Multi-stage Bias Mitigation for Individual Fairness in Algorithmic Decisions

Adinath Ghadage[1], Dewei Yi[1], George Coghill[1], and Wei Pang[2]

[1] University of Aberdeen, Aberdeen, United Kingdom
[2] Heriot-Watt University, Edinburgh, United Kingdom

Abstract. The widespread use of machine learning algorithms in data-driven decision-making systems has become increasingly popular. Recent studies have raised concerns that this increasing popularity has exacerbated issues of unfairness and discrimination toward individuals. Researchers in this field have proposed a wide variety of fairness-enhanced classifiers and fairness matrices to address these issues, but very few fairness techniques have been translated into the real-world practice of data-driven decisions. This work focuses on individual fairness, where similar individuals need to be treated similarly based on the similarity of tasks. In this paper, we propose a novel model of individual fairness that transforms features into high-level representations that conform to the individual fairness and accuracy of the learning algorithms. The proposed model produces equally deserving pairs of individuals who are distinguished from other pairs in the records by data-driven similarity measures between each individual in the transformed data. Such a design identifies the bias and mitigates it at the data preprocessing stage of the machine learning pipeline to ensure individual fairness. Our method is evaluated on three real-world datasets to demonstrate its effectiveness: the credit card approval dataset, the adult census dataset, and the recidivism dataset.

Keywords: Algorithmic bias · Algorithmic fairness · Fairness-aware machine learning · fairness in machine learning · Individual fairness.

1 Introduction

With the widespread use of machine learning algorithms in decision-making systems, concerns about trust in AI are growing in terms of its full adaptation. The decisions in many decision-making systems are based upon the predictions of the results of machine learning algorithms. The major challenge for policy-makers, stakeholders, and companies in adopting AI is the black box nature of AI-based decision-making systems [31]. Recently, some studies [23, 34] attempted to open the black box of AI. The researchers have proposed new fundamental pillars for trust in the AI system, including explainability, fairness, robustness, and lineage[4, 26, 3].

Fairness is one of the fundamental principles of a trustworthy AI system [32]. Systematically understanding bias specifically against each individual as well as

group members in the dataset defined by their protected attributes like age, gender, race, and nationality is the first step to achieving fairness in decision-making and building trust in AI in general. The protected attributes, concerned in this paper follow the fairness guidelines given by the Information Commissioner's Office (ICO) in terms of Equality Law 2010 in the UK. Among these guidelines, ICO proposes potentially protected attributes including age, disability, gender, marital status, maternity, race, religion, sex, and sexual orientation. According to a recent study, machine learning algorithms treat people or groups of people who have the above-mentioned protected attributes unfairly. Our research attempts to identify potential research gaps between existing fairness approaches and possible techniques to address the fair classification of machine learning algorithmic decision-making systems to contribute to building trust in AI. Many studies have been done on group fairness (which is also called statistical parity). This family of definitions fixes a small number of protected groups, such as gender, race, and then approximate parity of some statistical measure across all of these groups. Some of the popular measures include the false positive rate and the false negative rate, which are also known as equalise odd and equality of opportunity [13, 19, 8, 9, 21, 18, 33], respectively, but fewer are concerned with individual fairness. The group fairness approaches (e.g. statistical parity, equal opportunity, and disparate mistreatment) are used to investigate discrimination against members of protected attributes such as age, gender, and race. A fair classifier tries to achieve equality across the protected groups like statistical parity [13], equalised false positive and false negative rates, and calibrations [6].

Most of the group fairness definitions are subjected to learning statistical constraints [10] or averaged over the protected groups to satisfy fairness definitions [10, 16]. As group fairness is measured by aggregating over male or female or any other protected attribute, this constraint-based definition may harm some of the individuals within that group. Individual fairness is an alternative approach that satisfies the constraints for specific pairs of individuals defined by their task similarity. The notion of individual fairness is defined by *"similar individuals should be treated similarly."* [13]. Here, similarity is defined in terms of task-specific similarity metrics, where a classifier maps individuals to the probability distribution of outcomes. For example, if x_i and x_j are similar to each other, then their classification predictions y_i and y_j need to be the same.

In this paper, we address the individual notion of fairness based on the work done in [13, 22, 34, 1], in the sense that we attempt to understand more fundamental questions about how an individual is classified as fair/unfair in task-specific similarity. A model is identified to be fair to individuals if similar pairs of individuals yield similar outcomes in prediction. The similarity of individuals is determined by the closeness of distance between the data points in the input space, which satisfies the Lipschitz property (i.e., distance preservation). The model is unfair to individuals if similar individuals are treated discriminatorily in their predictions. That is, for two similar individuals $< x_i, x_j >$, their classification predictions are different, that is, $y_i \neq y_j$. To estimate a model's individual unfairness, we can use a pool of similar individuals generated by a human spec-

ified process from the original data and/or from the transformed data and their discrimination in treatment among these individuals. In our proposed approach, we revisit the notion of individual fairness proposed by Dwork *et al* [13], that is, similar individuals are treated similarly on the same given task. We learn to generalise a representation of the original data into a transformed representation. The transformed representation learnt by our model preserves fairness awareness similarity among data points with multiple protected attributes considered rather than the single protected attribute used in much of the previous individual fairness research work [30]. In this research, the words "sensitive" and "protected" are used interchangeably for the same purpose to specify the list of possible attributes based on the individuals who can be treated discriminatorily in their predictions.

Furthermore, in our work, we aim to identify and mitigate the bias presented in the data by historical decisions. The pre-processing approach is enforced to reduce discrimination and make the model fairer to individuals. We applied pre-processing techniques on transformed data to identify and remove biased data points. The process of removing biased data in our method is to modify those outcome labels of similar pairs of individuals, where such pairs contribute more to the model's unfairness and leave all other data points and features unchanged. Modifying the outcome values of similar pairs yields a less biased dataset. A model is then trained on these less biased samples and produces a fairer outcome with less individual discrimination than the model trained on the original data. Our fairness model offers more adaptability and versatility to data with multiple sensitive attributes. The experiments performed on three real-world datasets with only individual fairness definitions are considered, excluding the group fairness definitions. More specifically, the contributions of this paper are summarised as follows:

- We propose a novel approach to improve individual fairness. To the best of our knowledge, this is the first attempt to provide individual fairness by considering multiple sensitive attributes to identify and mitigate biases in the dataset.
- We develop an application-agnostic feature transformation approach by learning transformed representations of data points that restore individually fair data and accuracy such that application-specific multi-valued multiple sensitive attributes are considered, rather than single binary protected attributes such as gender.
- Our method can identify and mitigate bias at the pre-processing stage of the machine learning pipeline.
- Our method is evaluated on classification and regression tasks, showing that strong individual fairness can indeed reconcile with a high utility on real-world datasets: the adult census, credit card approval dataset and recidivism dataset.

2 Background and Related Work

2.1 Statistical Definitions of Fairness

Most of the research on fairness attempts to deal with two missions: 1) developing methods to detect bias and discrimination in AI-based decision-making systems and 2) developing methods to mitigate these biases by using different criteria to improve fairness in AI-based systems. There are a wide variety of bias metrics and fairness definitions proposed in the literature [31] [17][10][23][24][14]. The group fairness [13] is a constraint-based approximation of parity across all groups with a statistical measure. Suppose a and a' are the classes of protected and unprotected groups, and the group fairness constraint to satisfy equal probability of prediction across these two groups is defined as $P[Y|A = a] = P[Y|A = a']$. Here, bias metrics quantify system error in the context of fairness and bias systematically, which provides advantages to privileged groups over disadvantages to unprivileged groups. Bias mitigation algorithms reduce unwanted bias in the data. Conditional parity [11] [15][25] [31] and the inequality indices [29][5][28] are the most commonly used statistical measures of fairness, where this definition satisfies the equal prediction of outcome in both protected and unprotected groups controlled by legitimate factors. However, it is impossible to implement all these definitions in practice as there is no guideline on which bias metrics and bias mitigation algorithms should be used to address any specific definition of fairness. Therefore, despite recent awareness of bias and fairness issues in AI development and deployment, there is no systematic operation in practice.

2.2 Definitions of Individual Fairness

Alternatively, the individual notion of fairness is a constraint over pairs of individuals rather than an average over group members. Individual fairness ensures that members who are similar are treated similarly.Here, the similarity is a task-specific similarity metric that must determine the basis of this notion of definitions. Assume that metrics-based fairness defines similarity on variables (i.e., input feature vectors) as follows: $m : V \times V \rightarrow R \quad m$ is a map function which maps each individual to distribution of outcomes. Hence, metric fairness for individuals with variable $v, v' \in V$ is a closeness in their decisions.

$$|f(v) - f(v')| \leq m(v, v') \tag{1}$$

This formulation is based on Lipschitz condition [13] [27].

The definitions of metrics in individual fairness change subsequently, authors in [13] define the task-specific similarity metrics over *individuals*. In the following research, metrics are defined over features, variables, and inputs to classifiers [27]. A construction space was introduced in [16] in addition to the observed space (OS) and decision space (DS) [22]. A construction space (CS) is a metric space consisting of individuals and their distances. Whereas, Observed space (OS) is a metric space which approximates metrics in CS with respect to

task, assuming $g : v \rightarrow v'$ that generates an entity $v' = g(v)$ from a person $v \in CS$.

Another notion of individual fairness is proposed in [27] where the authors approximate the metrics of fairness by marginalising a small probability of error in the similarity between two individuals. This is an extension of metrics fairness defined by Dwork et al. [13]. In this matrix approximate fairness [27] definition two constants α, γ are used to approximate in addition to the similarity metrics definition suggested by Dwork [13]. In AI Fairness360 [4], it implements individual fairness mapping as the author proposed methods that measure similar prediction of a given instance to its nearest neighbours [34]. Similarly, in average individual fairness, [20] method is inspired by oracle-efficient algorithms.

3 Multi-Stage Individual Fairness

In this section, we explain the whole framework of our proposed method to investigate unfairness in machine learning framework. After having a comprehensive literature review, our research contributions is mentioned in Section 1. To address these research contributions, several novel techniques are proposed in this research. The rest of this section describes the details of each stage and its responsibilities.

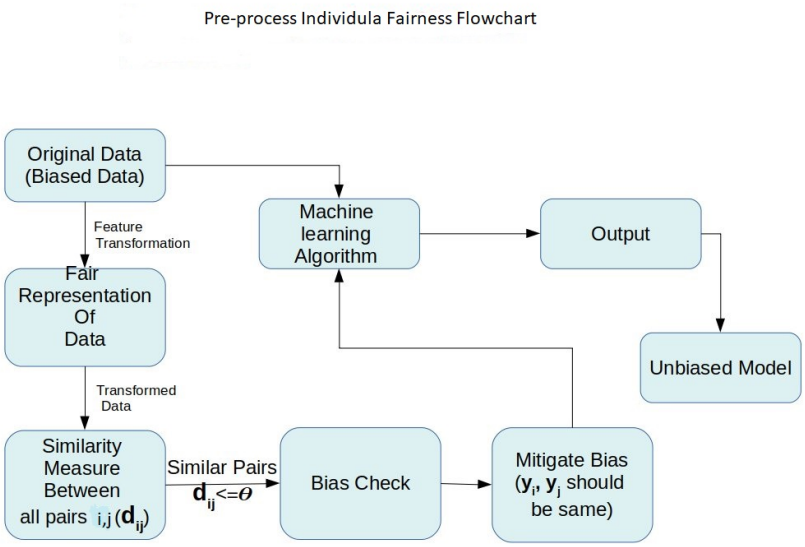

Fig. 1. Flowchart with step in our individual fairness model.

We consider that a fair algorithm should consider both protected and non-protected attributes while making fair decisions, and existing approaches give less attention to the multiple protected attributes in both group and individual fairness. Our work will be a further step towards individual fairness [13], fair [22] and perfect metrics fairness [27] with a focus on multiple protected attributes. These methods measure similarity by Lipschits mapping between a pair of individuals in unknown distributions over the classified outcomes with distance metrics. Perfect metrics fairness [27] is a generalised approach that approximates individuals' lying in both group and individual fairness. Here, Lipschits' mapping conditions map task-specific similarity metrics. Our proposed framework for individual fairness consists of detecting and mitigating bias at the preprocessing stage of the machine learning pipeline. The term "IndFair" is referred to as our defined approach to individual fairness in the paper. The proposed framework consists of five major components in the machine learning fairness pipeline. These components are as follows: fair representation of data, similarity measure, bias identification, mitigating identified bias, and the final unbiased model output. In the first component, a fair representation of features is computed using the transformation function detailed in this section. In the second component, we measure the similarity between a pair of individuals by using the Euclidean distance function. We identify a bias in the similarity measured data in the third component. Fairness measures and bias mitigation are performed in the preprocessing stage of the machine learning pipeline. A detailed description of the working principle behind each component is given in the rest of this section.

3.1 Notations

We define the notations used in the proposed model in which the input data X is represented as a $m \times n$ matrix where each individual in the populations as $Xi = 1, 2, 3, ..., n$ and m is the number of features. Each person x_i has m features (i.e. variables, input) $x_i \in X$, where features m is a combination of protected features p and non-protected features np. We assume the attributes $1...l$ protected attributes and the attributes $l+1...m$ are non-protected. A binary classification decision on each person is denoted as $\hat{Y} = f(X, Y)$, where f is a function of variables known at decision time $f : X \longrightarrow 0, 1$. In binary prediction based system a outcome variable (i.e. predictor) $\hat{Y}$ for each person is unknown at decision time and the actual outcome is denoted by $Y = (y_1, y_2, ..., y_n)$.

3.2 Transformed Representation Learning

We aim to transform features into fair representation by matrix multiplication in the transform stage. Here, the intuition of matrix multiplication is vectors of protected and non-protected attributes are multiplied by distributive property. Each person x_i has m features (i.e. variables, input) $x_i \in X$, where features m is a combination of protected features p and non-protected features np. We assume the attributes $1...l$ protected attributes and the attributes $l + 1...m$ are

non-protected. Each x_p is the vector representation of the protected attributes for individual i, similarly x_{np} is vector of non-protected attributes. The result of this transformation can be viewed as a low-level representation of individual i with $k = m - l - 2$ dimensions vector size of attributes. We perform the above operation for all data points. The mapping of $x_i \rightarrow \tilde{x_{i_k}}$ is given below:

$$\tilde{x_{i_k}} = \sum_{p=1}^{l} x_p(\sum_{np=l+1}^{m} x_{np}) \tag{2}$$

3.3 Similarity Measure

The similarity measure is an important component to achieve individual fairness in algorithmic decision making in the pre-processing stage. Similarity measure can be achieved through a distance measure between two individual records in the distribution of feature space. Mostly used distance functions are Euclidean distance, Manhattan distance, and Minkowski distance. In this paper, we focus on the euclidean distance function to measure the similarity between all pairs of individuals.

$$d(x_i, x_j) = \sqrt{\sum_{i,j=1}^{n} (x_i - x_j)^2} \tag{3}$$

The above distance functions d presented in Eq. (3) is applicable to original records x_i and transformed records $\tilde{x_{i_k}}$to measure the similarity among all pair of records.

3.4 Fairness Measure

Mapping individual bias: The individual fairness is to preserve the fairness-aware distances between a pair of individuals i and j in a given matrix space. The mapping of individual bias in the given matrix space is measured as the *consistency* of outcome between the pairwise similar individuals in transformed data and original data. The matrices used to measure the individual bias captures the intuition of individual similarity definition, that is, the similar individuals should be treated similarly. For instance, if two individual records x_i and x_j are similar, then we check the consistency of the output variable y_i and $\tilde{y_i}$. We adapted the bias mapping metrics defined in iFair [22] and [34], where our formula is different from the one used in iFair. The distance d is on a paired records x_i, x_j of input data X, and the transformed data $\tilde{X}$ with each pair of the records $\tilde{x_i}, \tilde{x_j}$ whereas $\tilde{d}$ is pairwise distance on transformed records. The mapping of bias in input data and transformed data is performed using fairness loss F_{loss} is as given below,

$$F_{loss}(X, \tilde{X}) = 1 - \sum_{i,j=1}^{n} (d(x_i, x_j) - \tilde{d}(\tilde{x_i} - \tilde{x_j}), \tag{4}$$

Mitigating Bias: The bias mapping seeks to identify any unfair distortion in the original data, transformed data and output data for mitigating the bias in the pre-processing stage. The equation 4 a fairness loss that is, F_{loss} which addresses a *systematic or structural bias* presented in the data. The intuition to mitigate bias in fair individual classifications is based on the similarity of the outcome of two individuals i and j. If the similarity distance function $d(x_i, x_j)$ indicates that individual records, i and j are similar in transformed data then their outcome value y_i and y_j should be similar. If these individuals outcome is not similar then we modify their outcome value to mitigate bias using the below formula.

$$Y_i, Y_j = \{\, 1 \,, \frac{1}{n}(y_i, y_j = 1) \geq 0.5; 0, otherwise \tag{5}$$

In the above, y_i, y_j are the mitigated values for similar individuals. Here, the intuition is for similar pair of individuals we take an average value of binary outcome either 0 or 1. If this average values of outcome for individuals holding positive outcome e.g., $\frac{1}{n}y_i, y_j = 1 \geq 0.5$ then the outcome of all these similar individuals is replaced as 1 and 0 otherwise. We used the same bias mitigation strategy in the $pre-processing$ stage of fairness pipeline defined in [4].

3.5 Optimisation

Utility: The utility measured as accuracy (Acc) for each of the support vector machine classifiers(SVM) and logistic regression(LR) for the classification task. We adapted the utility measure from iFair [22], as given below:

$$Utility(Y, \tilde{Y}) = \sum_{i=1}^{n} (y_i - \tilde{y_i})^2 \tag{6}$$

Overall Objective Function: Combining the data loss i.e., utility 6 and the fairness loss 4 yields our final objective function, the learned representation is to minimise the objective function. The utility minimises data loss and the individual fairness is fairness-aware treatment in input and output space.:

$$L_{total}(\Theta) = Utility(Y, \tilde{Y}) + F_{loss}(X, \tilde{X}) \tag{7}$$

4 Data And Experiment

4.1 Datasets

We evaluate our method on three real-world datasets, which are publicly available for researcher use. In this experiment, we examine the structural bias in the three different datasets including Adult data[7], Recidivism data[2] and Credit approval dataset [12] for fair individual classification.

- Adult dataset is a census income data in the US [7] which consists of 48,842 records. In this dataset, a target variable Y is individual income with more than 50 thousand dollars and the protected attributes that we used in our framework are *age, marital status, sex and race* for the binary classification task.
- Credit Approval dataset is a collection of 690 records in credit approval application[12]. The binary outcome value Y represents if the individual is default or not. The protected attributes that we used in our experiments are *sex, age, married, ethnicity and citizen.*
- Recidivism dataset is a widely used test case in experiments on fairness algorithms. We have used 11 attributes including 3 protected attributes, namely *age, sex and race*. The target variable two years of recidivism is the binary outcome value.

The data are randomly split into three parts: train set, test set and validation set to learn model parameters and the validation set is used for validation. We use the same setting in our experiments and compare all methods. We train and evaluate the data by using a support vector machine and logistic regression as classifiers and our individual fairness method (IndFair) and learning fair representation[34]. In our setting, the data are used to compute the accuracy and fairness at the pre-processing stage by using the above-mentioned machine learning algorithms and individual fairness definitions. We have only considered support vector machine classifier(SVM) and logistic regression(LR) in our experiments and experiments using neural classifiers will be considered in the future work. These setting of data are given in the Table 1 with attribute name *Method* which contains original data, pre-processed, post-processed, and optimal. The data setup is described in the following section.

Data Setup

- The original data consists of all the attributes including the protected attributes and the non-protected attributes.
- The result of transformed data are *Pre-processed* data given in the Table 1. The data is a transformed representation of original data which preserves the fairness-aware distance between pairs of individuals learned by applying transformed representation learning. We then check the accuracy and fairness of the data.

4.2 Evaluation Measures

- **Utility**: This is measured as accuracy (Acc) on tested classifiers, including the support vector machine (SVM) and logistic regression (LR) where these classifiers work on the binary classification tasks on three different setups of the data as mentioned above.
- **Individual Fairness:** The individual fairness, that is *IndFair*, is measured by the consistency of outcome for the individually fair pairs. This means if

the pair of records are similar to each other based on the fact that the distance value is less than the given threshold, then the predicted classification of similarly paired individuals should be the same. We categorise the similar pair individuals into three parts based on their output value y for all the individual records below the threshold distance. The first part of individuals have both positive outcome that is $y_i, y_j = [1, 1]$, in the second part both individuals have a outcome values $y_i, y_j = [0, 0]$, and the third part of individuals have either positive or negative but do not have same outcome value for each pair of individuals such that $y_i, y_j = [1, 0] or [0, 1]$. The mapping of individual fairness is given as follows:

$$IndFair(Y) = 1 - \sum_{i,j=1}^{n} (y_i, y_j) - (\tilde{y_i}\tilde{y_j}) \tag{8}$$

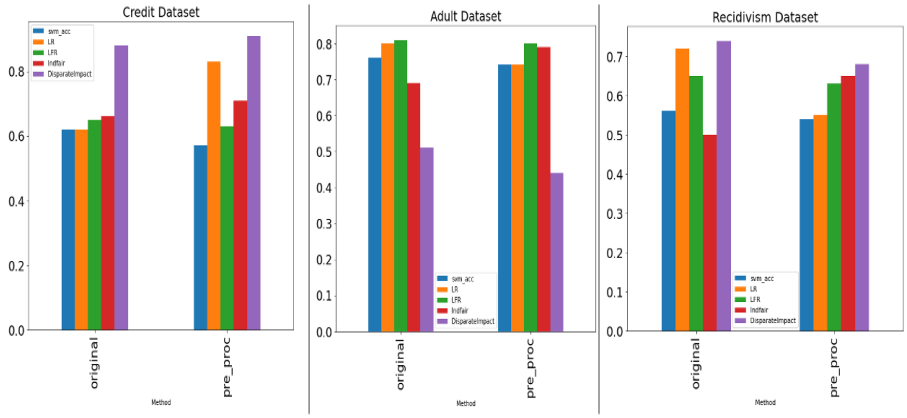

Fig. 2. Experiment result on three datasets with original and pre-processed data setting

4.3 Experimental Results

In this section, we demonstrate the effectiveness of our proposed model on the classification task. The Fig. 2 shows the results for all the methods on three datasets, plotting the accuracy and individual fairness. We observe that there is a considerable amount of unfairness in the original data, therefore the IndFair accuracy is lower in all the datasets. However, the IndFair is significantly increased in pre-processed data. The overall performance of IndFair is better than representation learned by learning fair representation (LFR) in terms of identifying bias and improving the individual fairness in the pre-processing stage of the machine learning pipeline. Table 1 shows the results of the accuracy and fairness trade-off of the machine learning classifier at three categories of data distribution as well as their fairness measure using IndFair and LFR. The optimal results are the harmonic mean of the results in all methods in the datasets setup. The pre-processed method shows an improvement in fairness; however, there is still a considerable size of unfairness hidden in the data.

Table 1. Experimental results for classification and individual Fairness task.

Dataset	Method	SVM ACC.	LR Acc.	IndFair	LFR	DisparateImpact
Adult Data	Original Data	0.76	0.80	0.69	0.81	0.51
	Pre-processed	0.74	0.74	0.79	0.80	0.44
	Post-processed	0.74	0.74	0.76	0.80	0.65
	Optimal	0.74	0.75	0.74	0.80	0.53
Credit Data	Original Data	0.62	0.62	0.66	0.65	0.88
	Pre-processed	0.57	0.83	0.71	0.63	0.91
	Post-processed	0.60	0.81	0.69	0.63	0.87
	Optimal	0.59	0.74	0.68	0.63	0.88
Recidivism Data	Original Data	0.56	0.72	0.50	0.65	0.74
	Pre-processed	0.54	0.55	0.65	0.63	0.68
	Post-processed	0.54	0.55	0.63	0.63	0.62
	Optimal	0.54	0.59	0.58	0.63	0.68

5 Conclusion

In this paper, we propose a generic and flexible method to achieve better individual fairness. It is framework to perform a transformation of data into individually fair representations. Our method accommodates two important criteria. First, we view fairness from an application-agnostic prospect, which allows us to incorporate it in a wide variety of tasks, including general classifiers. Second, we consider multiple protected attributes along with the non-protected attributes to facilitate fair treatments of individuals through transformed representation

of data. Our proposed model is evaluated on three real-world datasets including Adult income data, Credit data and Recidivism dataset, which demonstrates that the consistency with utility and individual fairness can reach a promising degree by using our model. With applying the representations of our individual fair model on classifier, it leads that algorithmic decisions through our approach are substantially more fairer than the decisions made on the original data.

References

1. Ahn, Y., Lin, Y.R.: Fairsight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics **26**(1), 1086–1095 (2019)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias risk assessments in criminal sentencing. ProPublica, May **23** (2016)
3. Arnold, M., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K.N., Olteanu, A., Piorkowski, D., et al.: Factsheets: Increasing trust in ai services through supplier's declarations of conformity. IBM Journal of Research and Development **63**(4/5), 6–1 (2019)
4. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018)
5. Bellù, L.G., Liberati, P.: Inequality analysis: The gini index. FAO, EASYPol Module **40** (2006)
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research p. 0049124118782533 (2018)
7. Blake, C.: Cj merz uci repository of machine learning databases. University of California at Irvine (1998)
8. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery **21**(2), 277–292 (2010)
9. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
10. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810 (2018)
11. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 797–806. ACM (2017)
12. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
13. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226. ACM (2012)
14. Dwork, C., Ilvento, C.: Fairness under composition. arXiv preprint arXiv:1806.06122 (2018)
15. Dwork, C., Ilvento, C.: Group fairness under composition (2018)
16. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236 (2016)

17. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017)
18. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29**, 3315–3323 (2016)
19. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: 2011 IEEE 11th International Conference on Data Mining Workshops. pp. 643–650. IEEE (2011)
20. Kearns, M., Roth, A., Sharifi-Malvajerdi, S.: Average individual fairness: Algorithms, generalization and experiments. arXiv preprint arXiv:1905.10607 (2019)
21. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
22. Lahoti, P., Gummadi, K.P., Weikum, G.: ifair: Learning individually fair data representations for algorithmic decision making. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). pp. 1334–1345. IEEE (2019)
23. Mitchell, S., Potash, E., Barocas, S.: Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:1811.07867 (2018)
24. Narayanan, A.: Translation tutorial: 21 fairness definitions and their politics. In: Proc. Conf. Fairness Accountability Transp., New York, USA (2018)
25. Ritov, Y., Sun, Y., Zhao, R.: On conditional parity as a notion of non-discrimination in machine learning. arXiv preprint arXiv:1706.08519 (2017)
26. Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. Ieee Access **8**, 42200–42216 (2020)
27. Rothblum, G.N., Yona, G.: Probably approximately metric-fair learning. arXiv preprint arXiv:1803.03242 (2018)
28. Shorrocks, A.F.: Inequality decomposition by population subgroups. Econometrica: Journal of the Econometric Society pp. 1369–1385 (1984)
29. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: 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. pp. 2239–2248. ACM (2018)
30. Verma, S., Ernst, M., Just, R.: Removing biased data to improve fairness and accuracy. arXiv preprint arXiv:2102.03054 (2021)
31. Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). pp. 1–7. IEEE (2018)
32. Wing, J.M.: Trustworthy ai. Communications of the ACM **64**(10), 64–71 (2021)
33. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. International World Wide Web Conferences Steering Committee (2017)
34. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)