Training Fair ML Models through Causal Multi-Objective Optimization

Jason Lee, Keyu Li

Introduction

Problem Overview (Motivation, Background Information)

Bias in machine learning models affect domains such as healthcare, justice, etc. For example, Northpointe's COMPAS model came under public scrutiny when it was reported to racially discriminate against African Americans in its assessment of recidivism.

In order to mitigate algorithmic bias, one line of literature directly optimizes for fairness via multi-objective optimization (MOO). In this line of work, we train a model that optimizes for both loss and fairness objectives simultaneously. Because fairness metrics are often at odds with one another, existing literature optimizes for multiple associational fairness metrics [1]. However, using multiple fairness objectives makes the search space for a pareto-optimal solution more complex. It also makes it harder for decision-makers to decide on an optimal solution. In our project, we want to determine whether we can identify an universal fairness objective that can optimize for conflicting fairness objectives simultaneously.

To this end, counterfactual fairness [2] throws light in this scenario. In the paper "Causal Reasoning for Algorithmic Fairness," [3] the authors suggest that causal metrics like counterfactual fairness is superior to associational metrics of fairness, since satisfying counterfactual fairness acts as a proxy for both group-based and individual fairness. Therefore, we attempt to train a fairer machine learning model using causal multi-objective optimization.

Project Outline

By using causal metrics like counterfactual fairness [2] as a universal fairness objective, we can resolve the issue of using multiple, conflicting fairness metrics. We compare the performance of a multi-objective optimization model that uses both equalized odds and predictive parity as fairness objectives (as suggested by this paper) [4] and our proposed model that uses only counterfactual fairness as a fairness objective to determine whether causal reasoning is a sufficient metric for MOO of fair models. We also compare multi-objective methods to two baseline methods: 1) fairness through unawareness and 2) counterfactual fairness.

We choose to analyze the COMPAS recidivism dataset to conduct our experiments, and will treat race as the sensitive attribute for our project. Specifically, in line with

ProPublica’s analysis on the COMPAS dataset, Caucasians will be our unprotected group while African Americans will be our protected group. We first analyze the current COMPAS score using global feature importance generated by an Explainable Boosting Machine. Then, we compare the performance of our four baseline and proposed models using the COMPAS dataset. We mainly use 1) classification accuracy 2) F1-score to evaluate our predictive performance while using 1) equalized odds 2) predictive parity 3) race-wise FP(False Positive Rate) and FN(False Negative Rate) to evaluate our fairness.

Methodology

In this section, we provide a condensed summary of literature review we did to design & formulate the experimental framework for the project (more in-depth review can be found in the midterm proposal)

Choice of Multi-Objective Optimization Method
As part of literature review, we first read up on different MOO strategies. There were a few notable approaches. First is the **weighted sum method**, where multiple objectives are linearly combined into a single objective function.

$$\textbf{minimize} \quad F(\boldsymbol{x}) = \sum_{m=1}^{M} w_m f_m(\boldsymbol{x}),$$

The above equation illustrates the weighted sum method. M is the number of objective functions for our model, $f_m(x)$ the m-th objective function, and w_m the corresponding weight that trades off the relative importance between objectives.

Since the nature of our project isn’t focused on comparing the performance of different MOO methodologies, rather comparing the choice of objective functions, we simply need to standardize our choice of training methodology across models. Therefore, we opt to use the **weighted sum method** as our MOO methodology.

Defining Fairness Objectives
Next, we had to concretely define the objective functions for our fairness objectives. Given S is the sensitive attribute (protected group), O the outcome of the model, Y the true outcome, the optimization problem for equalized odds and predictive parity can be defined as follows:

$$PPV_{diff} = P(Y = 1 \mid S = 1, O = 1) - P(Y = 1 \mid S = 0, O = 1)$$

$$EO_{diff} = \frac{1}{n} \sum P(O = i \,|\, S = 1, Y = 1 - i) - P(O = i \,|\, S = 0, Y = 1 - i)$$

$$i \in I = \{0,1\}, n = |I| = 2$$

These objective functions draw directly from the definition of predictive parity and equalized odds, and would require us to have a batch size > 1 during training (mini-batch or batch learning) in order to compute the objective function.

It is a much harder task to design an objective function for counterfactual fairness during model training. We define counterfactual fairness as follows:

$$P(\hat{Y} = \hat{y}_{A=a} | A = a, Z = z) = P(\hat{Y} = \hat{y}_{A=a'} | A = a, Z = z)$$

where the predicted class label must stay the same in the counterfactual world where an individual belonged to a different class. Several methods exist in literature to measure counterfactual fairness.

One is to define a structural causal model (SCM) that models the causal relationship between endogenous and exogenous variables in the dataset, and directly compute counterfactual fairness. This involves intervening on the value of the protected attribute, and keeping the unobserved (latent) variables constant. In this method, our resulting objective function will involve minimizing the difference between the probability distributions of the actual and the counterfactual prediction. We define the counterfactual objective as follows:

$$\mu_j(f, \mathbf{x}_i, a_i, a') := \max\{0, \left| f(\mathbf{x}_{i,A^j \leftarrow a_i}, a_i) - f(\mathbf{x}_{i,A^j \leftarrow a'}, a') \right| - \epsilon\}$$

Much like the hinge loss, we evaluate whether the difference in model predictions between the original and counterfactual input exceeds a threshold ϵ. All differences below the threshold ϵ incur a loss of 0.

Another line of work focuses on counterfactual explanations. For the purpose of this project, we decided to define a SCM for measuring counterfactual fairness. However, the use of counterfactual explanations to measure counterfactual fairness is an interesting domain for MOO problems.

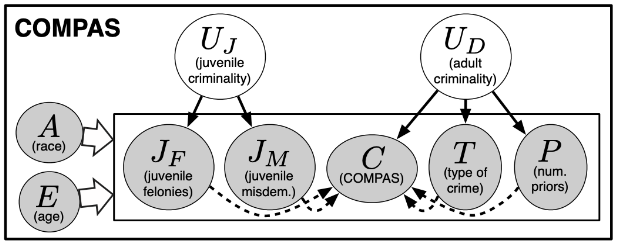
Experiment Setup

Dataset

In line with ProPublica's analysis, we treat score = 'Low' as unlikely to recidivate and score = 'Medium' and 'High' as likely to recidivate. Since we assume that the COMPAS scores are racially biased, we use an individual's true recidivism outcome as the training label for our models, and compare the difference in COMPAS scores vs. our scores.

Structural Causal Model

Both training a counterfactually fair algorithm and using counterfactual fairness as a training objective require a structural causal model, as outlined in our methodologies. To this end, we leverage a SCM proposed by Kushner et. al in their paper:



$$
\begin{aligned}
T &\sim \text{Bernoulli}(\phi(b_T + w_C^{U_D} U_D + w_C^E E + w_C^A A) \\
C &\sim \mathcal{N}(b_C + w_C^{U_D} U_D + w_C^E E + w_C^A A + w_C^T T + w_C^P P + w_C^{J_F} J_F + w_C^{J_M} J_M, \sigma_C) \\
P &\sim \text{Poisson}(\exp(b_P + w_P^{U_D} U_D + w_P^E E + w_P^A A)) \\
J_F &\sim \text{Poisson}(\exp(b_{J_F} + w_{J_F}^{U_J} + w_{J_F}^E E + w_{J_F}^A A)) \\
J_M &\sim \text{Poisson}(\exp(b_{J_M} + w_{J_M}^{U_J} + w_{J_M}^E E + w_{J_M}^A A)) \\
[U_J, U_D] &\sim \mathcal{N}(0, \Sigma)
\end{aligned}
$$

Figure 1: Casual graph and Structural Equations for the SCM proposed by Kushner et al. for the COMPAS dataset

The proposed structural causal model uses age, race, # juvenile felonies, # juvenile misdemeanors, type of crime committed, and number of priors as features to predict the COMPAS score. It also assumes latent features 'juvenile criminality' and 'adult criminality' that aren't affected by race/age to impact our covariates.

Using our data, we fit the parameters to the structural equations provided in Figure 1.

Fairness through Unawareness

A fairness through unawareness model trains a fair classifier by removing sensitive attributes from the feature space. We thus train a model that does not include race. This means that our resulting features are age, # juvenile felonies, # juvenile misdemeanors, type of crime committed, and number of priors.

We compared performance with support vector machines and explainable boosting machines to ultimately train a logistic regression model with binary cross entropy as loss to classify whether an individual is likely to recidivate within the next two years or not.

Multi-Objective Optimization

Given the fairness objectives we defined in Methodologies, we can define our multi-objective optimization objectives as follows:

To train a MOO problem with both predictive parity and equalized odds as fairness objectives, we define an objective function

$$\lambda Loss + (1-\lambda)(\epsilon PPV_{diff} + (1-\epsilon)EO_{diff})$$
$$\lambda = \epsilon = 0.5$$

where loss is the binary cross-entropy loss.

To train a MOO problem with just counterfactual fairness as a fairness objective, we define an objective function:

$$\lambda Loss + (1-\lambda)CF$$
$$\lambda = 0.5$$

where again, our loss is the binary cross-entropy loss.

Results

Analysis of Existing COMPAS Score

We use an explainable boosting machine(EBM) to determine the influential features that affect the COMPAS scores in our dataset. Results show that race is an influential factor to recidivism predictions, indicating that existing COMPAS scores are racially biased.

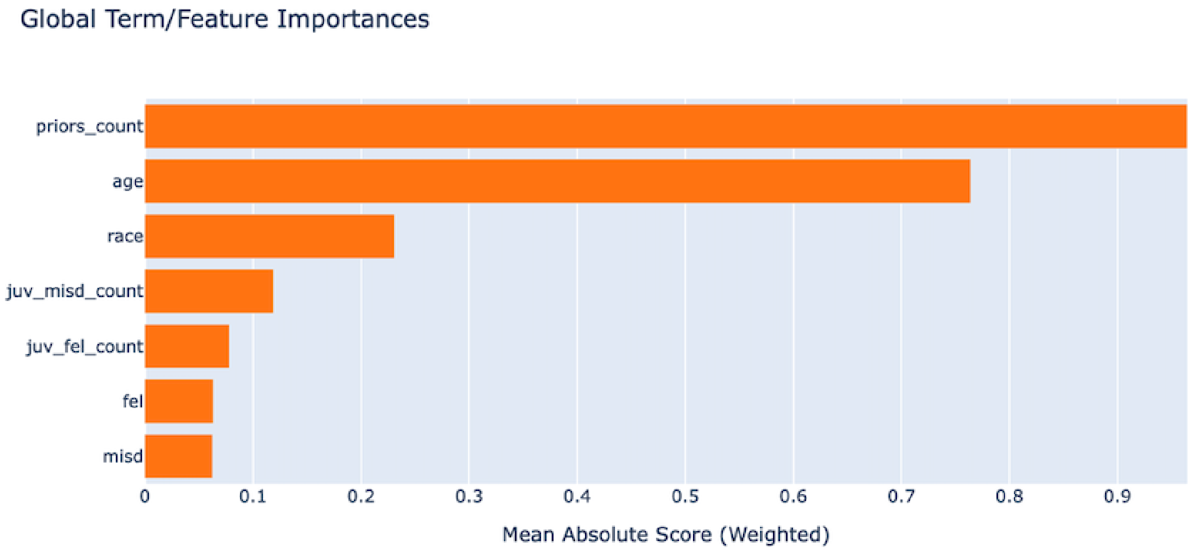


Figure 2: Feature importance of the existing COMPAS score measuring by Explainable Boosting Machine(EBM)

Comparison of Results and Discussion

We first compare the relative performance of our four models

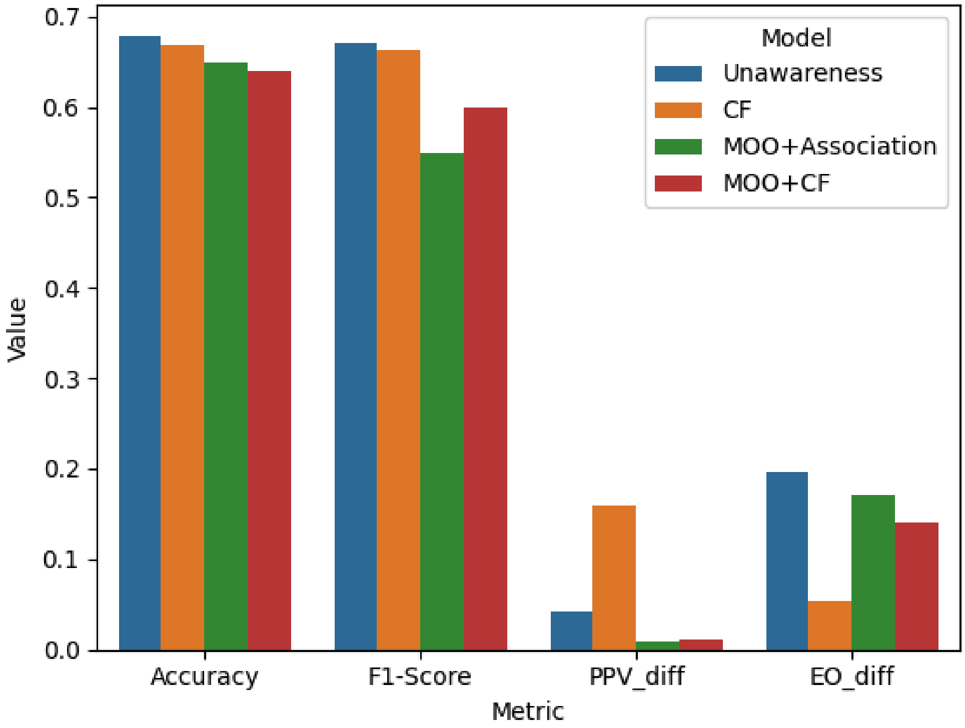


Figure 3: Comparison of results for different models: 1) Unawareness 2) Counterfactual Fairness(CF) 3) Muti-Objective Optimization with two association metrics(MOO+Association) 4) Muti-Objective Optimization with counterfactual fairness(MOO+CF)

We note two things from our results. First, compared with our baseline Unawareness model, using MOO with associational fairness objectives led to a tradeoff in accuracy and fairness — we sacrifice predictive performance for improved equalized odds and predictive parity. However, while we were able to reduce the difference in predictive parity to near-zero, we weren't able to satisfy equalized odds.

This supports existing literature which show equalized odds and predictive parity are at odds with one another. Second, while our proposed MOO+CF model was able to further minimize the difference in equalized odds, we also weren't able to satisfy its definition. However, with relatively similar predictive performance, we argue that our proposed model is at least a simpler alternative to existing MOO approaches.

Then, in line with ProPublica's analysis, we compare the per-race false positive and false negative rates (FPR and FNR) across our four models to see how much fairer our recidivism classifications are compared to the original COMPAS labels. In COMPAS' case, we see that African Americans had higher false positive rate (meaning they were wrongly classified as likely to recidivate more frequently) while Caucasians had a higher False Negative rate (meaning they were wrongly classified as not likely to recidivate more frequently). **A fair model would misclassify any individual from each demographic at an equal rate.** This is compatible with the definition of equalized odds.

	African-American		Caucasian	
	FP	FN	FP	FN
COMPAS	0.45	0.28	0.23	0.48
Unawareness	0.32	0.40	0.15	0.60
CF	0.22	0.41	0.27	0.47
MOO+Association	0.21	0.49	0.06	0.69
MOO+CF	0.18	0.57	0.07	0.71

Tabel 1: Comparison of FP and FN for African-American and Caucasian for 5 recidivism labels: 1) COMPAS 2) Fairness through Unawareness(Unawarenes) 3) *Counterfactual Fairness(*CF) 4*) Multi-Objective Optimization with two associational fairness metrics(MOO+Associational) 5) Multi-Objective Optimization with counterfactual fairness(MOO+CF)*

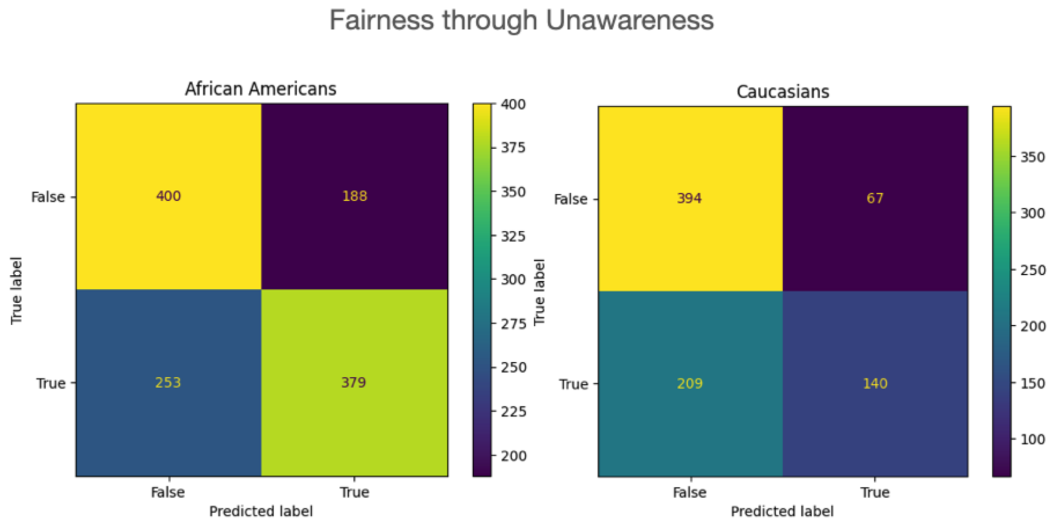


Figure 4(a): Confusion matrix for Fairness through Unawareness model

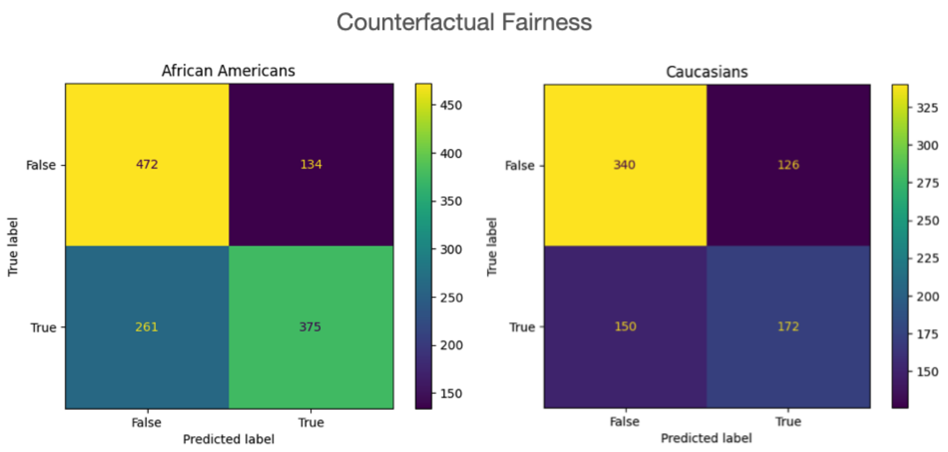


Figure 4(b): Confusion matrix for Counterfactual Fairness model

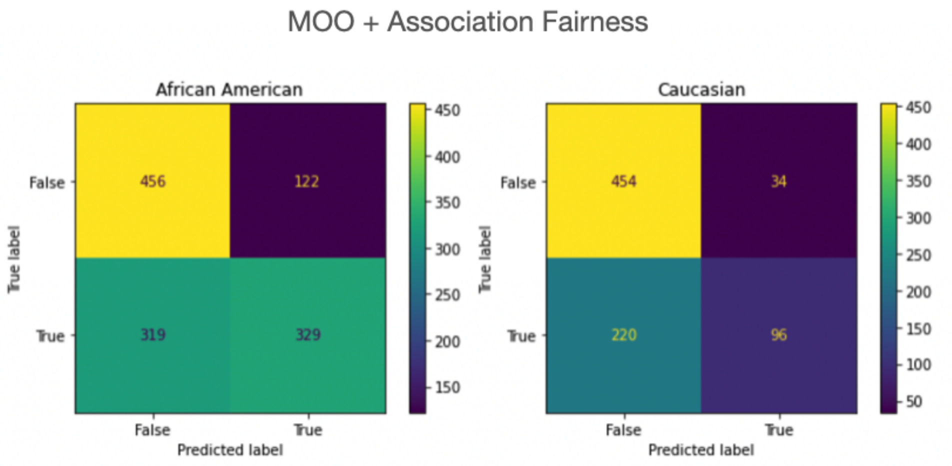


Figure 4(c): Confusion matrix for MOO+Associational Fairness model

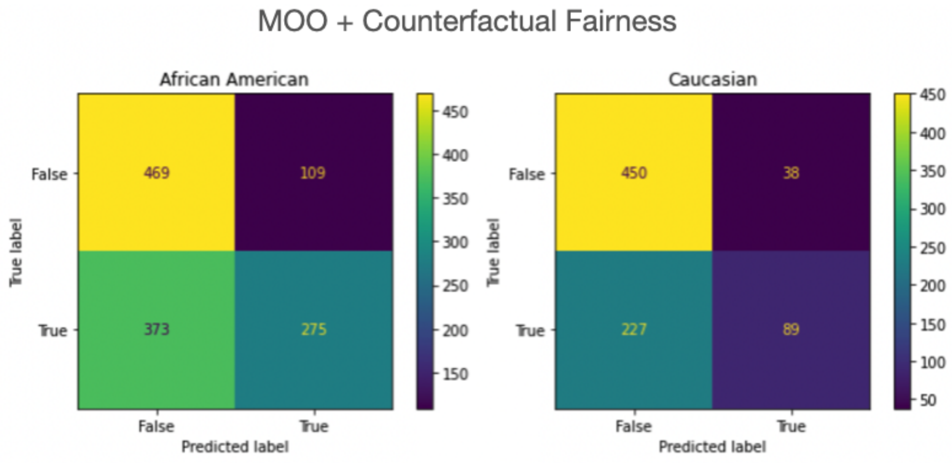


Figure 4(d): Confusion matrix for MOO+Associational Fairness model

Therefore, we want to minimize the difference in FPR and FNR between the protected and unprotected classes. To this end, a simple counterfactual fairness algorithm is best — it has the smallest difference in FPR and FNR between the protected and unprotected groups. However, while our counterfactual fairness algorithm optimizes for equalized odds, it doesn't do a great job minimizing predictive parity, as shown in Figure 3.

Among multi-objective optimization approaches, we note that the FNR for the protected and unprotected groups is the highest in our proposed MOO+CF model. However, the *difference* in FPR and FNR is notably smaller compared to the MOO+Association model. Therefore, we can conclude that to train a fairer model, using counterfactual fairness as a single objective may be better. However, it comes at the expense of accuracy.

Conclusion

In this research project, we analyze the COMPAS dataset to determine the most effective fair machine learning strategies. We first demonstrated that racial bias does exist in the original COMPAS system by evaluating global feature importance. We then compared popular fair machine learning methods such as counterfactual fairness, fairness through unawareness, and MOO with associational fairness objectives. We also proposed a multi-objective optimization model with counterfactual fairness as the single fairness objective.

Comparing predictive parity and equalized odds of our four models, we determine that the best way to simultaneously optimize both fairness metrics is our proposed MOO+CF model. However, this does come with a tradeoff in predictive performance.

Comparing the race-wise FPR and FNR of the biased COMPAS labels with our model predictions, we determine that a counterfactual fairness model does the best job mitigating the difference in FPR and FNR across the protected/unprotected groups. However, much like how ProPublica's existing analysis was scrutinized for only considering equalized odds, the counterfactual fairness model also fails to account for predictive parity.

In conclusion, we determine that counterfactual fairness cannot be the singular fairness objective to jointly optimize for conflicting fairness metrics. Not only does our proposed MOO+CF model fail to satisfy both fairness metrics, but a counterfactual fairness model guarantees optimal equalized odds at the expense of predictive parity. While counterfactual fairness takes a step towards the right direction, it cannot be the universal solution practitioners should rely on.

Division of Labor

Literature Review: Jason Lee and Keyu Li
MOO theory: Jason Lee
Analysis of COMPAS score: Keyu Li
Experiments:

1) Fairness through unawareness: Keyu Li
    a) Logistic Regression
    b) Support Vector Machine
    c) Explainable Boosting Machine
2) Counterfactual Fairness: Keyu Li
    a) Logistic Regression
3) MOO + Association: Jason Lee
    a) Logistic Regression
4) MOO + Counterfactual Fairness: Jason Lee
    a) Logistic Regression

Visualization: Keyu Li
Paper writing: Jason Lee and Keyu Li

References

[1] Yu, Guo, et al. "Towards Fairness-Aware Multi-Objective Optimization." *arXiv preprint arXiv:2207.12138* (2022).

[2] Kusner, Matt J., et al. "Counterfactual fairness." *Advances in neural information processing systems* 30 (2017)

[3] Loftus, Joshua R., et al. "Causal reasoning for algorithmic fairness." *arXiv preprint arXiv:1805.05859* (2018).

[4] Zhang, Qingquan, et al. "Fairer machine learning through multi-objective evolutionary learning." *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part IV*. Cham: Springer International Publishing, 2021/

[5] Russell, Chris, et al. "When worlds collide: integrating different counterfactual assumptions in fairness." *Advances in neural information processing systems* 30 (2017).