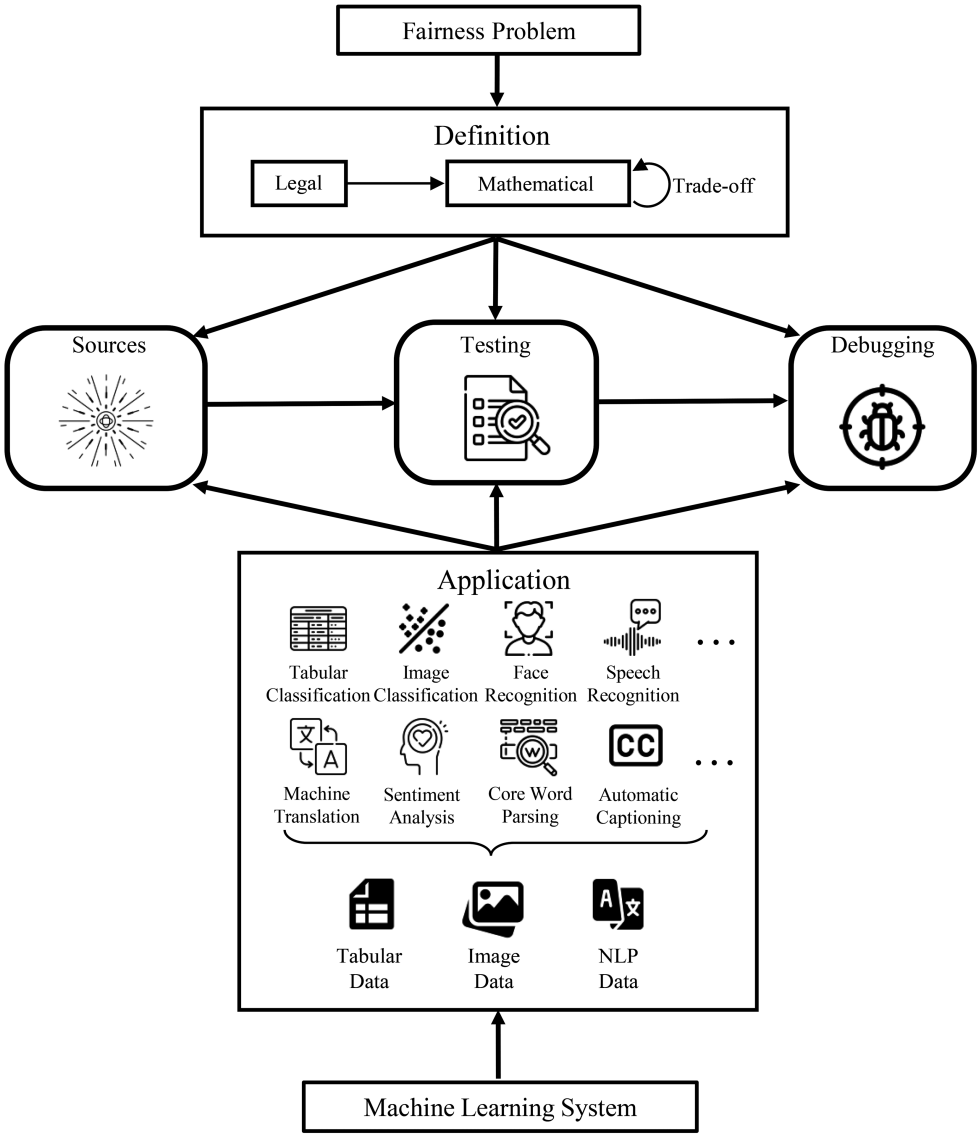

Figure 1 Overview of this study.

models, in order to identify any shortcomings and provide meaningful guidance for researchers in related fields.

Existing surveys mainly focus on the sources and concepts of fairness, and fairness related to tabular data. However, the current landscape of ML encompasses diverse data types, including tabular data, image data, natural language data, and multimodal data consisting of a combination of them. These data exhibit distinct data structures and yield varying task outputs, consequently leading to different definitions and considerations of fairness issues. There is a lack of comprehensive research that addresses emerging fairness issues in image data and NLP data. This paper aims to fill this gap by analyzing the similarities, differences, and connections of fairness issues in three fields: tabular data, image data, and NLP data. The goal is to promote the further development of fairness in ML and provide guidance and references to ensure the widespread application of ML-related technologies.

The overview of this paper is shown in Figure 1. To discuss and address the fairness problem of ML systems, we begin by establishing fundamental definitions for the prevalent categories of fairness issues. Considering the entirety of the ML system's lifecycle, we discuss the existing techniques aimed at addressing fairness problems in a structured three-step process, i.e., sources-testing-debugging. The sources of fairness problems affect the object chosen for testing, i.e., whether the ML algorithm or the training data. Subsequently, based on the specific issues and anomalies identified during testing, we

employ appropriate debugging algorithms to enhance fairness. Notably, a primary highlight of this paper is its categorization of common ML applications into three types: tabular-based, image-based and NLP-based systems. This categorization is pivotal because distinct application types manifest fairness issues differently, influencing various aspects such as test case generation and model repair methodologies.

More detailed, the paper is organized as follows. Section 2 provides a review of the origins and mathematical definitions of common definitions of fairness in ML. Section 3 summarizes, analyzes, and discusses the causes of fairness problems in ML. Section 4 reviews the research on testing and repairing fairness in ML, and discusses the effectiveness and limitations of existing research. Section 5 presents a review and analysis of existing datasets on ML fairness in academia. Section 6 discusses the similarities, differences, and connections of fairness issues in different fields, and analyzes the challenges and future research directions in ML fairness. Section 7 concludes the paper. By addressing the gaps in the current literature, this paper provides valuable insights into the evolving field of fairness in ML.

2 Fairness definition

To study fairness in ML, it is essential to define what fairness means and clarify its mathematical form and practical implications. In this section, we will analyze legal and mathematical definitions of fairness and examine the trade-offs of fairness indicators.

The definition of fairness may vary among different disciplines and scholars. John Locke's "Two Treatises of Government" [3] regarded fairness as one of the foundations of social and political order. He argued that every individual has natural rights, including property rights and freedom, and the role of government is to protect these rights and ensure fair distribution of resources and opportunities. John Rawls's "A Theory of Justice" [4] introduced the famous "difference principle", which states that social and economic inequalities are justifiable only if they benefit the least advantaged and are combined with the fairness of opportunities for all. He believed that fairness requires the achievement of social justice through the just distribution of opportunities and resources. Amartya Sen's "The idea of justice" [5] explored the multidimensional nature of fairness and justice. He emphasized that fairness goes beyond equal distribution of wealth and resources and includes equal treatment of people's capabilities, rights, and freedoms. Mehrabi et al. [6] argued that fairness refers to the absence of bias or favoritism in the decision-making process based on an individual's inherent or acquired characteristics. While fairness is a widely accepted desirable quality, it is challenging to define or achieve in practice. Researchers define fairness based on sensitive or protected characteristics such as race, gender, sexual orientation, religion, disability status, or place of birth. These characteristics have been subjected to systematic discrimination in the past. However, the dataset's feature set cannot always be clearly divided into "neutral" and "sensitive" categories. Seemingly neutral features can often accurately predict group membership, and a classifier for sensitive features can be built using the remaining features if sensitive features are considered as the target variable in a classification problem [7]. For instance, until the 1960s, literacy tests were used as a method that appeared race-neutral on the surface, but was actually aimed at depriving African Americans and other groups of their voting rights [8].

2.1 Legal definition

The selection of sensitive characteristics in ML models can have significant implications as it affects which groups are focused on and the conclusions drawn from the data analysis. If the taxonomy used is too general, too specific, misleading, or inaccurate, it can lead to harm. Furthermore, the act of categorizing identities into sensitive categories and gathering relevant data can be problematic itself. To understand why there are so many definitions of fairness, it is important to analyze the various types of discrimination that can occur.

(1) Direct discrimination. Direct discrimination occurs when an individual is treated unfairly due to their sensitive characteristics, such as race, gender, age, religion, disability, or nationality [9]. Discrimination against these protected characteristics is illegal according to many laws. For example, the Fair Housing Act and the Equal Credit Opportunity Act in the United States provide protection against discrimination based on certain sensitive characteristics.

(2) Indirect discrimination. Indirect discrimination, also referred to as disparate impact, occurs when a policy or measure seems to be neutral but can have a more negative impact on certain groups compared to others. While individuals may seem to be treated fairly based on seemingly neutral and unprotected

characteristics, protected groups or individuals can still be treated unfairly due to the implicit effects of their protected characteristics. For instance, a person's residential zip code can be used in decision-making processes, such as loan applications. However, this can still result in racial discrimination since zip codes, although seemingly insensitive, are likely associated with race due to varying demographic distributions of residences [9]. To prevent direct discrimination as much as possible, ML models can be trained without directly using data that contains sensitive features (i.e., explicitly protected features). However, this approach may not necessarily eliminate the possibility of indirect discrimination.

(3) Systemic discrimination. Systemic discrimination refers to policies, practices, or behaviors that are embedded within the culture or structure of an organization and may perpetuate discrimination against certain demographic subgroups [10]. Research by Kline et al. [11] found that employers tend to favor candidates who share similar experiences and preferences to their own culture, which may result in discrimination against qualified candidates who do not belong to these subgroups if the decision makers belong overwhelmingly to certain subgroups.

(4) Statistical discrimination. Statistical discrimination occurs when a decision maker uses the average statistics of a group to judge individuals who belong to that group. This typically happens when a decision maker, such as an employer or a law enforcement officer, uses a person's obvious, identifiable characteristics as a proxy for hidden or more difficult to identify characteristics that may be relevant to the outcome [12].

(5) Explainable discrimination. Explainable discrimination refers to differences in treatment or outcomes between different groups that can be justified and explained by certain characteristics or factors. It is considered acceptable and legal because the differences can be explained by legitimate reasons. For example, if men and women have different average earnings, this can be explained by the fact that women work fewer hours than men. If an organization were to ignore this factor and try to make the earnings equal, it would result in reverse discrimination, where male employees would earn less than female employees. Quantifying explainable and illegitimate discrimination is important to ensure that decisions are fair and unbiased. Kamiran et al. [13] proposed a method to measure discrimination in data or classifier decisions that directly consider both illegal and explainable discrimination. They explained how to quantify and measure discrimination in the data or classifier decisions that directly consider illegal and explainable discrimination.

(6) Unexplained discrimination. Unexplained discrimination refers to discrimination against a group that cannot be justified or explained by any legitimate characteristics or factors. It is considered unlawful and unjustified. Kamiran et al. [13] proposed a technique that can identify and remove only unexplained or illegal discrimination, while preserving explainable differences in decision-making. This approach ensures that discrimination is eliminated only when it is unjustifiable, while allowing legitimate differences that can be explained by relevant characteristics.

2.2 Mathematical definition

The legal definition of fairness has its roots in sociological considerations, but to apply this definition, a mathematical formalization of fairness is necessary. We first introduce the most commonly used group fairness definition, and then the introduce individual fairness definition.

(1) Disparate impact [14]. It is a mathematical expression of the legal definition of indirect discrimination, which is among the most widely used group fairness definition. This definition necessitates that the ratio of the predictive accuracy of the two groups is near 1, signifying that both groups are equally accurately predicted by the model:

$$\frac{P[\hat{Y}=1|S\neq 1]}{P[\hat{Y}=1|S=1]}\geqslant 1-\varepsilon, \tag{1}$$

where $P[\cdot]$ represents the probability or frequency of occurrence of a certain condition, ε represents a small enough tolerance, S represents the sensitive features, and $\hat{Y}$ represents whether the model gives a positive result. A higher value of this metric indicates greater equity between the groups. For instance, in the context of job applications, if the model predicts a positive outcome (i.e., acceptance) for individuals with a value of 1 for the characteristic S, this metric requires that the proportion of accepted applicants is not significantly different across groups [15]. The "80% rule" [16] of the disparate impact approach is commonly applied to this concept, requiring that the acceptance rate for any racial, gender, or ethnic group is at least 80% of that of the group with the highest acceptance rate.

(2) Demographic parity [17,18]. Also known as statistical parity, this definition is similar to disparate impact, but takes the mathematical form of a difference rather than a ratio.

$$|P[\hat{Y}=1|S\neq 1]-P[\hat{Y}=1|S=1]|\leqslant\varepsilon. \tag{2}$$

The lower the value of the metric, the more similar the acceptance rate and therefore the better the fairness performance.

Both of these metrics aim to ensure that positive predictions are distributed to different groups in a similar proportion. However, a clear disadvantage of both criteria is that a perfectly equal classifier can result in inequity when actual differences in ability exist between groups [19]. Moreover, in order to satisfy the statistical equality criterion, two similar individuals may receive different results simply because they belong to different groups, which also leads to inequity.

(3) Equalized odds [20]. This metric aims to address the limitations of the previously mentioned metrics. It computes the difference between false positive and true positive rates for various groups. In other words, the probability of accurately predicting a positive outcome for a positive class and the probability of incorrectly predicting a positive outcome for a negative class should be equal for both protected and unprotected groups (such as male and female).

$$|P[\hat{Y}=1|S\neq 1,Y=a]-P[\hat{Y}=1|S=1,Y=a]|\leqslant\varepsilon,\quad a\in\{0,1\}. \tag{3}$$

Unlike demographic parity and disparate impact metrics, this definition requires both true positive and false positive rates to be equal for both protected and unprotected groups. However, the assumption underlying this indicator is that the underlying ratios of the two groups are representative and not obtained in a biased manner. An example of a system that was found to violate this indicator is the COMPAS (correctional offender management profiling for alternative sanctions) algorithm used in the U.S. criminal justice system [15]. The algorithm was found to have different odds of accuracy for African Americans and Whites in predicting recidivism by prior offenders. Specifically, it incorrectly predicted future offenses for African Americans at twice the rate predicted for Whites (false positive rate) and significantly underestimated the likelihood of future offenses for Whites (true positive rate).

(4) Equal opportunity [20]. This metric is similar to equality of odds, but focuses only on the true positive rate, and calculates the ratio of true positives between different groups.

$$|P[\hat{Y}=1|S\neq 1,Y=1]-P[\hat{Y}=1|S=1,Y=1]|\leqslant\varepsilon. \tag{4}$$

These definitions focus on the fairness of the whole group and are therefore called group fairness definition. However, there is another important type of fairness definition known as individual fairness, which emphasizes that similar individuals should receive similar outputs from a classifier.

$$d(M(x),M(y))\leqslant d(x,y). \tag{5}$$

This means that if two individuals, x and y, are similar according to a defined similarity metric, then their predicted outcomes by the classifier, M, should also be similar. To measure individual fairness, a similarity metric needs to be defined, which can be derived from various individual fairness metrics [21,22].

Besides, causal inference-based individual fairness metrics are another type of individual fairness measure [23]. For instance, counterfactual fairness is a definition of fairness based on causal inference, which requires testing whether the predicted values of a counterfactual world, where sensitive features are reversed, are the same as those of the real world [18]. If they are the same, the algorithm is considered to be fair to individuals.

2.3 Fairness trade-off

Multiple notions of fairness cannot be satisfied simultaneously by the same system. For example, if there are real differences in ability between different groups, a classifier cannot achieve both population parity and equality of odds. In the case of the COMPAS algorithm, to achieve population parity, the algorithm needs to select equal proportions of male and female prisoners for parole, which may potentially result in the mistaken incarceration of females with lower risks [15]. As a result, the principle of population parity can increase false positive errors in one group and false negative errors in the other, thus undermining equality of odds. To address this inherent incompatibility, it is recommended to prioritize one fairness

metric based on the specific application's requirements. Pleiss et al. [24] suggested that any chosen measure of algorithmic fairness should be evaluated in the appropriate legal, social, and ethical context.

Individual and group fairness can also be incompatible, as illustrated by the example of the female prisoners who would be treated unfairly even if her data were similar to that of a male. Another challenge is the trade-off between fairness and accuracy, where improving one metric may come at the expense of the other. Previous research has analyzed this trade-off and found that adding a fairness constraint to the optimization problem typically leads to a decrease in classifier accuracy [8, 25–27]. However, there are cases where increasing fairness can improve accuracy, and a perfectly accurate classifier is considered perfectly fair when using equality of odds as a fairness metric [28, 29]. Nevertheless, some recent studies have shown that the two metrics of equilibrium accuracy and equality of odds may be incompatible [15].

Summary. Existing definitions of fairness have limitations and are not universally applicable. There is no one way to measure fairness, and researchers must tailor their approach to real-world scenarios. For instance, in predicting recidivism, the short- and long-term effects of risk assessment on public safety and the incarcerated population size, as well as whether an instrument is consistent with due process principles, should be considered. Similarly, in lending, the immediate and long-term effects on community development and lending programs should be taken into account. Mathematical measures of fairness cannot fully address these concerns, and there is a need for better quantification and balance of the costs and benefits associated with algorithmic interventions.

3 Sources of fairness issues

The concepts of bias and discrimination are closely related to equity. A system is considered fair when its results do not discriminate based on certain sensitive characteristics such as gender or nationality. In ML evaluation, discrimination can be estimated by analyzing the confusion matrix for different groups, including positive rate, true positive rate, false positive rate, and other metrics. Large disparities in these rates suggest that a predictive system may be behaving unfairly or exhibiting bias in its decisions about a sample with a particular characteristic. However, it is worth noting that bias is not always a source of unfairness, as statistical biases can favor inference. For example, a car in disrepair is more likely to be involved in an accident; a patient with a chronic disease may have a greater risk of deterioration. What constitutes unfair bias depends on historical and social context, as well as legal requirements.

In this section, we aim to outline the sources of fairness problems in tabular ML systems that are based on tabular data, then expand upon these definitions and define the sources of fairness problems for image systems that are based on image data, as well as NLP systems that are based on text data.

3.1 Sources of fairness issues for tabular systems

In this paper, the term "tabular ML systems" refers to ML systems trained using tabular data. Tabular data is a type of formatted data that can be represented in rows and columns, where each row in a table represents a data sample, and each column represents a feature. Tabular data is the most commonly used type of data in ML and is the type of data that most decision-making systems currently need to handle. Below, we will introduce the sources of biases in tabular data. Note that biases in image data or NLP data can also stem from these aspects.

3.1.1 *Historical biases*

Historical biases are biases and socio-technical issues that already exist in society and can be introduced through the data generation process, even in the presence of perfect sampling and feature selection. Such biases are already present in the dataset used for learning and are typically due to inaccuracies in device measurements, historically biased human decisions, misreporting, or other factors. These biases, even when they accurately reflect reality, can cause harm to vulnerable populations and therefore need to be addressed. For instance, in 2018, only 5% of Fortune 500 CEOs were women, but the image search results for "CEO" should not reflect this reality by the same percentage [30].

3.1.2 *Representation biases*

Data, especially big data, is often unevenly generated by different subpopulations with their own behavioral characteristics, which can bias the data and affect the effectiveness of ML. Representation bias

occurs when some parts of the input space are underrepresented, leading to a learned distribution that is biased. For example, a dataset collected through a smartphone app may be underrepresented for low-income or elderly groups who may not own smartphones. Similarly, census data from thirty years ago may not reflect today's population, and ImageNet's dataset contains significantly more images from North America and Western Europe than other regions, resulting in certain classifiers being less accurate on images from other regions [31].

3.1.3 *Training biases*

The model training process may also introduce biases. One such bias is rooted in the algorithm's objective, which is to minimize the total prediction error and may thus favor the majority group over the minority group. This often leads to sensitive characteristics that distinguish privileged from non-privileged groups, such as race, gender, and age, being used illegitimately in decision-making to achieve higher accuracy [32]. The second bias is rooted in model quality, where the difficulty of obtaining data for minority groups or their low commercial value leads to poorer quality models for these groups that do not provide the same level of service as the majority group. This is exemplified by the case of face recognition systems.

3.2 Sources of fairness issues for image-based systems

Image-based ML systems, which have been one of the most successful applications in recent years, have diverse sources of fairness problems. One source of bias is the model itself, as the training procedure may introduce fairness issues. For example, deep generative models tend to be biased with respect to key demographic factors, as found by Choi et al. [33]. Image caption models also tend to amplify the gender bias present in the training data, as reported by Hendricks et al. [34].

Another source of bias is the long-tail problem, which refers to datasets with large disparities in the amount of data between different classes. This can lead to poorer prediction performance of models for rare classes, reducing fairness. Xu et al. [35] and Benz et al. [36] have found a similar phenomenon in adversarial training.

The input data itself can also induce predictive discrimination, even if protected features such as race, gender, and age are not explicitly included as inputs [37]. For example, Kärkkäinen and Joo [38] found that existing public face datasets have a strong preference for Caucasians, while other races are significantly underrepresented. Manjunatha et al. [39] found that visual question answering (VQA) systems often exhibit a bias towards factors such as gender. Buolamwini and Gebru [2] discovered that commonly used commercial gender classification systems have a classification error rate for dark-skinned females that is tens of times higher.

3.3 Sources of fairness issues for NLP systems

Deep learning models for NLP have diverse applications across various domains, including web search engines, virtual assistants, chatbots, language translation tools, sentiment analysis platforms, and more. NLP systems rely on language models, which estimate the probability of a sequence of words to predict the most likely next word in a given sequence. These models are used for tasks like machine translation, text generation, and text classification. However, stereotype bias can arise in language models due to prior words associated with protected features. For example, Bolukbasi et al. [32] found that word embedding techniques exhibit gender stereotypes, such as associating "programmer" with "male" and "housewife" with "female". In addition, equity issues can also arise in other NLP applications, such as speech recognition systems. Tatman [40] found that the accuracy of Google's speech recognition system for generating captions was influenced by gender and dialect, with lower accuracy for women. Moreover, NLP systems can inadvertently reinforce existing biases and inequalities. For example, biased search results or recommendations can lead to users being exposed to more biased content, creating a feedback loop that perpetuates unfairness [41, 42].

4 Fairness testing & debugging

ML models can suffer from various biases, which can significantly impact their prediction effectiveness, work quality, and have negative social consequences. Therefore, it is essential to systematically test and fix the fairness of ML models. This section delineates the fairness testing and debugging methods for

ML systems, specifically tailored to tabular data, image data, and NLP data. This division is warranted for several reasons. Firstly, it aligns with the categorization of mainstream ML application scenarios into these three categories. Secondly, distinct data structures necessitate the development of unique testing and remediation processes for each category. Lastly, task outputs differ: tabular and image data predominantly involve classification tasks, whereas NLP data encompasses a broader range of task types, including translations and automatic subtitles. These inherent disparities underscore the importance of introducing and deliberating upon their individual characteristics. The overview of this chapter is shown in Table 1 [1, 2, 32–34, 37, 43–89].

4.1 Tabular testing

ML systems based on tabular data are able to discriminate and change sensitive features relatively simply, and therefore their research results are more numerous and their testing tools are simpler. Biswas and Rajan [43] and Valentim et al. [44] tested whether data processing methods of ML algorithms introduced fairness errors using causal inference. Zhang et al. [46] detected paths related to fairness in decision trees and random forest models. NeuronFair [49] identified biased neurons responsible for unfairness through neuronal analysis and then generated discriminative instances with the aim of increasing the activation difference values of unidentified biased neurons. DeepFAIT [48] applied significance testing to detect the difference in activation between disadvantaged and non underprivileged neurons' activation differences to identify fairness-related neurons. Vig et al. [45] used causal mediation analysis [90] to test which parts of a DNN model are causally related to its unfair predictions. Gao et al. [47] proposed FairNeuron, which identifies conflicting paths of ML models through neuron slicing techniques.

Other approaches for detecting fairness errors in ML models involve manipulating the input data. The Themis [50], for example, generates random test inputs and checks if the output of the model is the same for individuals with only different sensitive feature values. Aequitas [51] and ExpGA [52] search the input space for discriminatory instances that lead to unfair predictions. Chakraborty et al. [53] used k-nearest neighbors to interpret model predictions and reveal bias. White box testing methods, such as ADF [54,55] and EIDIG [56], examine the internal structure of the model and exploit gradient information to test input generation for model fairness.

4.2 Tabular debugging

There are various approaches to address the fairness problem in tabular ML systems. Fairway, proposed by Chakraborty et al. [57], optimizes ML fairness by searching for model hyperparameters. Tizpaz-Niari et al. [58] proposed Parfait-ML, which uses dynamic search algorithms to approximate the Pareto optimum of the model hyperparameters. Zhang et al. [46] utilized the MaxSMT solver to identify which paths in a decision tree model can be flipped or refined, and refine the decision tree paths to improve fairness. FairNeuron [47] identifies biased instances and enforces the model to learn important fairness features. Additionally, researchers have retrained models based on test sample generation in fairness testing to improve model fairness [51, 55].

4.3 Image testing

In the context of ML systems using tabular data, sensitive features are often represented as discrete or binary variables. However, when working with image data, the high-dimensionality of the feature space typically precludes explicit representations of individual features. Instead, more abstract or implicit representations, such as embeddings or feature maps, may be used to capture relevant patterns and relationships among the underlying data. Consequently, test and repair procedures for image data often involve intricate preprocessing steps and possess a more extensive repertoire of fairness optimization tools.

Joo and Kärkkäinen [59] conducted a black box test for counterfactual fairness on image classification APIs from major companies, such as Google, Amazon, and Clarifai. Meanwhile, McDuff et al. [60] proposed a simulation-based testing method that uses facial image generation models with Bayesian optimization to evaluate the fairness of facial image classification systems. Buolamwini and Gebru [2] introduced cross-sectional demographics as a means of detecting subgroup bias in automated facial analysis algorithms and datasets. Similarly, Schaaf et al. [37] performed bias assessments on image classification models based on convolutional neural networks (CNN) by synthesizing biased datasets manually and generating attribution graphs.

Table 1 Comparison of typical techniques for ML fairness testing and debugging

Function description	Method category	Method description	Related work
Tabular system testing	Causal method	Assessing fairness of preprocessing stage based on causal inference	Biswas & Rajan [43]
		Assessing fairness of data preparation procedures based on causal inference	Valentim et al. [44]
		Testing the fairness of DNN model components based on causal mediation analysis	Vig et al. [45]
	Path method	Test fairness-related paths in the model	Zhang et al. [46]
		Identifying model conflict paths based on neuron slicing technique	Gao et al. [47]
		Apply significance tests for neuronal activation differences to assess fairness	Zhang et al. [48]
		Testing for the presence of biased neurons based on neuronal analysis	Zheng et al. [49]
	Black box test generation	Randomly generate test data and compare the output	Angell et al. [50]
		Search input space and find discrimination instances	Udeshi et al. [51]
		Search for discrimination instances based on genetic algorithm guidance	Fan et al. [52]
		Assessing model fairness based on k-nearest neighbor approach	Chakraborty et al. [53]
	White box test generation	Search for discrimination instances based on gradient information	Zhang et al. [54, 55]
		Successive iterations for instance generation under gradient guidance	Zhang et al. [56]
Tabular system debugging	Hyperparameter search	Search for model hyperparameters to optimize fairness	Chakraborty et al. [57]
		Dynamic search for hyperparameters to approximate Pareto optimality	Tizpaz-Niari et al. [58]
	Path method	MaxSMT solver-based refinement of decision trees for paths	Zhang et al. [46]
		Identify discrimination instances and force the model to learn important features	Gao et al. [47]
	Retraining	Search the input space to find discrimination instances and retrain	Udeshi et al. [51]
		Search for discrimination instances based on genetic algorithm and retrain	Fan et al. [52]
		Search for discrimination instances based on gradient information and retrain them	Zhang et al. [54, 55]
		Successive iterative generation of instances under gradient guidance and retraining	Zhang et al [56]
Image system testing	Generative method	Counterfactual black box testing of image classification API	Joo & Kärkkäinen [59]
		Generate facial images to test facial image classification systems	McDuff et al. [60]
		Artificially synthesize datasets and generate attribution maps to test image systems	Schaaf et al. [37]
	Model compression test	Testing the effect of pruning and quantification on gender bias	Hooker et al. [61]
		Testing the effects of distillation and pruning on the fairness of generative models	Xu & Hu [62]
		Testing the impact of model compression on facial expression recognition systems	Stoychev & Gunes [63]
	Statistical method	Evaluating Gender Bias in Automated Facial Analysis Algorithms	Buolamwini & Gebru [2]
Image system debugging	Discriminative model testing	Introduction of new models to mitigate gender bias	Hendricks et al. [34]
		Construct balanced classifiers based on undersampling and oversampling	Chawla et al. [64]
		Build classifiers based on oversampling of classes near the classification boundary line	Han et al. [65]
		Combining integrated learning with data generation to repair classifiers	Guo & Viktor [66]
	Generative model testing	Preprocessing datasets to improve generative image fairness	Sattigeri et al. [67]
		Introduce supervised signals to mitigate image generation bias	Choi et al. [33]
		Altering image semantics to meet fairness criteria	Quadrianto et al. [68]
		Separating sensitive attribute signals for counterfactual intervention	Joo & Kärkkäinen [59]

(To be continued on the next page)

(Continued)

Function description	Method category	Method description	Related work
NLP fairness testing	Word similarity test	Check for bias in word embedding	Caliskan et al. [69]
		Check for stereotype associations between groups of words	Dev et al. [70]
	Sentence template testing	Association testing based on sentence encoders	May et al. [71]
		Test for attribute words based on sentence templates	Kurita et al. [72]
		Testing with templates with two empty spaces	Webster et al. [73]
	Crowdsourced dataset testing	Measure the degree of stereotyping of the model in contextual associations	Nadeem et al. [74]
		Contrast model bias for minimum distance sentence combinations	Nangia et al. [75]
	Extrinsic bias test	Testing gender bias in occupational prediction systems	De-Arteaga et al. [76]
		Testing the bias of the occupational prediction system on name	Romanov et al. [77]
		Testing gender bias in sentiment analysis systems	Kiritchenko et al. [78]
		Testing core word parsing systems for gender bias	Rudinger et al. [79]
		Testing the accuracy of the core word parsing system for different genders	Zhao et al. [80]
		Testing the core word parsing system for name bias	Webster et al. [81]
		Testing machine translation systems for gender bias	Stanovsky et al. [82]
		Testing gender bias when translating pronouns with Google Translate API	Prates et al. [1]
NLP fairness debugging	For raw dataset	Mitigating data bias based on counterfactual data augmentation	Zmigrod et al. [83]
		Avoiding text duplication based on random substitution	Maudslay et al. [84]
		Reducing dataset bias through rebalancing strategies	Dixon et al. [85]
	For word embedding process	Mitigate word embedding gender bias based on projection	Bolukbasi et al. [32]
		Iteratively build linear classifiers to mitigate potential bias	Ravfogel et al. [86]
		Orthogonalize sensitive feature spaces to mitigate bias and retain information	Dev et al. [87]
		Use projections at the sentence level to mitigate bias	Liang et al. [88]
	For training process	Force the model to maintain gender-neutrality when learning word embedding information	Zhao et al. [89]
		Mitigate model gender bias based on dropout regularization	Webster et al. [73]

Given the large model sizes of image-based systems, it is often necessary to compress these models for deployment on platforms with limited computational resources. However, model compression may introduce fairness issues, as researchers have shown. For example, Hooker et al. [61] demonstrated that pruning and quantization amplify gender bias when classifying hair color in computer vision datasets. In contrast, Xu and Hu [62] tested the effects of distillation and pruning on generative model bias and found empirical evidence to suggest that distillation models exhibit less bias. Finally, Stoychev and Gunes [63] identified fairness errors introduced by different model compression algorithms in various facial expression recognition systems.

4.4 Image debugging

There are various methods for debugging fairness in image-based systems, which can be broadly categorized into two types: discriminative models and generative models.

4.4.1 *Discriminative models*

Hendricks et al. [34] proposed the Equalizer model, which reduces gender bias in image caption prediction systems. Chawla et al. [64] proposed the synthetic minority over-sampling technique (SMOTE) to build classifiers from imbalanced datasets. By oversampling minority classes and undersampling majority classes, SMOTE achieves better classifier performance than undersampling alone. Han et al. [65] improved SMOTE by introducing Borderline-SMOTE, which oversamples classes near the classification boundary line to improve classifier performance. Guo and Viktor [66] proposed the DataBoost-IM

method, which combines integrated learning and data generation to improve classifier performance on imbalanced datasets.

4.4.2 *Generative models*

Sattigeri et al. [67] proposed fairness-GAN, a generative adversarial network (GAN) for high-dimensional image data that improves population parity and equality of opportunity by preprocessing the dataset. Choi et al. [33] proposed a weakly supervised algorithm that mitigates demographic bias in GAN by using a small, unlabeled reference dataset as a supervised signal. Quadrianto et al. [68] ensured fairness criteria by semantically altering images in the dataset, such as migrating male images to female. Joo and Kärkkäinen [59] performed counterfactual fairness intervention by separating the sensitive feature signals from the rest of the image and synthesizing facial images that differ only in the sensitive feature dimension.

4.5 NLP testing

Typically, NLP models are fine-tuned by the deployer on downstream tasks, building upon pre-trained models. This approach eliminates the need to train new models from scratch. Hence, fairness biases in NLP systems can be classified into two categories, intrinsic bias and extrinsic bias, based on their origin during the pre-training and fine-tuning phases, respectively. In the context of NLP systems, the challenge of mitigating fairness concerns arises because the sensitive attributes of the models are not consistently discernible, a situation akin to that encountered with image data. Furthermore, NLP data exhibits a heightened degree of task diversity and intricacy compared to tabular and image data [91]. These features necessitate NLP fairness methods to possess greater granularity, finer categorization, and expedited update iterations.

4.5.1 *Intrinsic bias*

Intrinsic bias testing focuses on detecting and addressing fairness issues that arise during the pre-training phase of NLP models. These tests can be broadly categorized into three main types: word similarity-based testing, sentence template-based testing, and crowdsourced dataset-based testing.

(1) Word similarity-based testing. Word vector embeddings have been found to contain and amplify biases present in the data they extract. To detect such biases, the word embedding association test (WEAT) has been proposed [69]. WEAT performs hypothesis testing on two sets of target words (e.g., programmer, engineer, and nurse, teacher) and two sets of attribute words (e.g., man, male, and woman, female) to determine whether there is a difference in the relative similarity of these words. The results are used to evaluate the presence of bias in word embeddings. The embedding coherence test (ECT) [70] assesses whether groups of words are stereotypically associated, instead of assessing the similarity of exact words such as gender words and occupational words. ECT first calculates the mean of gender words separately, then the cosine similarity of the mean of occupational words and gender words separately, and finally the Spearman coefficient after ranking the similarity list by ranking similarity degree to measure the fairness of word embedding.

(2) Sentence templates-based testing. In order to expand the WEAT to encompass sentence-level semantics, May et al. [71] proposed the sentence encoder association test (SEAT). Kurita et al. [72] developed the log-probability bias score (LPBS) algorithm, which utilizes sentence templates for attribute words and target words, and then applies the WEAT to these sets of sentences. The LPBS algorithm incorporates prior probabilities based on sentence templates, thereby ensuring that any observed differences between attribute words can be attributed to the attribute words themselves, rather than to their prior probabilities. Webster et al. [73] proposed the DisCo algorithm, which is also based on sentence templates, but with the difference that DisCo employs a template with two empty spaces (e.g., "[A] likes [B]"). By filling in the first empty space with the target word and calculating the model's prediction for the second empty space, we can determine whether the model is biased or not.

(3) Crowdsourced datasets-based testing. Nadeem et al. [74] introduced the contextual association test, which measures the extent of model stereotype bias. This approach provides a crowdsourced dataset called StereoSet, where each entry comprises a sentence with masking (e.g., "Our housekeeper is [a word]") and three candidate complementary options for the sentence, one of which is a stereotype, one of which is an anti-stereotype, and one of which is a completely irrelevant word. The degree of stereotyping by a

model can be assessed by calculating the percentage of stereotype options that a model favors. CrowS-Pairs [75], also designed to evaluate stereotype bias, has each data entry consisting of a pair of sentences with minimal distance, where one sentence contains a historical stereotype of a disadvantaged group that is not present in the other sentence, such as "People who live in slums are alcoholics" and "People who live in mansions are alcoholics". The degree of model bias is quantified by measuring the proportion of sentences in each pair for which the model prefers the stereotype sentence.

4.5.2 *Extrinsic bias*

Extrinsic bias testing aims to investigate how biases in pre-trained models affect downstream tasks and their potential impact on the real world, rather than focusing on the biases within the models themselves.

(1) Job searching and recruitment systems. De-Arteaga et al. [76] examined gender bias in occupational classification by predicting occupations based on resumes and using the difference in accuracy between male and female applicants as a test metric. Romanov et al. [77] also investigated the relationship between occupational classification and names using the same resume-occupation prediction task.

(2) Sentiment analysis systems. Kiritchenko and Mohammad [78] evaluated automatic sentiment analysis systems using the emotion in context (EEC) corpus, which contains sentences that differ by only one sensitive feature (e.g., gender), and measured the deviation between the sentiment scores as a test indicator.

(3) Core word parsing systems. Most core word parsing systems are evaluated on the Winograd Schema Challenge task, which asks machines to identify the antecedent object referred to by ambiguous pronouns in a statement. Rudinger et al. [79] created the Winogender dataset, which explores pronoun gender only and evaluates several publicly available core word parsing systems. Zhao et al. [80] introduced the WinoBias dataset, which uses the accuracy of core word parsing systems in handling different gender pronouns as a test metric, while Webster et al. [81] developed the GAP dataset, which examines the pronoun-name relationship to measure the gender bias of core word parsing systems.

(4) Machine translation systems. Stanovsky et al. [82] evaluated gender bias in machine translation using a dataset based on the core word parsing system. Prates et al. [1] analyzed the gender tendencies of the Google Translate API when translating sentences into English, finding a strong male default bias in the translation output.

4.6 NLP debugging

To mitigate the fairness problem of NLP systems, numerous fairness debugging techniques have been proposed. These techniques can be categorized into three broad categories based on their implementation during various stages of training and deployment of NLP systems.

(1) For raw dataset. Counterfactual data augmentation [83] is a data-level fairness debugging technique that rebalances the corpus by swapping biased attribute words in the dataset, such as "he" versus "she" or "church". Maudslay et al. [84] proposed a counterfactual data augmentation based on name pairing with random substitution to avoid textual repetition. Dixon et al. [85] suggested a rebalancing strategy by unsupervised balancing of the training dataset to reduce the bias embedded in the dataset.

(2) For word embedding process. Bolukbasi et al. [32] proposed a hard debiasing algorithm, which targets bias and uses a mechanism similar to linear projection but extends the distance of each word pair after projection according to the distance before the operation, thus better guaranteeing the classification performance. Ravfogel et al. [86] designed the iterative nullspace projection algorithm, which follows the direction of the sensitive feature vector, finds the most extreme words at both ends, builds a linear classifier to separate these two words optimally, and projects all words to the normal of this classifier. Then, the algorithm finds the most extreme words at both ends using the normal as the new direction and iterates the above process to randomize the potential biased association between words. However, Dev et al. [87] pointed out that this approach may destroy information that needs to be preserved, such as the term "grandfather" referring to male grandparents, and proposed the orthogonal subspace correction (OSCaR) algorithm to avoid this problem. The OSCaR algorithm attempts to orthogonalize the subspace of sensitive features with other features, such as the subspace of representative gender and the subspace of representative occupation, so that the bias can be reduced while preserving the original subspace information. Deviations are calculated by subtracting the projection generated by the sensitive feature subspace from the original sentence representation to perform debiasing.

Table 2 Comparison of popular datasets in ML fairness research

Data type	Dataset name	Task scope	Data size
Tabular	Census [93]	Income prediction	48842
	German Credit [94]	Credit risk level prediction	1000
	COMPAS [95]	Re-offence prediction	7214
Image	ImageNet [96]	Image classification	∼14000 k
	CIFAR-10 [97]	Image classification	∼60 k
	CIFAR-100 [97]	Image classification	∼60 k
	MNIST [98]	Image classification	∼70 k
	Fashion-MNIST [99]	Image classification	∼70 k
	CelebA [100]	Face recognition	∼200 k
	FairFace [38]	Face recognition	∼100 k
	Adience [101]	Face recognition	∼30 k
	MS-COCO [102]	Automatic captioning	∼300 k
	VQA [103]	Visual question answering	∼300 k images, 1000 k questions
NLP	Translation template [104]	Machine translation	∼1 k
	Equity evaluation corpus [78]	Sentiment analysis	∼90 k
	Winogender [78]	Core word parsing	∼700
	Winobias [80]	Core word parsing	∼3 k
	GAP Coreference [81]	Core word parsing	∼9 k
	In-Situ [105]	Named entity recognition	∼50 k
	TIMIT [106]	Speech recognition	∼6 k

(3) For training process. Zhao et al. [89] proposed a new training procedure, GN-GloVe, that forces the model to maintain gender-leaning neutrality while learning word embedding information. Webster et al. [73] suggested using dropout regularization [92] as a bias mitigation technique. They investigated the impact of increasing the attentional mechanism weights of BERT and ALBERT and changing the dropout parameter with an additional pre-training phase. Their experiments revealed that increasing dropout regularization reduced gender bias in these models. They argue that dropout's interference with the attention mechanism within BERT and ALBERT helps prevent them from learning poor associations between words.

5 Datasets

In the field of ML, datasets play a pivotal role. This section provides an overview of the most commonly used representative datasets in the field of fairness, analyzing their characteristics in terms of sources, task types, and application domains. We categorize different types of datasets according to their respective application areas. The overview of this chapter is shown in Table 2 [38, 78, 80, 81, 93–106].

5.1 Tabular datasets

The adult income dataset, the German credit card dataset, and the COMPAS dataset are the most used and studied datasets in the algorithmic fairness literature [107]. They serve as de facto fairness benchmark datasets and set the standard for subsequent fairness research and fairness dataset production. The Adult income dataset (Adult) [93], also known as the Census dataset (Census), is the most widely used dataset in fairness classification studies. The classification task defined on this dataset is often to determine whether a person's annual income exceeds $50000 based on that person's characteristics. The dataset consists of 48842 pieces of data, each containing 15 features, of which seven are categorical features, six are numerical features, and two are binary features. Some researchers have chosen to remove these data, while others have tried to exploit the missing parts of the data through interpolation or estimation.

The German credit card dataset (German credit) [94] is a small dataset for fairness studies. This dataset is often used for risk assessment prediction, i.e., to determine whether extending credit to a person is risky. The dataset consists of 1000 data from bank account holders. Each piece of data contains 13 categorical features, 7 numerical features, and 1 binary feature. In fairness studies of this dataset, gender is usually considered as a sensitive feature. Age can also be considered as a sensitive feature after binarization into young and old.

The COMPAS dataset [95] was published by ProPublica in 2016 for predicting the risk of offenders re-offending. Researchers typically perform binary classification tasks based on this dataset to predict whether an offender will be re-arrested within two years of his first arrest. The dataset contains 7214 data, each containing 31 categorical features, 6 binary features and 14 numerical features. It is important to note that the dataset is commonly missing features, with 6395 of the data containing missing values, so the dataset needs to be cleaned according to the user's needs.

5.2 Image datasets

The fairness of image datasets has been studied on several tasks, including image classification, face recognition, automatic caption generation, and VQA.

5.2.1 *Image classification datasets*

ImageNet is one of the most influential ML datasets currently available. Many important studies on computer vision, including early breakthroughs in deep learning, were sparked by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), a competition held annually from 2010 to 2017, and was most used to prepare data for the classification task of ILSVRC 2012, a dataset with 1000 classes. Recently, researchers have identified some fairness issues in ImageNet and have proposed methods to eliminate them [108–110].

CIFAR-10 and CIFAR-100 [97] are a labeled subset of a database of 80 million tiny images, consisting of 32×32 color images. This dataset is widely used in several fairness domains including fairness classification [111–113] and robustness studies [114]. In addition, common fairness image classification datasets include MNIST [98] and Fashion-MNIST [99].

5.2.2 *Face datasets*

The CelebA dataset [100] features celebrity images from the CelebFaces dataset and contains about 200000 face images of about 10000 celebrities, complemented by annotations of landmark locations and binary features. These features, ranging from highly subjective (e.g., attractive large nose) and potentially aggressive (e.g., double chin) to more objective features (e.g., dark hair) were annotated by a professional labeling company. Typically, researchers have conducted fairness studies using gender, age, and skin color as sensitive traits [115–117].

The FairFace [38] dataset is a dataset obtained by sampling images based on different races, genders, and ages, starting from a large public image dataset (Yahoo YFCC100M). The publisher considered seven categories: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino, ensuring diversity in terms of race by sampling. Their sensitive characteristics include race, age, gender and skin color.

The Adience dataset [101] is a dataset sampled from a Flickr album, opened in order to support research on automatic age and gender recognition from face images. The dataset contains about 30000 face images of about 2000 individuals, all of which are manually tagged with age, gender, and identity. Besides, other commonly used fairness face datasets are LFW [118], UTK Face [119], IJB-A [120], PPB [2], MS-Celeb-1M [121], DiF [122], MTFL [123], RFW [124], BUPT Faces [125].

5.2.3 *Other datasets*

The MS-COCO dataset contains more than 300000 labeled images collected from Flickr [102]. Each image is labeled according to whether it contains one or more of the 91 object types proposed by the authors, and segmentation semantic information is provided to indicate the region in which the object is located in each image. The publisher also provided five human-generated captions for each image, and the annotation, segmentation, and description were all done manually. Fairness studies typically focus on the presence of bias in automatic caption generation tasks based on this dataset [34].

The VQA dataset [103], a dataset containing both real images from MS-COCO and abstract scenes with the presence of people, is published as a research benchmark for open-ended VQA. Fairness researchers typically study the presence of bias in the image and text information they associate to gender [39].

5.3 NLP datasets

5.3.1 *Machine translation datasets*

The translation template dataset [104] was released to study the problem of gender bias in machine translation and consists of about a thousand sets of comma-separated sentence templates, with the part before the comma serving as a cue for a specific gender and the part after the comma containing information about gender and sentiment/occupation. An accurate translation should correctly match the grammatical gender before and after the comma in each word required by the target language.

5.3.2 *Sentiment analysis datasets*

The equity evaluation corpus (EEC) [78] is a dataset compiled to audit the sentiment analysis system for gender and race bias and contains about 9000 sentences. Each sentence includes a word related to gender or race, such as a name associated with African-American or European-American. Gender-related words include names, nouns, and pronouns.

5.3.3 *Core word parsing datasets*

Winogender [78] is a dataset published to systematically investigate gender bias in coreference systems. This resource follows the Winograd model, with sentence templates containing an occupation (nurse), a participant (patient), and a pronoun referring to either of them. The sentence templates have been carefully designed so that pronoun parsing can be performed explicitly based on contextual information, so that an unbiased system should have similar error rates regardless of the gender proportions of the different occupations. The base tokens for each sentence have been manually verified by crowdsourcing with an accuracy rate of over 99%. Similar datasets to Winobias [80] and GAP Coreference [81] are available.

5.3.4 *Named entity recognition datasets*

The In-Situ dataset [105] is a text dataset containing about 50000 sentences. The authors first built a list of 123 typical names of men and women of different races and religions based on the gender, race, and religion of the person represented by the entity by sampling census data. Next, they extracted 289 sentences referring to people from CoNLL 2003 NER test data [126] from Reuters news reports from the 1990s. Finally, a dataset for measuring bias in the named entity recognition algorithm was generated by replacing the entities in CoNLL 2003 with previously obtained names associated with different demographic groups.

5.3.5 *Speech recognition datasets*

TIMIT [106] is a dataset used for speech recognition studies and contains about 6000 sentences from about 600 speakers. The dataset features speakers of different American English dialects and includes annotations such as time-aligned orthography, phonetic symbols, and word transcriptions. For a more comprehensive overview of the available fairness dataset, please refer to the work of Le Quy et al. [127] and Fabris et al. [107]. The former surveyed tabular datasets used for fairness studies; the latter extended the survey to unstructured data (e.g., text and images), covering datasets from various fields such as social sciences, computer vision, health, economics, business, and linguistics.

6 Future gazing

Although academic research on fairness for ML models has made considerable progress, the field is still in its nascent stage, and several significant issues pertaining to definition generality, testing, and repair performance remain unaddressed. Moreover, there is no complete technical system or tool chain that has been developed yet. In this section, first, we conclude the similarities, differences, and connections of fairness issues in different fields, and then we outline the current challenges and future research opportunities in the field of ML fairness.

6.1 Analysis

Fairness issues can stem from biased data in all three fields. Biases present in the data used for training models can lead to discriminatory outcomes and perpetuate existing inequalities. Fairness concerns often revolve around the representation of different groups. In tabular data, image data, and NLP data, underrepresented or marginalized groups may be disadvantaged due to biased or limited data samples. Besides, the model training process in all three fields can exhibit biases, both in their results and in their underlying algorithms.

Meanwhile, the differences in data types introduce unique challenges and biases specific to each field. Tabular data often allows for more straightforward interpretability, as relationships between variables can be directly observed. In contrast, image and NLP models may have complex internal representations that make it harder to understand and explain their decision-making processes.

Mitigating fairness issues can involve similar techniques across the fields, such as data augmentation, model regularization, and adversarial training. Fairness considerations can influence feature selection and engineering across all three fields. Identifying and addressing biased features is crucial in ensuring fair outcomes in tabular data, image data, and NLP data. However, the specific implementations may differ due to the unique characteristics of each field.

Overall, fairness issues in tabular data, image data, and NLP data share similarities in terms of biases and the need for equitable outcomes, but the differences in data types, interpretability, and domain-specific challenges highlight the need for context-specific approaches in addressing fairness issues in each field.

6.2 Challenges

Currently, the main challenges in the field of fairness are concentrated in the areas of definition and implementation:

(1) Integration of definitions of fairness. In Section 1, we have demonstrated that some definitions of fairness are contradictory and cannot be satisfied simultaneously. However, there are also numerous definitions of fairness that overlap [26], and reconciling them is a critical task that needs to be addressed. Furthermore, there is no consensus on whether individual fairness or group fairness should take priority, and optimizing fairness between groups may exacerbate unfairness within groups [128]. Additionally, various communities and groups have different understandings of fairness. For instance, men and women may have different perspectives on whether to include gender as a sensitive attribute [129], and addressing one form of bias may lead to another form of bias [130]. Several recent studies have investigated how users perceive mathematically defined notions of fairness [131,132]. Interestingly, these studies have found that users often exhibit a preference for the simpler concept of demographic parity. This inclination may stem from the inherent complexity of grasping more intricate definitions of fairness.

(2) Equality and equity. Fairness and equity are two related but not identical concepts. Fairness emphasizes equal treatment, while equity emphasizes equal opportunities or equal outcomes, taking into account differences among individuals or groups. Most of the proposed definitions of fairness concentrate on equality, i.e., ensuring that each individual or group receives an equal share of resources. However, some scholars argue that ensuring that each individual or group receives the resources they need, rather than exactly equal resources, is more consistent with the concept of fairness [133].

(3) Cost of fairness. Considering fairness when designing algorithms comes at a cost. The accuracy of the model usually decreases when fairness improvements are implemented, though there are exceptions (see Subsection 2.3). If discrimination is the factor that increases accuracy, then the decline in accuracy is precisely what we would expect [14]. However, this may still raise concerns among users. Additionally, imposing fairness constraints on the model necessitates computational and time costs, which have not been adequately addressed, hindering the promotion of fairness methods. To mitigate the cost associated with implementing fairness measures, it is imperative to incorporate the concurrent optimization of fairness alongside other aspects of the algorithm's performance. Numerous existing studies have delved into this particular aspect [47, 134].

(4) Legal and practical constraints. Many fairness methods require access to data-sensitive features, which may be legally restricted and, therefore, infeasible [135, 136]. Causality-based methods are anticipated to effectively tackle this challenge. Causality serves as a valuable instrument for evaluating fairness, particularly in situations involving incomplete data or selection bias within datasets [137, 138].

Additionally, fairness repair methods are often challenging to employ in situations where model accuracy is critical, such as healthcare or criminal justice scenarios, where repairing the underlying data sample may be preferable to attempting to repair the model [136, 139, 140].

(5) Fairness in artificial intelligence generated content (AIGC) large models. With the rapid development of AIGC applications represented by LLM (e.g., ChatGPT), their fairness problem has become a new dilemma to be solved today [141–143]. While previous work has focused more on decision-making systems, AIGC generates creative content that may imply new fairness constraints and problem definitions [141]. At the same time, AIGC large models are continuously evolving with human interaction, which amplifies the risk of introducing biases at different stages [144]. Moreover, the size of AIGC large models can even be up to 30 billion [145], compared with the geometric growth of traditional models, previous testing and repair algorithms may no longer be applicable from the performance point of view.

6.3 Future opportunities

This paper provides a comprehensive overview of the current state of fairness research across various subfields of ML, including NLP, representational learning, and sentiment analysis. However, there are still many areas of ML that are actively developing fair methods, some areas, such as community detection [146] and graph embedding [147], remain understudied, presenting significant opportunities for future research in fairness. More urgently, as the wave of AIGC sweeps through and triggers changes in productivity, research on how to uncover bias and ensure fairness in AIGC will be topical.

With the increasing accessibility and popularity of ML tools and cloud solutions, there is a growing demand for fairness frameworks that can be easily integrated by both professional practitioners and beginners [148–151]. However, the majority of existing fairness frameworks, including AIF360 [152], FairLearn [153] and ML Fairness Gym are targeted toward experienced practitioners, and there is a lack of mature fairness problem-solving frameworks tailored for beginners. Thus, developing beginner-friendly fairness frameworks and facilitating the integration of fairness considerations into ML workflows in various fields of society are pressing and promising research topics.

7 Conclusion

The rapid development and widespread deployment of ML technology have brought about significant social impact, attracting researchers from academia and industry to investigate the fairness and social impact of ML models. Despite the fruitful results achieved thus far, the fairness of evolving iterations of ML models has not received sufficient attention, indicating a need to improve existing testing and repair methods related to model fairness. The constantly changing ML technologies pose additional opportunities and challenges for model fairness. To re-examine the research status of fairness of ML models in different application areas, this paper systematically investigates the definition, sources, testing and debugging methods, and common datasets of fairness in three fields: tabular data, image data, and NLP data. We review a large number of representative and influential research results, and summarize the related work scientifically. However, there are still challenges and difficulties that need to be addressed, and this paper identifies these issues and proposes possible trends and directions for future research to further promote the development of fairness research for ML.

Acknowledgements This work was partially supported by National Key R&D Program of China (Grant No. 2020AAA0107702), National Natural Science Foundation of China (Grant Nos. U21B2018, 62161160337, 62132011, 62206217, 62376210, 62006181, U20B2049, U20A20177), Shaanxi Province Key Industry Innovation Program (Grant Nos. 2023-ZDLGY-38, 2021ZDLGY01-02), China Postdoctoral Science Foundation (Grant Nos. 2022M722530, 2023T160512), and Fundamental Research Funds for the Central Universities (Grant Nos. xtr052023004, xtr022019002, xzy012022082).

References

1 Prates M O R, Avelar P H, Lamb L C. Assessing gender bias in machine translation: a case study with Google Translate. Neural Comput Applic, 2020, 32: 6363–6381
2 Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of Conference on Fairness, Accountability and Transparency, 2018. 77–91
3 Locke J. Two Treatises of Government. 1689
4 Rawls J. A Theory of Justice. Cambridge: Harvard University Press, 1971
5 Sen A. The idea of justice1. J Hum Dev, 2008, 9: 331–342
6 Mehrabi N, Naveed M, Morstatter F, et al. Exacerbating algorithmic bias through fairness attacks. AAAI, 2021, 35: 8930–8938

7 Cornacchia G, Anelli V W, Biancofiore G M, et al. Auditing fairness under unawareness through counterfactual reasoning. Inf Process Manage, 2023, 60: 103224
8 Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. 2018. ArXiv:1808.00023
9 Gregory J. Sex, race and the law: legislating for equality. Fem Rev, 1988, 30: 121–122
10 Cuevas A G, Ong A D, Carvalho K, et al. Discrimination and systemic inflammation: a critical review and synthesis. Brain Behav Immun, 2020, 89: 465–479
11 Kline P, Rose E K, Walters C R. Systemic discrimination among large U.S. employers. Quart J Econ, 2022, 137: 1963–2036
12 Altonji J G, Pierret C R. Employer Learning and Statistical Discrimination. Quart J Econ, 2001, 116: 313–350
13 Kamiran F, Mansha S, Karim A, et al. Exploiting reject option in classification for social discrimination control. Inf Sci, 2018, 425: 18–33
14 Feldman M, Friedler S A, Moeller J, et al. Certifying and removing disparate impact. In: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 259–268
15 Berk R, Heidari H, Jabbari S, et al. Fairness in criminal Justice risk assessments: the state of the art. Sociological Methods Res, 2021, 50: 3–44
16 Tobriner M. California FEPC. Hastings L J, 1964, 16: 333
17 Dwork C, Hardt M, Pitassi T, et al. Fairness through awareness. In: Proceedings of the 3rd Conference on Innovations in Theoretical Computer Science, 2012. 214–226
18 Kusner M, Loftus J, Russell C, et al. Counterfactual fairness. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 4069–4079
19 Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, 2018. 1–7
20 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. 3323–3331
21 Gupta S, Kamble V. Individual fairness in hindsight. J Mach Learn Res, 2021, 22: 6386–6420
22 Ilvento C. Metric learning for individual fairness. 2020. ArXiv:1906.00250
23 Holland P W. Statistics and causal inference. J Am Stat Assoc, 1986, 81: 945–960
24 Pleiss G, Raghavan M, Wu F, et al. On fairness and calibration. 2017. ArXiv:1709.02012
25 Bechavod Y, Ligett K. Learning fair classifiers: a regularization-inspired approach. 2017. ArXiv:1707.00044
26 Friedler S, Scheidegger C, Venkatasubramanian S, et al. A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019. 329–338
27 Menon A, Williamson R. The cost of fairness in binary classification. In: Proceedings of Conference on Fairness, Accountability and Transparency, 2018. 107–118
28 Wick M, Tristan J. Unlocking fairness: a trade-off revisited. In: Proceedings of Advances in Neural Information Processing Systems, 2019
29 Pessach D, Shmueli E. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. Expert Syst Appl, 2021, 185: 115667
30 Zarya V. The share of female CEOS in the Fortune 500 dropped by 25% in 2018. 2018. https://fortune.com/2018/05/21/women-fortune-500-2018/
31 Shankar S, Halpern Y, Breck E, et al. No classification without representation: assessing geodiversity issues in open data sets for the developing world. 2017. ArXiv:1711.08536
32 Bolukbasi T, Chang K, Zou J, et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. 4356–4364
33 Choi K, Grover A, Singh T, et al. Fair generative modeling via weak supervision. In: Proceedings of International Conference on Machine Learning, 2020. 1887–1898
34 Hendricks L, Burns K, Saenko K, et al. Women also snowboard: overcoming bias in captioning models. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 771–787
35 Xu H, Liu X, Li Y, et al. To be robust or to be fair: towards fairness in adversarial training. 2020. ArXiv:2010.06121
36 Benz P, Zhang C, Karjauv A, et al. Robustness may be at odds with fairness: an empirical study on class-wise accuracy. In: Proceedings of neurIPS 2020 Workshop on Pre-registration in Machine Learning, 2021. 325–342
37 Schaaf N, Mitri G P U, Kim H, et al. Towards measuring bias in image classification. 2021. ArXiv:2107.00360
38 Kärkkäinen K, Joo J. FairFace: face attribute dataset for balanced race, gender, and age. 2019. ArXiv:1908.04913
39 Manjunatha V, Saini N, Davis L. Explicit bias discovery in visual question answering models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 9562–9571
40 Tatman R. Gender and dialect bias in YouTube's automatic captions. In: Proceedings of the 1st ACL Workshop on Ethics in Natural Language Processing, 2017. 53–59
41 Hamilton W, Leskovec J, Jurafsky D. Diachronic word embeddings reveal statistical laws of semantic change. 2016. ArXiv:1605.09096
42 Garg N, Schiebinger L, Jurafsky D, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci USA, 2018, 115: E3635–E3644
43 Biswas S, Rajan H. 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. 981–993
44 Valentim I, Lourenço N, Antunes N. The impact of data preparation on the fairness of software systems. In: Proceedings of the 30th International Symposium on Software Reliability Engineering (ISSRE), 2019. 391–401
45 Vig J, Gehrmann S, Belinkov Y, et al. Investigating gender bias in language models using causal mediation analysis. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 12388–12401
46 Zhang J, Beschastnikh I, Mechtaev S, et al. Fairness-guided SMT-based rectification of decision trees and random forests. 2020. ArXiv:2011.11001
47 Gao X, Zhai J, Ma S, et al. FairNeuron: improving deep neural network fairness with adversary games on selective neurons. In: Proceedings of the 44th International Conference on Software Engineering (ICSE), 2022. 921–933
48 Zhang P, Wang J, Sun J, et al. Fairness testing of deep image classification with adequacy metrics. 2021. ArXiv:2111.08856
49 Zheng H, Chen Z, Du T, et al. NeuronFair: interpretable white-box fairness testing through biased neuron identification,

2021. ArXiv:2112.13214
50 Angell R, Johnson B, Brun Y, et al. Themis: automatically testing software for discrimination. In: Proceedings of the 26th ACM JOINT MEETING on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2018. 871–875
51 Udeshi S, Arora P, Chattopadhyay S. Automated directed fairness testing. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, 2018. 98–108
52 Fan M, Wei W, Jin W, et al. Explanation-guided fairness testing through genetic algorithm. 2022. ArXiv:2205.08335
53 Chakraborty J, Peng K, Menzies T. Making fair ML software using trustworthy explanation. In: Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2020. 1229–1233
54 Zhang P, Wang J, Sun J, et al. Automatic fairness testing of neural classifiers through adversarial sampling. IEEE Trans Softw Eng, 2022, 48: 3593–3612
55 Zhang P, Wang J, Sun J, et al. White-box fairness testing through adversarial sampling. In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering, 2020. 949–960
56 Zhang L, Zhang Y, Zhang M. Efficient white-box fairness testing through gradient search. In: Proceedings of the 30th ACM Sigsoft International Symposium on Software Testing and Analysis, 2021. 103–114
57 Chakraborty J, Majumder S, Yu Z, et al. 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, 2020. 654–665
58 Tizpaz-Niari S, Kumar A, Tan G, et al. Fairness-aware configuration of machine learning libraries. 2022. ArXiv:2202.06196
59 Joo J, Kärkkäinen K. Gender slopes: counterfactual fairness for computer vision models by attribute manipulation. In: Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia, 2020. 1–5
60 McDuff D, Ma S, Song Y, et al. Characterizing bias in classifiers using generative models. In: Proceedings of Advances in Neural Information Processing Systems, 2019
61 Hooker S, Moorosi N, Clark G, et al. Characterising bias in compressed models. 2020. ArXiv:2010.03058
62 Xu G, Hu Q. Can model compression improve NLP fairness. 2022. ArXiv:2201.08542
63 Stoychev S, Gunes H. The effect of model compression on fairness in facial expression recognition. 2022. ArXiv:2201.01709
64 Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res, 2002, 16: 321–357
65 Han H, Wang W, Mao B. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Proceedings of International Conference on Intelligent Computing, 2005. 878–887
66 Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation. SIGKDD Explor Newsl, 2004, 6: 30–39
67 Sattigeri P, Hoffman S C, Chenthamarakshan V, et al. Fairness GAN: generating datasets with fairness properties using a generative adversarial network. IBM J Res Dev, 2019, 63: 3:1–3:9
68 Quadrianto N, Sharmanska V, Thomas O. Discovering fair representations in the data domain. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 8227–8236
69 Caliskan A, Bryson J J, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science, 2017, 356: 183–186
70 Dev S, Phillips J. Attenuating bias in word vectors. In: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019. 879–887
71 May C, Wang A, Bordia S, et al. On measuring social biases in sentence encoders. 2019. ArXiv:1903.10561
72 Kurita K, Vyas N, Pareek A, et al. Measuring bias in contextualized word representations. 2019. ArXiv:1906.07337
73 Webster K, Wang X, Tenney I, et al. Measuring and reducing gendered correlations in pre-trained models. 2020. ArXiv:2010.06032
74 Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained language models. 2020. arXiv:2004.09456
75 Nangia N, Vania C, Bhalerao R, et al. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. 2020. ArXiv:2010.00133
76 De-Arteaga M, Romanov A, Wallach H, et al. Bias in BIOS: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019. 120–128
77 Romanov A, De-Arteaga M, Wallach H, et al. What's in a name? Reducing bias in BIOS without access to protected attributes. 2019. ArXiv:1904.05233
78 Kiritchenko S, Mohammad S. Examining gender and race bias in two hundred sentiment analysis systems. 2018. ArXiv:1805.04508
79 Rudinger R, Naradowsky J, Leonard B, et al. Gender bias in coreference resolution. 2018. ArXiv:1804.09301
80 Zhao J, Wang T, Yatskar M, et al. Gender bias in coreference resolution: evaluation and debiasing methods. 2018. ArXiv:1804.06876
81 Webster K, Recasens M, Axelrod V, et al. Mind the GAP: a balanced corpus of gendered ambiguous pronouns. Trans Assoc Comput Linguist, 2018, 6: 605–617
82 Stanovsky G, Smith N, Zettlemoyer L. Evaluating gender bias in machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 1679–1684
83 Zmigrod R, Mielke S, Wallach H, et al. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. 2019. ArXiv:1906.04571
84 Maudslay R, Gonen H, Cotterell R, et al. It's all in the name: mitigating gender bias with name-based counterfactual data substitution. 2019. ArXiv:1909.00871
85 Dixon L, Li J, Sorensen J, et al. Measuring and mitigating unintended bias in text classification. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018. 67–73
86 Ravfogel S, Elazar Y, Gonen H, et al. Null it out: guarding protected attributes by iterative nullspace projection. 2020. ArXiv:2004.07667
87 Dev S, Li T, Phillips J, et al. OSCaR: orthogonal subspace correction and rectification of biases in word embeddings. 2020. ArXiv:2007.00049
88 Liang P, Li I, Zheng E, et al. Towards debiasing sentence representations. 2020. ArXiv:2007.08100
89 Zhao J, Zhou Y, Li Z, et al. Learning gender-neutral word embeddings. 2018. ArXiv:1809.01496

90 Hicks R, Tingley D. Causal mediation analysis. Stata J, 2011, 11: 605–619
91 Bansal R. A survey on bias and fairness in natural language processing. 2022. ArXiv:2204.09591
92 Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res, 2014, 15: 1929–1958
93 Amarnath B, Balamurugan S, and Alias A. Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. J Eng Sci Technol, 2016, 11: 1639–1646
94 Kambal E, Osman I, Taha M, et al. Credit scoring using data mining techniques with particular reference to Sudanese banks. In: Proceedings of International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), 2013. 378–383
95 Angwin J, Larson J, Mattu S, et al. Machine bias. In: Proceedings of Ethics of Data and Analytics, 2016. 254–264
96 Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
97 Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009. http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf
98 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324
99 Xiao H, Rasul K, Vollgraf R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017. ArXiv:1708.07747
100 Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 3730–3738
101 Eidinger E, Enbar R, Hassner T. Age and gender estimation of unfiltered faces. IEEE Trans Inform Forensic Secur, 2014, 9: 2170–2179
102 Lin T, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Proceedings of European conference on computer vision, 2014. 740–755
103 Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6904–6913
104 Cho W, Kim J, Yang J, et al. Towards cross-lingual generalization of translation gender bias. In: Proceedings of the ACM CONFERENCE on Fairness, Accountability, and Transparency, 2021. 449–457
105 Mishra S, He S, Belli L. Assessing demographic bias in named entity recognition, 2020. ArXiv:2008.03415
106 Garofolo J, Lamel L, Fisher W, et al. Darpa timit acoustic-phonetic continuous speech corpus CD-ROM. NASA STI/Recon Technical Report N, 1993, 93: 27403
107 Fabris A, Messina S, Silvello G, et al. Algorithmic fairness datasets: the story so far. Data Min Knowl Disc, 2022, 36: 2074–2152
108 Prabhu V, Birhane A. Large image datasets: a pyrrhic win for computer vision? 2020. ArXiv:2006.16923
109 Yang K, Qinami K, Fei-Fei L, et al. Towards fairer datasets: filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2020. 547–558
110 Crawford K, Paglen T. Excavating AI: the politics of images in machine learning training sets. AI Soc, 2021, 36: 1105–1116
111 Zhao B, Xiao X, Gan G, et al. Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 13208–13217
112 Wang Z, Qinami K, Karakozis I, et al. Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 8919–8928
113 Jung S, Lee D, Park T, et al. Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 12115–12124
114 Nanda V, Dooley S, Singla S, et al. Fairness through robustness: investigating robustness disparity in deep learning. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2021. 466–477
115 Kim B, Kim H, Kim K, et al. Learning not to learn: training deep neural networks with biased data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 9012–9020
116 Zhang H, Davidson I. Towards fair deep anomaly detection. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 2021. 138–148
117 Amini A, Soleimany A, Schwarting W, et al. Uncovering and mitigating algorithmic bias through learned latent structure. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2019. 289–295
118 Huang G, Mattar M, Berg T, et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In: Proceedings of Workshop on Faces in 'Real-life' Images: Detection, Alignment, and Recognition, 2008
119 Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5810–5818
120 Klare B, Klein B, Taborsky E, et al. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1931–1939
121 Guo Y, Zhang L, Hu Y, et al. MS-Celeb-1M: a dataset and benchmark for large-scale face recognition. In: Proceedings of European Conference on Computer Vision, 2016. 87–102
122 Merler M, Ratha N, Feris R, et al. Diversity in faces. 2019. ArXiv:1901.10436
123 Zhang Z, Luo P, Loy C, et al. Facial landmark detection by deep multi-task learning. In: Proceedings of European Conference on Computer Vision, 2014. 94–108
124 Wang M, Deng W, Hu J, et al. Racial faces in the wild: reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 692–702
125 Wang M, Deng W. Mitigating bias in face recognition using skewness-aware reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 9322–9331
126 Sang E, de Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. 2003. ArXiv:cs/0306050
127 Le Quy T, Roy A, Iosifidis V, et al. A survey on datasets for fairness-aware machine learning. WIREs Data Min Knowl, 2022, 12: e1452
128 Speicher T, Heidari H, Grgic-Hlaca N, et al. A unified approach to quantifying algorithmic unfairness: measuring individual &group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018. 2239–2248
129 Pierson E. Gender differences in beliefs about algorithmic fairness. 2017. ArXiv:1712.09124
130 Kallus N, Zhou A. Residual unfairness in fair machine learning from prejudiced data. In: Proceedings of International

Conference on Machine Learning, 2018. 2439–2448
131 Grgic-Hlaca N, Redmiles E M, Gummadi K P, et al. Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: Proceedings of the World Wide Web Conference, 2018. 903–912
132 Srivastava M, Heidari H, Krause A. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019. 2459–2468
133 Johnson J M. Race and social equity: a nervous area of government. EQual Diversity Inclusion-An Int J, 2015, 34: 262–264
134 Li T, Xie X, Wang J, et al. Faire: repairing fairness of neural networks via neuron condition synthesis. ACM Trans Softw Eng Methodol, 2024, 33: 1–24
135 Agarwal A, Beygelzimer A, Dudik M, et al. A reductions approach to fair classification. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 60–69
136 Chen I, Johansson F, Sontag D. Why is my classifier discriminatory? In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018
137 Bareinboim E, Pearl J. Causal inference and the data-fusion problem. Proc Natl Acad Sci USA, 2016, 113: 7345–7352
138 Loftus J R, Russell C, Kusner M J, et al. Causal reasoning for algorithmic fairness. 2018. ArXiv:1805.05859
139 Ensign D, Friedler S, Neville S, et al. Runaway feedback loops in predictive policing. In: Proceedings of Conference on Fairness, Accountability and Transparency, 2018. 160–171
140 Liu L, Dean S, Rolf E, et al. Delayed impact of fair machine learning. In: Proceedings of International Conference on Machine Learning, 2018. 3150–3158
141 Li Y and Zhang Y. Fairness of ChatGPT. 2023. ArXiv:2305.18569
142 Zhang J, Bao K, Zhang Y, et al. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. 2023. ArXiv:2305.07609
143 Yaraghi N. ChatGPT and health care: implications for interoperability and fairness. Health Affairs Forefront, 2023. https://www.brookings.edu/articles/chatgpt-and-health-care-implications-for-interoperability-and-fairness/
144 Pedro R, Castro D, Carreira P, et al. From prompt injections to SQL injection attacks: how protected is your LLM-integrated web application? 2023. ArXiv:2308.01990
145 Rozière B, Gehring J, Gloeckle F, et al. Code LLAMA: open foundation models for code. 2023. ArXiv:2308.12950
146 Fortunato S. Community detection in graphs. Phys Rep, 2010, 486: 75–174
147 Cai H, Zheng V W, Chang K C C. A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans Knowl Data Eng, 2018, 30: 1616–1637
148 Hall M, Frank E, Holmes G, et al. The WEKA data mining software. SIGKDD Explor Newsl, 2009, 11: 10–18
149 Bisong E. Google autoML: cloud vision. In: Proceedings of Building Machine Learning and Deep Learning Models on Google Cloud Platform, 2019. 581–598
150 Feurer M, Klein A, Eggensperger K, et al. Efficient and robust automated machine learning. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015. 2755–2763
151 Thornton C, Hutter F, Hoos H, et al. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013. 847–855
152 Bellamy R K E, Dey K, Hind M, et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev, 2019, 63: 4:1–4:15
153 Bird S, Dudík M, Edgar R, et al. FairLearn: A Toolkit for Assessing and Improving Fairness in AI. Microsoft Technical Report MSR-TR-2020-32, 2020