Did You Do Your Homework? Raising Awareness on Software Fairness and Discrimination

Max Hort
University College London
London, United Kingdom
max.hort.19@ucl.ac.uk

Federica Sarro
University College London
London, United Kingdom
f.sarro@ucl.ac.uk

***Abstract*—Machine Learning is a vital part of various modern day decision making software. At the same time, it has shown to exhibit bias, which can cause an unjust treatment of individuals and population groups. One method to achieve fairness in machine learning software is to provide individuals with the same degree of benefit, regardless of sensitive attributes (e.g., students receive the same grade, independent of their sex or race). However, there can be other attributes that one might want to discriminate against (e.g., students with homework should receive higher grades). We will call such attributes anti-protected attributes. When reducing the bias of machine learning software, one risks the loss of discriminatory behaviour of anti-protected attributes. To combat this, we use grid search to show that machine learning software can be debiased (e.g., reduce gender bias) while also improving the ability to discriminate against anti-protected attributes.**

***Index Terms*—software fairness, discrimination, classification**

I. Introduction

In recent years, Machine Learning (ML) software has found a staggering rise in popularity and is nowadays used in a variety of decision making software, such as justice risk assessment [1], [2] and loan applicant filtering [3]. While ML software supports the decision making process, it has shown to exhibit discriminatory behaviours [4]. Such discriminatory behaviour of ML software can affect profits [5] and human rights [6], and can furthermore fall under regulatory control [4], [7], [8].

In 2018, Brun and Meliou [9] stated that ensuring the fairness of software systems (software fairness) is a critical software engineering problem to be tackled from multiple directions, and since then it has gained more and more attention from software engineering research [10], [11], [12], [13], [14], [15].

Software fairness aims to provide algorithms that operate in a non-discriminatory manner [16]. One way to achieve fairness is to treat individuals equally, such that they receive the same degree of benefit. However this can lead to uncertain situations, as we illustrate in Figure 1. In the two scenarios, we show a simplified school grading system that is used to assign grades to students, with "A" being the best grade. There are two types of students, distinguishable by their appearance (triangular and oval) which is used to represent a sensitive attribute (i.e., student should not receive a favourable grade based on their appearance). Additionally, we provide information on whether students did their homework or not. In Scenario 1, we can see that the two students in question are identical, except for their appearance. They receive the same grade and therefore the grading is unbiased. In Scenario 2, we can see that there still is no unfavourable treatment based on appearance, as the grades remain identical. However, we have to ask the question whether this scenario is as fair as Scenario 1. When only considering the impact of appearance on the grade, there is no bias to be found, however Student 2 received the same grade as Student 1 without doing the homework. Therefore, one could say that there is an unfair treatment in terms of "doing homework" which is usually not considered a sensitive attribute.

The concept we want to study further is that if one were to only consider sensitive attributes (e.g., grading is fair if everyone receives the same grade, as no differences in appearance can be seen), we lose the ability to discriminate against desired characteristics (e.g., did the student do homework?). Therefore, it can be beneficial to reduce discrimination in regards to sensitive attributes (e.g., appearance) and increase discrimination in regards to specific attributes such as "Homework". Nonetheless, one needs to be careful when choosing such attributes because they could be correlated to sensitive attributes and thereby directly impact and potentially negatively affect fairness. Therefore, we do not claim that we found definite attributes that one should distinguish against, but we want to clarify that only taking sensitive attributes into account when investigating bias can lead to undesired side effects. The choices of which type of discrimination is illegal, acceptable or desired therefore remains for law makers and domain experts to make [17], [18].

In this paper, we provide initial empirical evidence on the potential harm of bias reduction on what we call *anti-protected attributes* (Section II). To this end, we investigate a popular dataset in fairness research and specify two examples of anti-protected attributes (Section III). Our experiments (Section IV) show that increasing fairness of sensitive attributes might have a detrimental effect on the discriminatory power of anti-protected attributes. We also show that remedies can be taken to reduce such an effect and it is possible to simultaneously reduce bias and increase discrimination of anti-protected attributes by using grid search as a proof-of-concept. To the best of our knowledge, this work is the first to raise awareness

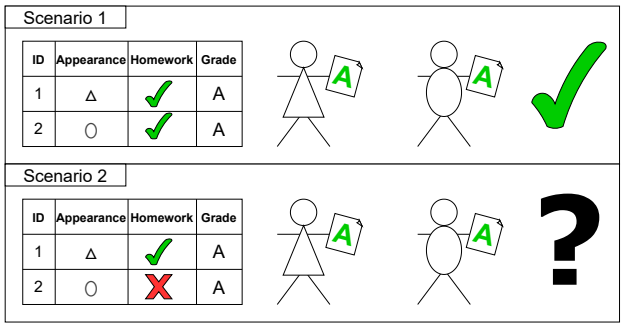


Fig. 1: Example scenarios of school grading system with the sensitive attribute "appearance".

on the problem and our results call for the community to undertake further research in this direction (Section V).

II. Background and Related Work

ML software can be deemed unfair if it favours individuals or groups of people based on sensitive attributes rather than merit. Sensitive attributes (e.g., sex, race, age) which divide the population in two groups (privileged and unprivileged) are called **protected** attributes. Thereby, individuals of the privileged group receive a favourable treatment in contrast to the unprivileged group. Non-critical attributes are called **unprotected** attributes.

To measure the degree of unfairness, fairness metrics have been introduced. These metrics can be divided in two different types: **Individual fairness** (similar individuals are treated equally) [19], **Group fairness** (population groups are treated equally) [20].

According to the scenario illustrated in Figure 1, by dividing the population in two groups based on appearance, we are interested in the fairness of privileged and unprivileged groups (i.e., group fairness). A fairness metric that can be used to measure group fairness is the *Statistical Parity Difference (SPD)*. SPD requires that predictions are made independent of protected attributes [21], ensuring that the ratio of positive and negative classifications are identical over the two groups. Given the predictions of a classification model $\hat{y}$, the probability Pr of favourable and unfavourable predictions for privileged and unprivileged ($D = privileged$, $D = unprivileged$) should be identical [19]:

$$\begin{aligned} SPD = Pr(\hat{y} = 1 | D = unprivileged) \\ - Pr(\hat{y} = 1 | D = privileged) \end{aligned} \tag{1}$$

The ability to determine the fairness of a classification model allows for the use of bias mitigation methods. Bias mitigation methods are algorithms that aim to reduce bias (according to fairness metrics) of classification models. There are three stages in which bias mitigation methods can be applied: pre-processing (i.e., adaptation of training data) [22], [23], [13], [24], in-processing (i.e., during model training) [25], [26], [27], [28], [21], post-processing (i.e., after the model has been trained) [29], [30], [31], [32], [33]. Previous work has shown that simply removing protected attributes from ML systems does not achieve fairness, as several other features can be correlated with the protected attribute [34], [35], [4].

Kamiran and Žliobaitė [17] and Žliobaitė et al. [18] introduced the concepts of two types of discrimination: **explainable** and **unexplainable**. Unexplainable discrimination is considered illegal, as it unjustifiably discriminates against population groups. Explainable discrimination on the other hand can be explained via other attributes in the datasets (i.e., attributes that are not protected). In contrast to Kamiran and Žliobaitė [17], and Žliobaitė et al. [18] who reduced unexplainable discrimination while allowing explainable discrimination to remain, we particularly aim at reducing bias **and** increasing discrimination that could be deemed "explainable". We call such attributes **anti-protected** attributes, as instead of reducing discrimination as done for protected attributes, the goal is to increase it (e.g., students who do homework should receive higher grades).

TABLE I: Features of the Adult dataset.

Name	Type	Values
age	continuous	17-90
workclass	categorical	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	continuous	12285-1490400
education	categorical	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education-num	continuous	1-16
marital-status	categorical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	categorical	Tech-support, Craft-repair, Other-service, Transport-moving, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Sales , Priv-house-serv, Protective-serv, Armed-Forces
relationship	categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	categorical	Female, Male
capital-gain	continuous	0-99999
capital-loss	continuous	0-4356
hours-per-week	continuous	1-99
native-country	categorical	41 countries, including: United-States, England, Japan, Cuba

III. The Adult Dataset

The **Adult Census Income (Adult)** [36] holds financial and demographic information about individuals from the 1994 U.S. census. The Adult dataset contains a total of $48,842$ instances with 14 features. A classification is made to determine whether individuals receive an annual income above $\$50,000$ a year. If the answer is "yes", individuals receive a favourable label of 1 and 0 otherwise. Table I provides a detailed description of the features.

There are two feature types present in the Adult dataset: categorical and continuous. The feature "education" is available both as a categorical and corresponding continuous feature. Whereas the continuous feature "education-num" interprets the level of education as a sequence. The feature "fnlwgt" represents the final weight of each row in the datasets, to show the estimated amount of people that each row represents. By default, the AIF360 framework [37], a popular framework in fairness research, does not weigh instances in the dataset and therefore removes the feature "fnlwgt".

A. Protected Attributes

The Adult dataset has two protected attributes [37]. The first one is based on the feature "sex", a categorical feature with values ("female", "male"). Therefore, the population can easily be divided in two groups, one for each category. In particular, "male" represents the privileged group and "female" the unprivileged group. The second protected attribute is the "race" of individuals. As shown in Table I, there are five different categorical values that individuals can be assigned to. To divide the population in two groups, some of these have to combined. The AIF360 framework [37] divides the population in "white" (privileged) and "non-white" (unprivileged) by combining the other four categories (Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black).

B. Anti-Protected Attributes

In the Adult dataset, we see two features that have the potential to be used as anti-protected attributes: education-num, hours-per-week.

Education-num can be used to distinguish between a "high" degree of education and "low" degree of education, which has shown to impact the average salary [38]. Furthermore, assuming that wages are paid hourly, hours-per-week is directly correlated with the income earned by an individual.

At first, we need to specify how, given a feature, we are able to divide the population in two groups. The two feature types can be treated as follows:

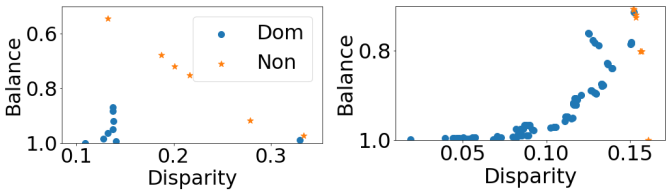
- categorical: combine categories until two remain;
- continuous: pick a threshold and split in to "smaller than threshold" and "larger than threshold".

To then divide the population in privileged and unprivileged groups, based on anti-protected attributes, two approaches can be followed: 1) create balanced groups (i.e., the size of privileged and unprivileged groups are similar); 2) maximize disparity (e.g., the proportion of individuals with a favourable outcome in the privileged group is higher than in the unprivileged group). Among the two approaches, we prioritize the creation of balanced groups, while still showing disparity among them. For example, Kamiran and Calers [34] picked a threshold of 25 when dividing the population in to "young" and "old", with the protected attribute "age". This threshold follows the first strategy, as the highest degree of disparity between the two groups can be found at this threshold.

Figure 2 illustrates our investigation of anti-protected attributes. As both analyzed features are continuous, we investigate each potential threshold (i.e., we choose every integer value possible, as a threshold t, that divides the population in two groups with one group $<= t$ and the other group $> t$). For each investigated threshold we calculate the following:

- Disparity: absolute rate difference of favourable outcomes for both groups (minimized at 0);
- Balance: size of the larger group (minimized at 0.5).

Among all possible thresholds t, we are only interested in non-dominated ones (i.e. there is no other threshold with lower disparity and better balance). In particular, we choose $t = 11$ for the feature "education-num" (i.e., every instance with $educationnum <= 11$ is one group, the remaining instances in the other), and $t = 41$ for the feature "hours-per-week".



(a) Education-num (b) Hours-per-week

Education-num			Hours-per-week		
t	Balance	Disparity	t	Balance	Disparity
15	0.971	0.334	2	0.999	0.161
14	0.916	0.279	50	0.802	0.157
13	0.752	0.216	49	0.801	0.156
12	0.719	0.201	45	0.725	0.154
11	**0.677**	**0.187**	44	0.719	0.153
10	0.546	0.132	42	0.707	0.152
			41	**0.706**	**0.152**

Fig. 2: Analysis of anti-protected attributes with different thresholds t. Figures (a) and (b) illustrate the Disparity and Balance achieved with different thresholds while highlighting non-dominated thresholds ("Non") with stars and dominated thresholds ("Dom") with circles. The specific values of each non-dominated data point from (a) and (b) are shown in the table underneath.

IV. Empirical Study

In this section, we describe the design of a preliminary study on anti-protected attributes for binary classification. With this analysis, we want to show the relevance and potential harm caused by the negligence of anti-protected attributes. At first, we measure the bias that protected and anti-protected attributes hold according to a fairness metric (Section II) when a classification model is trained for accuracy. Afterwards, we apply two existing bias mitigation methods to check whether the debiasing for a protected attribute leads to a loss in discrimination of anti-protected attributes. We furthermore propose a grid search approach to improve bias in regards of protected and anti-protected attributes.

A. Setup

We choose a Logistic Regression model, provided by `scikit`[39], for our evaluation, as this has been frequently used in fairness research [23], [13], [21], [40], [41], [33]. To determine the fairness of the Logistic Regression model, we train it on 70% data of the Adult dataset, use 15% of the data as the test set and the remaining 15% as a validation set. We repeat the training and testing procedure 50 times, with different data splits. SPD is then averaged over the 50 repetitions [27], [42].

As we are not concerned about which group (privileged or unprivileged) is more likely to receive a favourable outcome, but only in the degree of bias among the two groups, we measure the absolute values of SPD. Thereby, the highest degree of bias is 1 while 0 denotes no bias.

B. Bias Mitigation Methods

We investigate two bias mitigation methods and their impact on bias according to protected and anti-protected attributes. In particular, we choose **Reweighing** [24], [35] and **Equalized Odds Post-processing** (EqualizedOdds) [30] due to their ability to reduce bias on the Adult dataset and their ease of use (both are publicly available in AIF360 [37]). Reweighing is a pre-processing methods which applies weights to the instances of the training dataset such that a fair classifier could be trained. EqualizedOdds relabels prediction made by a classifier, in a post-processing stage, to improve fairness. The goal of EqualizedOdds is to achieve equal true positive and false positive rates across privileged and unprivileged groups.

C. Grid Search

To support the feasibility of our proposed idea, that protected and anti-protected attributes can be improved simultaneously, we apply a naive approach of grid search. Given sets of parameters, grid search performs an exhaustive search across each possible combination of parameters. In particular, we try to find configurations of Logistic Regression with better results (i.e., a decrease of SPD for protected attributes, and an increase for anti-protected attributes). As grid search grows exponentially with the number of search parameters [43], we limit our search as follows (default values are shown in bold):

- solver (algorithm used for optimization): newton-cg, **lbfgs**, liblinear, sag, saga;
- penalty (norm used in penalization): l1, **l2**, elasticnet, none;
- C (inverse of regularization strength): 0.5, **1**, 2, 3, 4, 5, 10;
- max_iter (maximum number of iterations for solver to converge): 50, 75, **100**, 125, 150.

In total, we investigate 315 viable parameter configurations out of 700 combinations.

D. Results

We investigated whether the reduction of bias in respect to a single protected attribute (e.g., Reweighing$_{race}$ aims to reduce the bias in respect to the feature "race") causes a Logistic Regression model to lose discriminative behaviour of anti-protected attributes (i.e., attributes that one would want to discriminate against). Table II illustrates the results. As can be seen, the two bias mitigation methods (Reweighing and EqualizedOdds) are able to reduce the SPD of both protected attributes. In both cases, optimizing for either of the two protected attributes (sex, race) reduces bias. However, we observe that while the bias of protected attributes is reduced, in comparison to the default Logistic Regression model, the SPD of anti-protected attributes are reduced as well. A reduction of SPD in anti-protected attributes is not desired, as one would want to discriminate against these.

We applied grid search to verify whether improvements in protected (reduce SPD) and anti-protected attributes (increase SPD) are achievable. Each configuration is trained on the 50 datasplits. Afterwards, we average the SPD achieved for the four attributes and check whether the performance on the validation set outperforms the default Logistic Regression model (e.g., we check whether the SPD on the validation set for "race" is lower than the default SPD on the test set). Among the 315 configurations of grid search, 7 are able to achieve improvements for all four attributes. The average SPD of the 7 configurations is shown in Table II. We can observe that, while the improvements are small, it is possible to achieve improvements in protected and anti-protected attributes simultaneously.

TABLE II: Performance of bias mitigation methods and grid search. The Statistical Parity Difference for two protected attributes (sex, race) and two anti-protected attributes (education-num, hours-per-week) are compared to the default Logistic Regression model. Improvements are shown in bold.

	Statistical Parity Difference			
	Sex	Race	Education	Hours
Default	0.171	0.081	0.421	0.252
Reweighing$_{Sex}$	**0.066**	**0.071**	0.375	0.237
Reweighing$_{Race}$	**0.167**	**0.040**	0.412	0.245
EqualizedOdds$_{Sex}$	**0.092**	**0.065**	0.392	0.221
EqualizedOdds$_{Race}$	**0.165**	**0.056**	0.412	0.244
Grid Search	**0.170**	**0.080**	**0.423**	0.251

V. Conclusions and Future Work

In this NIER paper, we investigate software fairness and potential issues that arise when bias mitigation is focused solely on reducing bias according to protected attributes (such as gender or race) and disregards other attributes. Such attributes can be non-critical, but there can also be attributes that practitioners actively want to discriminate against (e.g., more working-hours lead to a higher income), which previous work did not consider.

We analyzed the Adult dataset, a popular dataset in fairness research, for the potential use of anti-protected attributes (attributes that should be discriminated against). Initial empirical evidence, based on two popular existing bias mitigation methods, showed that while bias mitigation methods are able to reduce bias according to protected attributes, they also reduce the ability to discriminate against anti-protected attributes. We performed a grid search, as a proof-of-concept, to show that it is possible to achieve improvements discrimination in both protected and anti-protected attributes.

We hope that our investigation will spark further SE research in facing the problem of software fairness, as this an emerging field of study. In particular, by taking further attributes into account, bias mitigation gains relevance from a multi- or many-objective optimization point of view. Other areas of future work include the analysis of further datasets [44], [45]. Each dataset would require an individual analysis of potential anti-protected attributes.

Acknowledgements

M. Hort and F. Sarro are supported by the ERC grant no. 741278 (EPIC).

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias. propublica," *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

[3] A. Mukerjee, R. Biswas, K. Deb, and A. P. Mathur, "Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management," *International Transactions in operational research*, vol. 9, no. 5, pp. 583–597, 2002.

[4] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[5] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, "Detecting price and search discrimination on the internet," in *Proceedings of the 11th ACM workshop on hot topics in networks*, 2012, pp. 79–84.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[7] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[8] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," 2011.

[9] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2018, pp. 754–759.

[10] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, pp. 1–1, 2020.

[11] J. Zhang and M. Harman, "Ignorance and prejudice in software fairness," in *43th International Conference on Software Engineering (ICSE)*, 2021.

[12] J. Horkoff, "Non-functional requirements for machine learning: Challenges and new directions," in *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 2019, pp. 386–391.

[13] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, "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, pp. 654–665.

[14] J. Chakraborty, K. Peng, and T. Menzies, "Making fair ml software using trustworthy explanation," in *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2020, pp. 1229–1233.

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

[16] 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," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 329–338.

[17] F. Kamiran and I. Žliobaitė, "Explainable and non-explainable discrimination in classification," in *Discrimination and Privacy in the Information Society*. Springer, 2013, pp. 155–170.

[18] I. Žliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 992–1001.

[19] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[20] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

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

[22] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[24] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

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

[26] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 2564–2572. [Online]. Available: http://proceedings.mlr.press/v80/kearns18a.html

[27] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 319–328.

[28] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "A convex framework for fair regression," *FAT-ML Workshop*, 2017.

[29] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

[30] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[31] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[32] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 869–874.

[33] F. Kamiran, S. Mansha, A. Karim, and X. Zhang, "Exploiting reject option in classification for social discrimination control," *Information Sciences*, vol. 425, pp. 18–33, 2018.

[34] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, 2009, pp. 1–6.

[35] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 2009, pp. 13–18.

[36] R. Kohav, "Adult data set," http://archive.ics.uci.edu/ml/ datasets/adult.

[37] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[38] T. Stobierski. (2020, jun) Average salary by education level: The value of a college degree. [Online]. Available: https://www.northeastern.edu/bachelors-completion/news/average-salary-by-education-level/

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th International Conference on Data Mining*. IEEE, 2012, pp. 924–929.

[41] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2012, pp. 35–50.

[42] S. Biswas and H. Rajan, "Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness," *arXiv preprint arXiv:2005.12379*, 2020.

[43] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[44] propublica, "data for the propublica story 'machine bias'," https://github.com/propublica/compas-analysis/.

[45] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.