BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS

Pranay K. Lohia,^{*,1} *Karthikeyan Natesan Ramamurthy,*^{*,2} *Manish Bhide,*^{†,3}
Diptikalyan Saha,^{*,1} *Kush R. Varshney,*^{*,2} *and Ruchir Puri*^{†,2}

*IBM Research and †IBM Watson AI Platform
[1]Outer Ring Road, Embassy Manyatha B, Rachenahalli & Nagawara Villages, Bangalore, KA, India
[2]1101 Kitchawan Road, Yorktown Heights, NY, USA
[3]Mindspace 3A, Hi-Tech City, Madhapur, Hyderabad, TG, India

ABSTRACT

Whereas previous post-processing approaches for increasing the fairness of predictions of biased classifiers address only group fairness, we propose a method for increasing both individual and group fairness. Our novel framework includes an individual bias detector used to prioritize data samples in a bias mitigation algorithm aiming to improve the group fairness measure of disparate impact. We show superior performance to previous work in the combination of classification accuracy, individual fairness and group fairness on several real-world datasets in applications such as credit, employment, and criminal justice.

Index Terms— Classification, discrimination, algorithmic fairness, signal detection

1. INTRODUCTION

Fairness, non-discrimination, and unwanted bias have always been concerns in human decision making [1], but are increasingly in the limelight because historical human decisions are now being used as training data for machine learning models in high stakes applications such as employment, credit, and criminal justice [2]. Without bias mitigation, models trained on such decisions perpetuate and scale human biases and are thereby unsafe and untrustworthy [3, 4]. The last couple of years have seen a surge in papers on algorithmic fairness in the machine learning and data mining literature, with basic principles defined using detection, estimation theory and information theory [5, 6].

There are two main notions of fairness in decision making: *group fairness* and *individual fairness*. Group fairness, in its broadest sense, partitions a population into groups defined by *protected attributes* (such as gender, caste, or religion) and seeks for some statistical measure to be equal across groups. There are many different group fairness notions involving different statistical measures, one such notion being *disparate impact* [7]. Individual fairness, in its broadest sense, seeks for similar individuals to be treated similarly. Checking for group fairness is a fairly straightforward computation of statistical metrics [8], but checking for individual fairness is more computationally involved when there are many protected attributes with many values and scoring samples using a model is expensive [9, 10]. Unified metrics for both group and individual fairness have recently been proposed [11] based on inequality indices [12].

Machine learning pipelines contain three possible points of intervention to mitigate unwanted bias: the training data, the learning procedure, and the output predictions, with three corresponding classes of bias mitigation algorithms: pre-processing, in-processing, and post-processing [13]. Advantages of post-processing algorithms are that they do not require access to the training process and are thus suitable for run-time environments. Moreover, post-processing algorithms operate in a black-box fashion, meaning that they do not need access to the internals of models, their derivatives, etc., and are therefore applicable to *any* machine learning model (or amalgamation of models) [14].

The vast majority of bias mitigation algorithms address group fairness, but a few address individual fairness [15, 16]. Some pre-processing algorithms address both group and individual fairness [17, 18, 6], but to the best of our knowledge, all existing post-processing algorithms are only for group fairness [14, 19, 20, 21]. Our main contribution in this paper is to propose a post-processing bias mitigation algorithm that considers *both* group and individual fairness. Moreover, unlike the previous work, our proposal does not require any ground truth class labels in the validation samples while training the bias mitigation algorithm.

The general methodology of post-processing algorithms is to take a subset of samples and change their predicted labels appropriately to meet a group fairness requirement. An interesting observation about post-processing is that *any* sample can be altered to achieve group fairness requirements because the metrics are expectations. The papers [19, 20] choose the samples randomly, whereas [14] chooses the most uncertain samples (the ones in the reject option band [22, 23]), capturing the human intuition to give the benefit of the doubt to unprivileged groups. In the method we propose herein, we choose samples that have or are likely to have individual fairness issues and in this way are able to address both group and individual fairness together.

The starting point for our proposed approach is the individual bias detector of [10], which finds samples whose model prediction changes when the protected attributes change, leaving all other features constant. Despite a large set of efficiencies enacted in the algorithm, it is still computationally expensive. To overcome the limitation of not being able to run the detector continually, we check for individual fairness on a small set of points and generalize from them by training a classifier that is applied to new samples. The samples with likely individual bias are the ones considered for a change of predicted label. By doing so, we modify the idea of [14] from focusing on uncertainty to focusing on individual bias.

Our empirical results are promising. Compared to the state-of-the-art algorithms of [19] and [14], we have superior performance on the combination of classification accuracy, individual fairness, and group fairness in the preponderance of six different real-world classification tasks requiring non-discrimination. The results show very little reduction in classification accuracy with much improvement in

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

individual and group fairness measures.

The remainder of the paper is organized as follows. We first provide background on individual and group fairness definitions and detectors in Sec. 2. Next, in Sec. 3, we propose a new post-processing bias mitigation algorithm that accounts for both individual and group fairness. In Sec. 4, we provide empirical results on several real-world datasets including comparisons to [14, 19]. Finally, we conclude the paper in Sec. 5.

2. INDIVIDUAL AND GROUP FAIRNESS

In this section, we introduce notation, provide working definitions of individual and group fairness, and detail methods for detecting individual bias and mitigating group bias.

Consider a supervised classification problem with features $\mathbf{X} \in \mathcal{X}$, categorical protected attributes $\mathbf{D} \in \mathcal{D}$, and categorical labels $Y \in \mathcal{Y}$. We are given a set of training samples $\{(\mathbf{x}_1, \mathbf{d}_1, y_1), \ldots, (\mathbf{x}_n, \mathbf{d}_n, y_n)\}$ and would like to learn a classifier $\hat{y} : \mathcal{X} \times \mathcal{D} \to \mathcal{Y}$. For ease of exposition, we will only consider a scalar binary protected attribute, i.e. $\mathcal{D} = \{0, 1\}$, and a binary classification problem, i.e. $\mathcal{Y} = \{0, 1\}$.[1] The value $d = 1$ is set to correspond to the *privileged* group (e.g. whites in the United States in criminal justice applications) and $d = 0$ to *unprivileged* group (e.g. blacks). The value $y = 1$ is set to correspond to a *favorable* outcome (e.g. receiving a loan or not being arrested) and $y = 0$ to an *unfavorable* outcome. Based on the context, we may also deal with probabilistic binary classifiers with continuous output scores $\hat{y}_S \in [0, 1]$ that are thresholded to $\{0, 1\}$.

One definition of individual bias is as follows. Sample i has individual bias if $\hat{y}(\mathbf{x}_i, d = 0) \neq \hat{y}(\mathbf{x}_i, d = 1)$. Let $b_i = I[\hat{y}(\mathbf{x}_i, d = 0) \neq \hat{y}(\mathbf{x}_i, d = 1)]$, where $I[\cdot]$ is an indicator function. The individual bias score, $b_{S,i} = \hat{y}_S(\mathbf{x}_i, d = 1) - \hat{y}_S(\mathbf{x}_i, d = 0)$, is a soft version of b_i. To compute an individual bias summary statistic, we take the average of b_i across test samples.

One notion of group fairness known as *disparate impact* is defined as follows. There is disparate impact if

$$\frac{\mathbb{E}[\hat{y}(\mathbf{X}, D) \mid D = 0]}{\mathbb{E}[\hat{y}(\mathbf{X}, D) \mid D = 1]} \tag{1}$$

is less than $1 - \epsilon$ or greater than $(1 - \epsilon)^{-1}$, where a common value of ϵ is 0.2.

2.1. Test Generation for Individual Bias Detection

There are two distinct problems in individual bias detection: first, determining whether there are any cases of individual bias, and second, determining the individual bias status of all samples. In our earlier work [10], we presented a technique for the first problem that systematically explores the decision space of any black box classifier to generate test samples that have an enhanced chance of being biased. The method uses two kinds of search: (a) a global search which explores the decision space such that diverse areas are covered, and (b) a local search which generates test cases by intelligently perturbing the values of non-protected features of an already found individually-biased sample. The key idea is to use dynamic symbolic execution, an existing systematic test case generation technique for programs that generates search constraints by negating the constraints in a program path and uses a constraint solver to find new search paths [24]. This algorithm is useful in solving the second of the distinct problems from a computational perspective when used on a batch of samples in settings involving a large number of attributes and attribute values.

2.2. Post-Processing to Achieve Group Fairness

To achieve acceptable group fairness, various post-processing methods may be applied to change the label outputs of the classifier $\hat{y}_i$ to other labels $\check{y}_i \in \mathcal{Y}$. The reject option classification (ROC) method of [14] considers *uncertain* samples with $|\hat{y}_S - 0.5| < \theta$ (assuming 0.5 is the classification threshold) for some margin parameter θ and assigns $\check{y}_i = 1$ for samples with $d_i = 0$ and assigns $\check{y}_i = 0$ for samples with $d_i = 1$. For *certain* samples outside the so-called reject option band, $\check{y}_i = \hat{y}_i$. The θ value may be optimized to achieve the requirement on disparate impact.

The algorithm proposed by [19], equalized odds post-processing (EOP), is targeted to a different group fairness measure: equalized odds rather than disparate impact. Perfect equalized odds requires the privileged and unprivileged groups to have the same false negative rate and same false positive rate. The algorithm solves an optimization problem to find probabilities with which to assign $\check{y}_1 = 1 - \hat{y}_i$. There are four such probabilities for the following four combinations: $(d_i = 0, \hat{y} = 0)$, $(d_i = 0, \hat{y} = 1)$, $(d_i = 1, \hat{y} = 0)$, and $(d_i = 1, \hat{y} = 1)$. With these probabilities, the individual points whose prediction is flipped is a random draw. The methods of [20, 21] are refinements of [19] and share the key characteristics.

3. PROPOSED ALGORITHM

The new fairness post-processing algorithm we propose is inspired by and not radically different from [14] in form. The key observation in post-processing for group fairness metrics like disparate impact is that since they are defined as expectations, the individual samples are exchangeable. Kamiran et al. [14] elect to change values of $\hat{y}_i$ to $\check{y}_i$ in a reject option band to conform to one type of human sensibility, but the same effect on disparate impact can be achieved using the same numbers of samples from elsewhere in $\mathcal{X}$. And that is precisely what we propose: elect samples from parts of $\mathcal{X}$ that likely have individual bias. In this section, we first describe individual bias detection and then how we wrap that in a post-processing bias mitigation algorithm.

3.1. Individual Bias Detector

Consider a classifier $\hat{y}$ already trained on a training dataset partition. We can evaluate the individual bias definition provided in Sec. 2 on a validation partition that has no labels to go alongside. Some of these validation samples will have individual bias and some will not. Under an assumption of some coherence or smoothness of individual bias in $\mathcal{X}$, we can learn a classifier or detector for individual bias from this validation set that will generalize to unseen samples whose individual bias is unknown. One may use any classification or anomaly detection algorithm here that provides score outputs. We use logistic regression in the empirical results.

Formally, by perturbing the d_j of validation set samples $(\mathbf{x}_j, d_j)$, $j = 1, \ldots, m$, that belong to the unprivileged group ($d_j = 0$), we obtain individual bias scores $b_{S,j}$. We construct a further dataset $\{(\mathbf{x}_1, \beta_1), \ldots, (\mathbf{x}_m, \beta_m)\}$, and use it to train an individual bias detector $\hat{b}(\cdot)$. β_j is 1 for the samples that have the highest individual bias scores, and 0 for the rest. This assignment is determined by a

[1] In many realistic settings, these simplifications do not hold, which motivate the individual bias detector component described in Sec. 3.1.

threshold τ on the individual bias scores chosen based on the disparate impact constraint on the entire validation set. This is similar to the ROC algorithm where the margin parameter is adjusted based on disparate impact requirements.

One may argue that a trained individual bias detector is unnecessary and one should simply compute b_i for all samples as they come at run-time because doing so only involves scoring using the black-box classifier model. This may be true, but with the following caveats. Firstly, in the exposition of the paper, we have assumed d_i to be scalar and binary, when in many instances it is not. Therefore, computing b_i may require several model evaluations which could be prohibitive, especially in the industrial usage we imagine in which each sample that is scored costs a certain amount of money to be paid by the entity deploying the model and remediating the bias. Secondly, we compute the binary β_j values based on the group fairness constraint, which ensures that only examples with highest individual bias scores are considered for debiasing, and there is no overcompensation. This level of control is not possible if we consider all examples with $b_i = 1$ to be equally biased.

3.2. Overall Algorithm

Once we have the individual bias detector $\hat{b}$ trained on the validation set, the bias mitigation algorithm applied in run-time to test samples is as follows. Each sample from the unprivileged group ($d_i = 0$) is tested for individual bias and if it is likely to have individual bias, i.e., $\hat{b}_i = 1$, then this sample is assigned the outcome it would have received if it were in the privileged group, i.e., $\check{y}_i = \hat{y}(\mathbf{x}_k, 1)$. To encode a human sensibility similar to ROC, all other samples are left unchanged, including samples from the privileged group.

The proposed algorithm is summarized below:

Algorithm 1 Individual+Group Debiasing (IGD) Post-Processing

```
Given classifier $\hat{y}$ trained on training set $\{(\mathbf{x}_i, d_i, y_i)\}$, and
Given validation set $\{\mathbf{x}_j \mid d_j = 0\}$, compute individual bias scores $\{b_{S,j} \mid d_j = 0\}$.
if $b_{S,j} > \tau$ then
    $\beta_j \leftarrow 1$
else
    $\beta_j \leftarrow 0$
end if
Construct auxiliary dataset $\{(\mathbf{x}_j, \beta_j) \mid d_j = 0\}$.
Train individual bias detector $\hat{b}$ on auxiliary dataset.
for all run-time test samples $(\mathbf{x}_k, d_k)$ do
    $\check{y}_k \leftarrow \hat{y}(\mathbf{x}_k, d_k)$
    if $d_k == 0$ then
        $\hat{b}_k \leftarrow \hat{b}(\mathbf{x}_k)$
        if $\hat{b}_k == 1$ then
            $\check{y}_k \leftarrow \hat{y}(\mathbf{x}_k, 1)$
        else
            $\check{y}_k \leftarrow \hat{y}_k$
        end if
    else
        $\check{y}_k \leftarrow \hat{y}_k$
    end if
end for
```

4. EMPIRICAL RESULTS

We evaluate our proposed algorithm on three standard datasets: UCI Adult (an income dataset based on a 1994 US Census database; 45,222 samples; favorable outcome: income greater than \$50,000; protected attributes: sex, race), UCI Statlog German Credit (a credit scoring dataset; 1,000 samples; favorable outcome: low risk; protected attributes: sex, age), and ProPublica COMPAS (a prison recidivism dataset; 6,167 samples. favorable outcome: does not reoffend; protected attributes: sex, race). Each of the three datasets has two binary protected attributes that we consider as two different problems, yielding six problems overall. We compare our proposed individual+group debiasing (IGD) algorithm with ROC [14] and EOP [19] using the implementations of ROC and EOP provided in the AI Fairness 360 toolkit [25].

We process and load each dataset using the AI Fairness 360 toolkit and randomly divide it into 60% training, 20% validation and 20% testing partitions. We conduct experiments with 25 such random partitions of the datasets, allowing us to provide error bars in the empirical results that follow. Using the training partition, we fit both ℓ_2-regularized logistic regression and random forests as black-box classifiers. For random forests, we set the number of trees to 100 and the minimum samples per leaf node to 20.

The parameters of all three bias mitigation approaches are optimized on the validation partition of the dataset. Both the ROC and the EOP approaches require ground truth class labels in the validation set, whereas the proposed IGD approach, being a pure run-time method, does not. ROC and IGD are optimized to achieve disparate impact in the range $(0.8, 1.25)$, i.e., $\epsilon = 0.2$. EOP, being designed for equalized odds rather than disparate impact, cannot be optimized for ranges of disparate impact.

In the subsections that follow, we first demonstrate the efficacy of the individual bias detector used in the proposed IGD algorithm and then compare the three algorithms for classification accuracy, disparate impact, and individual fairness.

4.1. Validation Results on Individual Bias Generalization

We verify the generalization performance of the individual bias detector on unseen test data. Since the individual bias detector is used only on unprivileged group samples ($d = 0$), its performance measure is only computed for this subset. The ground truth labels for the bias detector are obtained by actually computing the individual bias scores ($b_{S,k}$) for all unprivileged group samples in the test data, and identifying the ground truth bias labels (β_k) based on the disparate impact constraint. These labels are compared with the labels predicted by the bias detector ($\hat{b}_k$), and the balanced classification accuracy is computed.

This performance of the bias detector is shown in Fig. 1 for all dataset and protected attribute combinations when the black-box classifier is logistic regression. All accuracy values are more than 0.85, which illustrates its clear effectiveness for the purpose at hand. The detector performs similarly when the black-box classifier is random forests, with a minimum accuracy of approximately 0.80.

4.2. Fairness Comparisons

We use three measures for comparing EOP, ROC, and IGD: (a) individual bias, (b) disparate impact, and (c) balanced classification accuracy. These measures are computed using the post-processed predictions $\check{y}$. The individual bias measure is the summary statistic discussed in Sec. 2, the disparate impact measure is defined in

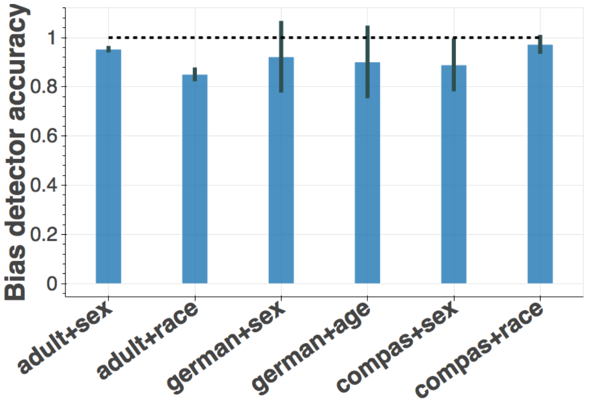

Fig. 1. Balanced accuracy of the bias detector when the black box classifier is a Logistic Regression model. The bar shows the mean accuracy, and the vertical lines show the extent of ± 1 standard deviation. The dotted horizontal line shows the best possible performance.

(1), and balanced classification accuracy is the mean of true positive and true negative rates obtained for the predictions $\tilde{y}$ with respect to the true labels y. We also obtain these measures for the original (Orig.) predictions $\hat{y}$. As shown in Fig. 2, Fig. 3, and Fig. 4, the proposed IGD approach is the only one that consistently improves both fairness measures while keeping the accuracy close to that of the original classifier. All results are show for logistic regression as the black-box classifier, but similar results are also observed for random forests (omitted due to space constraints).

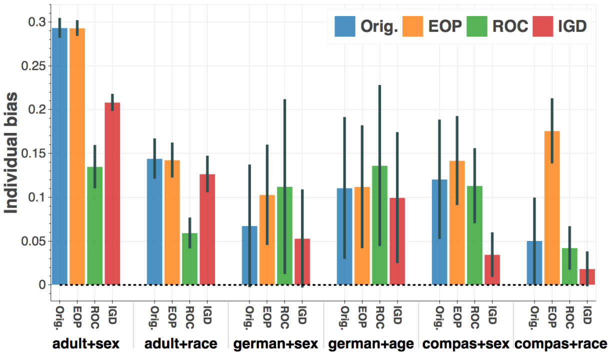

Fig. 2. Individual bias of the original model and the compared post-processing algorithms. The bar shows the mean value, and the vertical lines show the extent of ± 1 standard deviation. The dotted horizontal line shows the ideal fair value (0.0).

In individual bias, the proposed IGD method performs the best for the German and COMPAS datasets. The ROC method performs the best for the Adult dataset, at the expense of reducing the balanced accuracy. Sometimes the EOP and ROC methods increase the individual bias, which is never the case with IGD. The proposed IGD method also consistently improves disparate impact over the original predictions, although outperformed by the ROC method in five out of six cases. The strong performance of the ROC approach is likely because it does not also optimize for individual bias. The EOP method performs poorly on disparate impact, likely because it was designed to equalize odds, which may or may not always result in improved disparate impact [26]. The proposed IGD method is also the best in preserving the balanced classifier accuracy compared to the original predictions even though no ground truth labels are used in the validation partition.

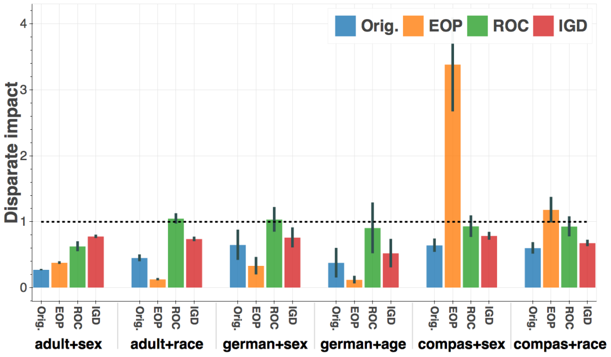

Fig. 3. Disparate impact of the original model and the compared post-processing algorithms. The bar shows the mean value, and the vertical lines show the extent of ± 1 standard deviation. The dotted horizontal line shows the ideal fair value (1.0).

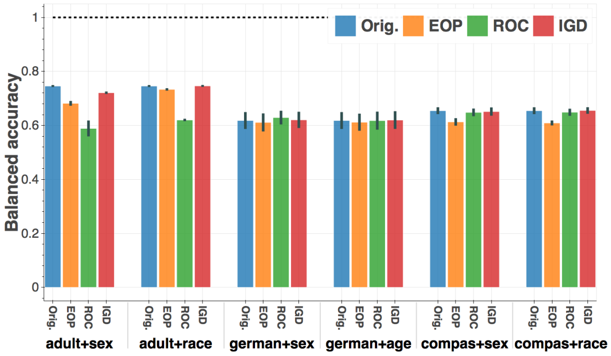

Fig. 4. Balanced classification accuracy of the original model and the compared post-processing algorithms. The bar shows the mean value, and the vertical lines show the extent of ± 1 standard deviation. The dotted horizontal line is the best possible accuracy (1.0).

5. CONCLUSION

Algorithmic fairness is an important topic for business and society, and developing new bias mitigation algorithms that address as many facets of fairness as possible is critical. In this paper, we have developed a new post-processing algorithm that targets samples with individual bias for remediation in order to improve *both* individual and group fairness metrics and shown that it does so empirically on several real-world datasets without much loss in classification accuracy. From our experience, the machine learning industry is moving towards paradigms in which there will be a separation between model building and model deployment. This will include a limited ability for deployers to gain access to the internals of pre-trained models. Therefore, post-processing algorithms, especially ones that can treat a classifier as a complete black-box are necessary. In comparison to previous work, our proposed algorithm not only tackles both individual and group fairness, but also is a pure run-time approach because it does not require ground truth class labels for the validation set.

6. REFERENCES

[1] L. R. Varshney and K. R. Varshney, “Decision making with quantized priors leads to discrimination,” *Proc. IEEE*, vol. 105, no. 2, pp. 241–255, Feb. 2017.

[2] B. A. Williams, C. F. Brooks, and Y. Shmargad, “How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications,” *J. Inf. Policy*, vol. 8, pp. 78–115, 2018.

[3] K. R. Varshney and H. Alemzadeh, “On the safety of machine learning: Cyber-physical systems, decision sciences, and data products,” *Big Data*, vol. 5, no. 3, pp. 246–255, Sep. 2017.

[4] M. Hind, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, and K. R. Varshney, “Increasing trust in AI services through supplier’s declarations of conformity,” arXiv:1808.07261, Aug. 2018.

[5] A. K. Menon and R. C. Williamson, “The cost of fairness in binary classification,” in *Proc. Conf. Fairness Accountability Transp.*, New York, USA, Feb. 2018, pp. 107–118.

[6] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, “Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis,” *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 1106–1119, Oct. 2018.

[7] A. Narayanan, “Translation tutorial: 21 fairness definitions and their politics,” in *Proc. Conf. Fairness Accountability Transp.*, New York, USA, Feb. 2018.

[8] I. Źliobaităź, “Measuring discrimination in algorithmic decision making,” *Data Min. Knowl. Disc.*, vol. 31, no. 4, pp. 1060–1089, Jul. 2017.

[9] S. Galhotra, Y. Brun, and A. Meliou, “Fairness testing: Testing software for discrimination,” in *Proc. Joint Meeting Found. Softw. Eng.*, Paderborn, Germany, Sep. 2017, pp. 498–510.

[10] A. Agarwal, P. Lohia, S. Nagar, K. Dey, and D. Saha, “Automated test generation to detect individual discrimination in AI models,” arXiv:1809.03269, Sep. 2018.

[11] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar, “A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, London, UK, Aug. 2018, pp. 2239–2248.

[12] N. Hurley and S. Rickard, “Comparing measures of sparsity,” *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[13] B. d’Alessandro, C. O’Neil, and T. LaGatta, “Conscientious classification: A data scientist’s guide to discrimination-aware classification,” *Big Data*, vol. 5, no. 2, pp. 120–134, Jun. 2017.

[14] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” in *Proc. IEEE Int. Conf. Data Min.*, Brussels, Belgium, Dec. 2012, pp. 924–929.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in *Proc. Innov. Theo. Comp. Sci. Conf.*, Cambridge, USA, Jan. 2012, pp. 214–226.

[16] C. Dwork and C. Ilvento, “Individual fairness under composition,” in *Proc. Fairness Accountability Transp. Mach. Learn. Workshop*, Stockholm, Sweden, Jul. 2018.

[17] R. Zemel, Y. L. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proc. Int. Conf. Mach. Learn.*, Atlanta, USA, Jun. 2013, pp. 325–333.

[18] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, “Optimized pre-processing for discrimination prevention,” in *Adv. Neur. Inf. Process. Syst. 30*, Long Beach, USA, Dec. 2017, pp. 3992–4001.

[19] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Adv. Neur. Inf. Process. Syst. 29*, Barcelona, Spain, Dec. 2016, pp. 3315–3323.

[20] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, “On fairness and calibration,” in *Adv. Neur. Inf. Process. Syst. 30*, Long Beach, USA, Dec. 2017, pp. 5680–5689.

[21] R. Canetti, A. Cohen, N. Dikkala, G. Ramnarayan, S. Scheffler, and A. Smith, “From soft classifiers to hard decisions: How fair can we be?” arXiv:1810.02003, Oct. 2018.

[22] C. K. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 41–46, Jan. 1970.

[23] K. R. Varshney, “A risk bound for ensemble classification with a reject option,” in *Proc. IEEE Stat. Signal Process. Workshop*, Nice, France, Jun. 2011, pp. 769–772.

[24] P. Godefroid, N. Klarlund, and K. Sen, “DART: Directed automated random testing,” in *Proc. ACM SIGPLAN Conf. Program. Lang. Design Impl.*, Chicago, USA, 2005, pp. 213–223.

[25] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” arXiv:1810.01943, Oct. 2018.

[26] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” arXiv:1802.04422, Feb. 2018.