CausalFairnessInAction : A Novel Open Source Python Library For Causal Fairness Analysis

A Practical Guide to Diagnosing Bias with Causal Fairness
Forthcoming at: https://github.com/amazon-science/causal-fairness-in-action

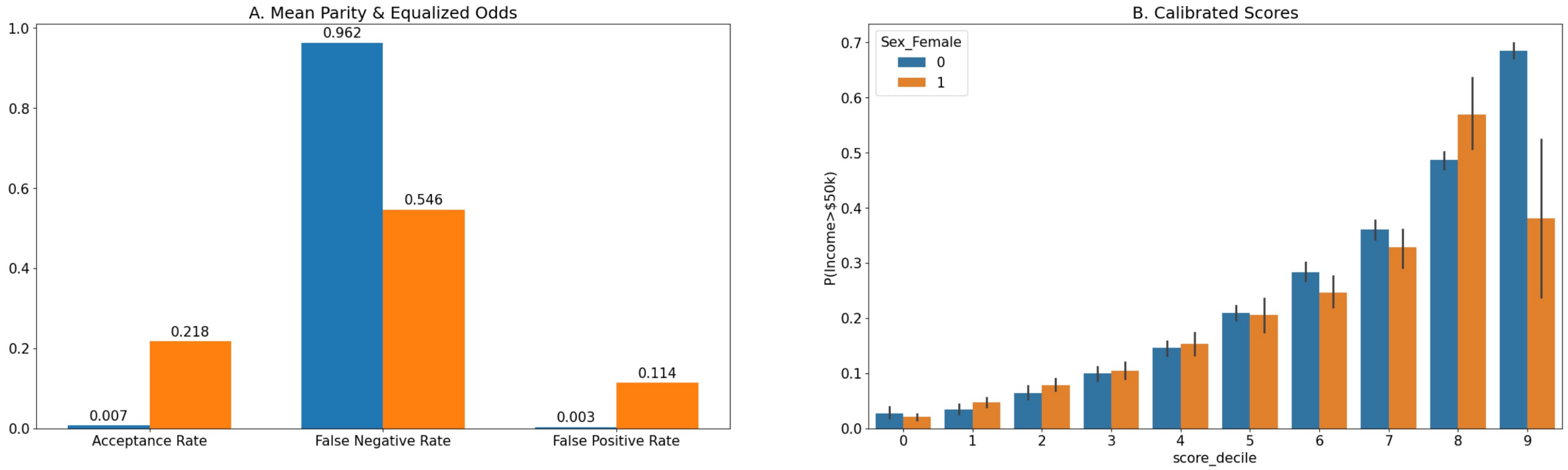
Kriti Mahajan, Amazon
kritimhj@amazon.com

Motivation

Our model shows disparity. But is it discrimination?

- As machine learning enters high-stakes domains, assessing fairness becomes vital
- Statistical fairness metrics are widely used : They tell us *what* disparities exist in a model's outcome

Figure 1:Statistical Fairness Metrics (Ex: Adult Income Dataset)



Commonly used statistical fairness metrics :

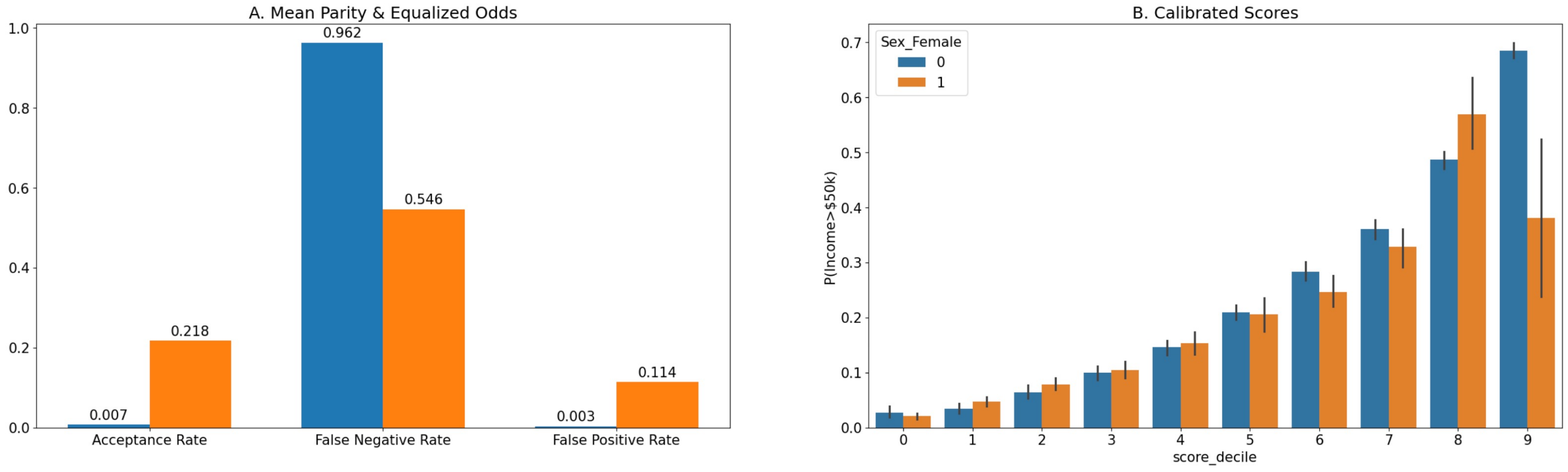
- Mean Parity
- Equalized Odds
- Calibrated Scores

Motivation

Our model shows disparity. But is it discrimination?

- But the typically used statistical fairness metrics have a key limitation: **They are associations - conditional probabilities - thus cannot explain why these disparities occur**
- Need methods that identify *why* disparities occur

Figure 1:Statistical Fairness Metrics (Ex: Adult Income Dataset)



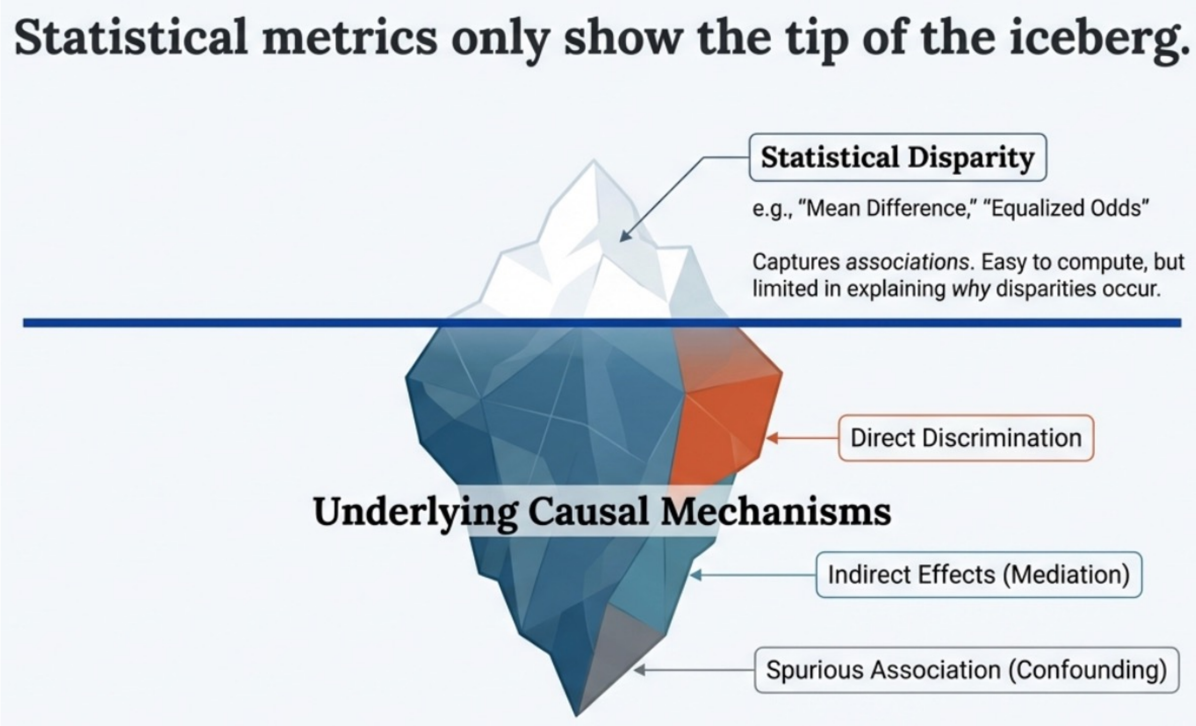
Commonly used statistical fairness metrics :

- Mean Parity
- Equalized Odds
- Calibrated Scores

Motivation

Our model shows disparity. But is it discrimination?

- **Causal Fairness metrics** based on Structural Causal Models (SCMs) attribute observed disparities to specific sources - protected attributes, mediators or confounders
- But they have limited adoption due to **technical** & **computational complexity.**



Contribution : CausalFairnessInAction

The first open-source Python package for computing causal fairness metrics

CausalFairnessInAction implements generalizable algorithms for three key metrics in the causal fairness literature:

- Counterfactual Effects for Mean Parity

Plecko, D. &; Bareinboim, E.,2024. "Causal fairness analysis.". In: Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 1–238

- Counterfactual Equalized Odds

Zhang, J. &; Bareinboim, E., 2018. "Equality of opportunity in classification: A Causal approach." In: Advances in Neural Information Processing Systems.

- Counterfactual Fairness

Kusner, M.J. et al., 2017. "Counterfactual fairness" In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems

Contribution : CausalFairnessInAction

The first open-source Python package for computing causal fairness metrics

- **Practical:**
  - Applicable across classification and regression tasks
  - Designed to work with minimal identifiability constraints
  - Doesn't require fully specified SCMs

- **Comprehensive:**
  - Computes metrics at both **group** & **individual** levels
  - Supports **intersectional analysis**

- **Efficient:** Optimized for scalability using Gaussian Mixture Models, parallelization to reduce latency

Enables **actionable audits** by decomposing statistical fairness metrics into causal components.

A Brief Literature Review

Broadly, there are three branches in the causal fairness literature:

- **Counterfactual measurement** (focus of this package) : helps answering what-if cause-effect questions without running randomized control trials.
- **Sensitivity analysis** : how sensitive a model is to latent / confounding variables
- **Impact evaluation :** measures the long-term consequences of automated decision-making systems through the use of interventions.

What gets measured gets managed - causal identification of discrimination is crucial before moving on to remedial actions and impact analysis.

Methodology & Framework

The Standard Fairness Model

The `CausalFairnessDecomposition` class is built on the standard fairness model (Zhang, J. &; Bareinboim, E., 2018):

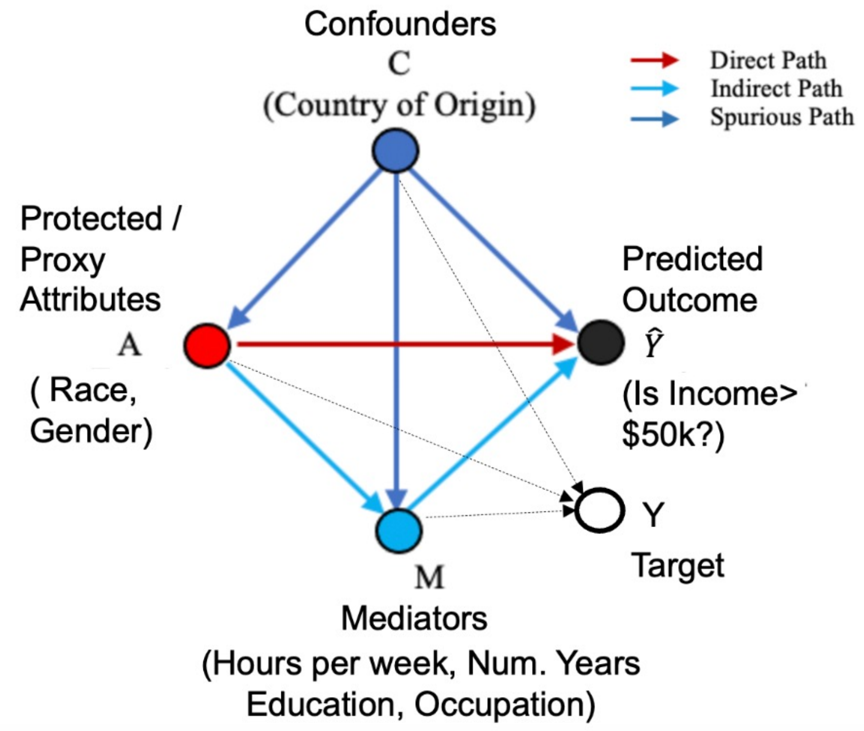


Figure 2: Standard Fairness Model (Ex: Adult Income Dataset; source - Zhang, J. &; Bareinboim, E., 2018)

- DAG includes protected attribute A, mediators M, confounders C, true outcome Y, prediction Ŷ

- Three paths from A → Y / Ŷ:
  1) **Direct (Causal Pathway)**: A → Y/Ŷ (disparate treatment)
  2) **Indirect (Causal Pathway)**: A → M → Y/Ŷ (mediated disparate impact)
  3) **Spurious (Non-Causal Pathway)**: : A ← C → Y/Ŷ (confounding: spurious correlation)

Implemented Metrics and Methods

`CausalFairnessDecomposition` class has 3 methods, each corresponding to a different causal fairness metric in the literature

Counterfactual Effects for Mean Parity

Group Level

1. `analyse_mean_difference`

Question Addressed →

What would a group's acceptance rate be if they had the identity, mediators, or confounders of another group?

Decomposes

Statistical Parity into Counterfactual Direct Effect (Ctf-DE), Indirect Effect (Ctf-IE), and Spurious Effect (Ctf-SE).

Counterfactual Equalized Odds

Group Level

2. `analyse_equalized_odds`

Question Addressed →

How would a group's error rate (FPR/FNR) change under counterfactual conditions?

Decomposes

Equalized Odds into Counterfactual Direct and Spurious Error Rates.

Counterfactual Fairness

Individual Level

3. `analyse_counterfactual_fairness`

Question Addressed →

If we changed an *individual's* protected attribute, would their predicted outcome change?

Measures

Whether a prediction is counterfactually fair for each person in the dataset.

Implemented Metrics and Methods

Counterfactual Effects : Calculating Disparate Treatment, Disparate Impact and Explaining the Causal Mechanism Behind Statistical Parity

- **Counterfactual Direct Effect (Ctf-DE):** measures direct discrimination along $A \rightarrow \hat{Y}$ by holding M and C constant. **Symmetric Ctf-DE** is the difference between the positive and negative effect of protected group membership. **Direct discrimination exists if Symmetric Ctf-DE > 0**
  → **disparate treatment**
- **Counterfactual Indirect Effect (Ctf-IE):** measures indirect discrimination along $A \rightarrow M \rightarrow \hat{Y}$ by holding A and C fixed. **Symmetric Ctf-IE** is the difference between the positive and negative effect of having the group's meditating characteristics. **Indirect discrimination exists if Symmetric Ctf-IE > 0**
  → **disparate impact**
- **Counterfactual Spurious Effect (Ctf-SE):** measures confounding impact along $A \leftarrow C \rightarrow \hat{Y}$, varying C while fixing A and M
  → **disparate impact**

Implemented Metrics and Methods

Counterfactual Effects : Calculating Disparate Treatment, Disparate Impact and Explaining the Causal Mechanism Behind Statistical Parity

Mean Difference Causal Decomposition: Using these three counterfactual metrics, (Plecko, D. &; Bareinboim, E.,2024) show that mean difference can be broken down into direct, indirect and spurious components as follows

Mean Difference = Symmetric Ctf-DE + Symmetric Ctf-IE + Ctf-SE

Implemented Metrics and Methods

Counterfactual Effects : Pseudo-algorithm for `analyse_mean_difference`

(Zhang, J. &; Bareinboim, E., 2018) provide empirical formulas for estimation from observed data using conditional probabilities

→ **fully specified SCM not needed**

For each combination m ∈ M and c ∈C, we get a subset of D defined by (m, c).

- For each (m, c) compute the expected outcome **E(y | a, m, c)** for $\mathbf{a_0}$, $\mathbf{a_1}$.
- For each **m**, calculate probability of **m** when c is fixed under $\mathbf{a_0}$, $\mathbf{a_1}$: $\mathbf{P(m \mid a_0, c)}$, $\mathbf{P(m \mid a_1, c)}$
- For each **c**, calculate probability of **c** under $\mathbf{a_0}$, $\mathbf{a_1}$: $\mathbf{P(c \mid a_0)}$, $\mathbf{P(c \mid a_1)}$.

Inputs: D, A, M, C, a_0, a_1, y

1. For each $(m, c) \in D$:
  - Compute: $\mathbb{E}(Y = y \mid a_0, m, c)$
  - Compute: $\mathbb{E}(Y = y \mid a_1, m, c)$

2. Estimate via GMM:
  $P(m \mid a_0, c), P(m \mid a_1, c)$
  $P(c \mid a_0), P(c \mid a_1)$

3. Combine expectations and probabilities to compute the counterfactual effects

Implemented Metrics and Methods

Counterfactual Equalized Odds

- **Counterfactual Direct Error Rate**
- **Counterfactual Indirect Error Rate**
- **Counterfactual Spurious Error Rate**

Equalized Odds Causal Decomposition: Using these three counterfactual error metrics, (Zhang, J. &; Bareinboim, E., 2018) show that equalized odds can be broken down into direct, indirect and spurious components as follows:

Equalized Odds = Counterfactual Direct Error Rate + Counterfactual Indirect Error Rate + Counterfactual Spurious Error Rate

Implemented Metrics and Methods

Counterfactual Equalized Odds : Pseudo-algorithm for `analyse_equalized_odds`

Limitation: cannot reliably estimate direct, indirect, and spurious effects in the presence of mediators due to lack of identifiability from conditioning on both Y and Ŷ

Solution : **Refit estimator without M to get direct & spurious effects but no indirect effect**

- For each c ∈ C use the fitted estimator to compute predicted outcomes under $\mathbf{a_0}$ and $\mathbf{a_1}$: $\hat{\mathbf{y}}_{\mathbf{a_0},\mathbf{c}}$ and $\hat{\mathbf{y}}_{\mathbf{a_1},\mathbf{c}}$.
- For each **c**, compute how its probability under $\mathbf{a_0}$, $\mathbf{a_1}$: $\mathbf{P(c \mid a_0, y)}$, $\mathbf{P(c \mid a_1, y)}$.

Inputs: $D, A, C, a_0, a_1, y, \hat{f}$

1. For each $c_j \in D$:
  - Predict: $\hat{f}(c_j, a_0), \hat{f}(c_j, a_1)$
  - Obtain: $P(\hat{y}_{a_0, c_j}), P(\hat{y}_{a_1, c_j})$

2. Estimate via GMM:
  $P(c \mid a_0), P(c \mid a_1)$

3. Combine predictions and probabilities to compute the Cft-EO

Implemented Metrics and Methods

Counterfactual Fairness

Counterfactual Fairness is an individual level metric which is achieved if changing an individual i’s protected attributes doesn’t change the individual’s predicted outcome

Counterfactual Fairness Decomposition: Not supported

Implemented Metrics and Methods

Counterfactual Fairness : Pseudo-algorithm in `analyse_counterfactual_fairness`

Define a structural causal model (SCM) and get observed $\mathbf{A_{obs}}$ and counterfactual $\mathbf{A_{cf}}$ for i ∈ D

- Generate two datasets by intervening on **A**:
  do(A = A$_{obs}$) → observed world
  do(A = A$_{cf}$) → counterfactual world
- Fit the model on each dataset to get predictions $\mathbf{\hat{Y}_{obs}}$ and $\mathbf{\hat{Y}_{cf}}$
- The model is **not counterfactually fair** if $\mathbf{\hat{Y}_{obs} \neq \hat{Y}_{cf}}$

Inputs: A, M, C, a_0, a_1, DAG
1. Fit SCM using DAG and dataset D **2.** For each individual $i \in D$: - Get A_{obs} (observed) and A_{cf} (counterfactual) - Sample from SCM under: $do(A = A_{obs}) \Rightarrow D_{obs}$ $do(A = A_{cf}) \Rightarrow D_{cf}$ - Predict: $\hat{f}(D_{obs})$, $\hat{f}(D_{cf})$ - Check: $Y_{obs} \neq Y_{cf}$

Implemented Metrics and Methods

Example API Call Applied To The Adult Income Dataset

```
cfd = CausalFairnessDecomposition(**{"X": X_train,
                                     "y_true" : y_train.values,
                                     "y_pred": X_train["prediction"],
                                     "model": trained_logistic_regression_classifier,
                                     "protected_attr": ["Sex_Female"],
                                     "mediators": ([x for x in X_train.columns if "Occupation_" in x]
                                                   + ["EducationNum"]
                                                   + ["HoursPerWeek"]),
                                     "confounders": [x for x in X_train.columns if "Country_" in x],
                                     "yi": 1,
                                     "advantage_group": 0,
                                     "disadvantage_group": 1,
                                     "continuous": False})
mean_diff_decomposition = cfd.analyse_mean_difference(how='decompose')
fig_md, ax_md = plot_mean_diff_waterfall(mean_diff_decomposition,mean_difference)
```

Implemented Metrics and Methods (Contd.)

Example API Call Applied To The Adult Income Dataset

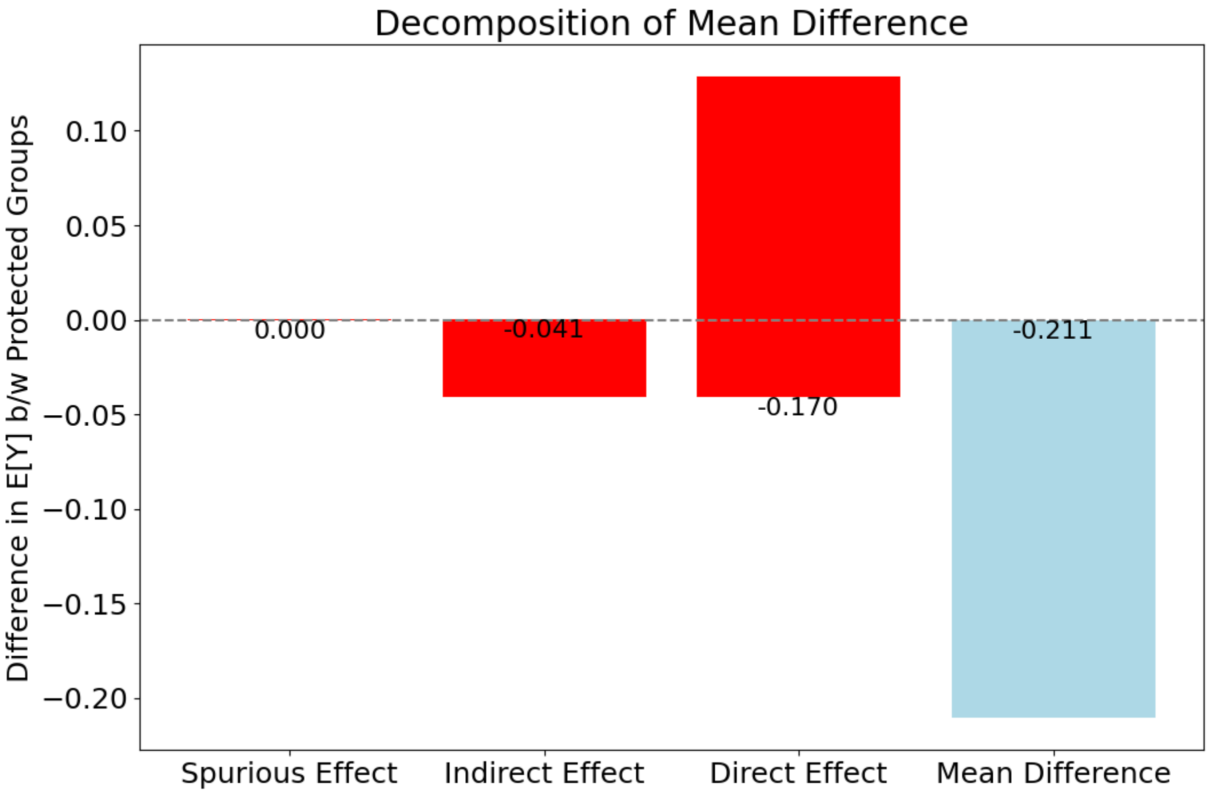


Figure 3: Counterfactual Effects Decomposition of Mean Difference for Adult Income Dataset

Mean Difference = Symmetric Ctf-DE + Symmetric Ctf-IE + Ctf-SE

Implemented Metrics and Methods

Computational Optimizations

- GMMs for conditional probability estimation
- Parallelization
- No specialized hardware requirement

NEURAL INFORMATION PROCESSING SYSTEMS

Application To Benchmark Datasets

Overview of Findings

We benchmarked the library on 3 datasets : **Adult Income, COMPAS, LSAC**

- **Direct discrimination is the primary contributor** to mean difference and equalized odds across all 3 datasets
- The classifier for Adult Income, COMPAS is not counterfactually fair but is counterfactually fair for LSAC i.e. **group fairness can differ from individual fairness**
- **Intersectional Analysis (Race x Sex) worsens direct discrimination** across all three datasets

Table 1: CausalFairnessInAction Benchmarking Results

Dataset	Protected Attribute	Mean Difference	FNR	FPR	$\mathrm{DE}_a^{\mathrm{sym}}(y\|a)$	$\mathrm{IE}_a^{\mathrm{sym}}(y\|a)$	$\mathrm{SE}_{a_0,a_1}(y\|a)$	ER^d	ER^i	ER^s	Counterfactual Fairness
Adult Income	Gender	0.203	0.410	-0.104	0.165	0.039	0.000	0.000	0.000	0.000	-0.031
Adult Income	Intersectional	0.221	0.445	-0.115	0.152	0.069	0.000	0.000	0.000	0.000	-0.068
COMPAS	Race (Black)	0.326	-0.310 (-42)	-0.253 (-0.41)	0.154	0.071	0.101	FPR: -0.297, FNR: -0.265	0	FPR: 0.113, FNR: 0.162	0.055
COMPAS	Intersectional	0.620	-0.620	-0.518	0.513	0.081	0.027	-	-	-	0.640
LSAC	Race (Black)	0.978	-	-	0.554	0.429	0.000	-	-	-	0.001
LSAC	Intersectional	0.990	-	-	0.531	0.458	0.000	-	-	-	-0.007

Application To Benchmark Datasets

Adult Income Dataset

Task : logistic regression fit to predict
P(Income > \$50k)

- **Counterfactual Effects:** On average, women are 20.3% less likely than men to be predicted as earning above \$50k. Most of this disparity (16.5%) is due to disparate treatment (direct effect)
- **Counterfactual Equalized Odds:** Not identifiable due to mediator issues
- **Counterfactual Fairness:** Not counterfactually fair—changing female→male increases predicted P(Income > \$50k)

DAG & Variables used for Fitting `CausalFairnessDecomposition`

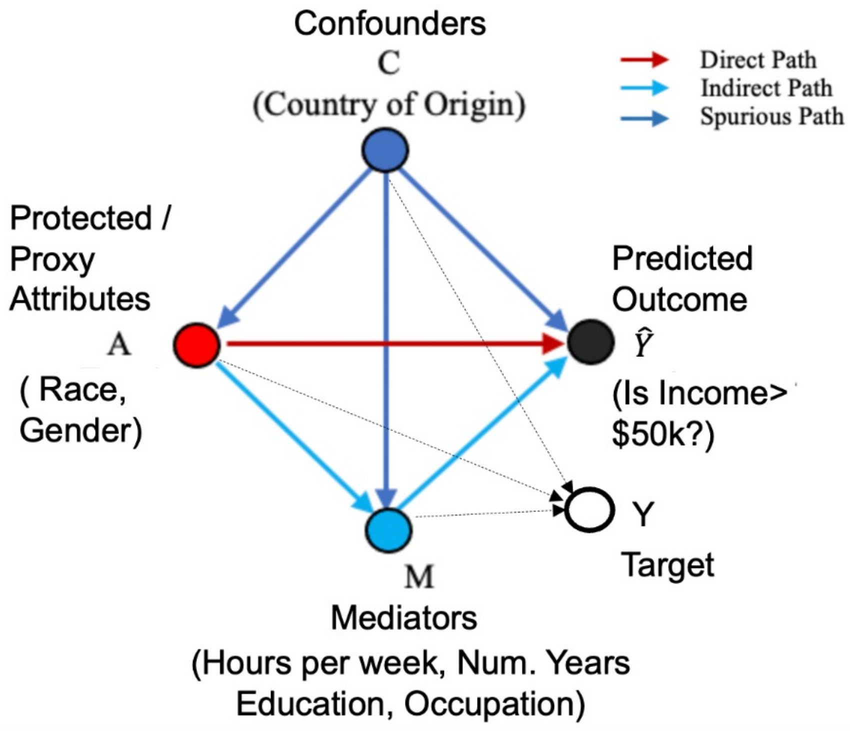


Figure 2: Adult Income Standard Fairness Model (Zhang, J. &; Bareinboim, E., 2018)

Application To Benchmark Datasets

COMPAS Dataset

Task : logistic regression fit to predict P(Recidivism)

- **Counterfactual Effects:** Black individuals are 32.6% more likely than white individuals to be predicted as high-risk for recidivism. Majority is attributed to disparate treatment (15.4%). Disparate impact comes from both M and C: confounders raise risk by~10% (spurious effect); M contributes an additional 7.1% (indirect effect)
- **Counterfactual Equalized Odds:** Excluding M does not make the model naive, though it increases error rates. Decomposing FPR/FNR shows most of the disparity stems from direct discrimination: 29.7% of the 41% FPR and 26.5% of the 42% FNR
- **Counterfactual Fairness:** Not counterfactually fair—changing black → white decreases predicted P(Recidivism)

DAG & Variables used for Fitting `CausalFairnessDecomposition`

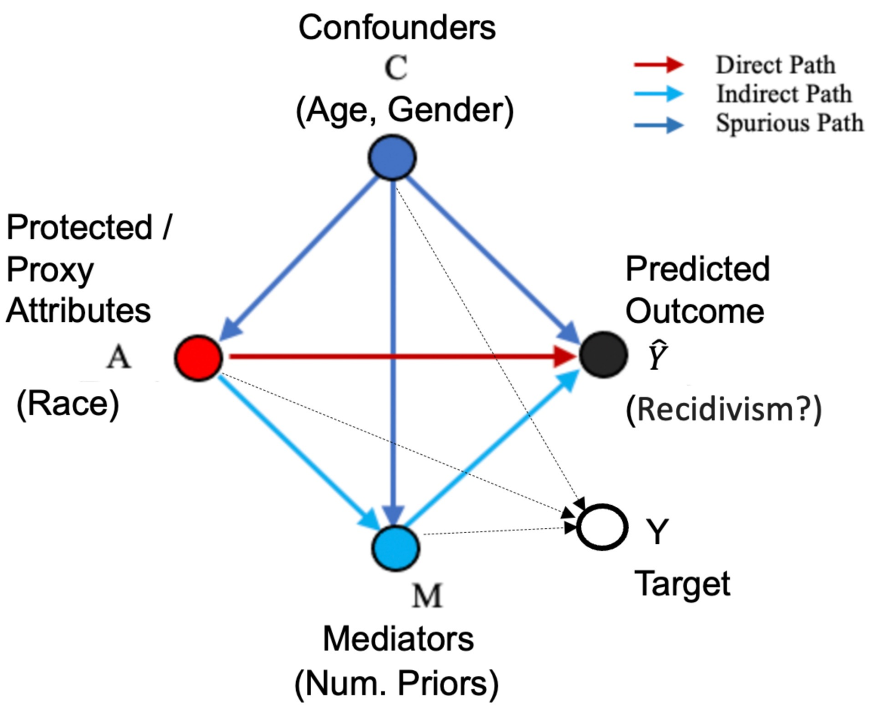


Figure 4: COMPAS Standard Fairness Model (Zhang, J. &; Bareinboim, E., 2018)

Application To Benchmark Datasets

LSAC Dataset

Task : fit Random Forest regressor to **predict** average grade

- **Counterfactual Effects:** The predicted average grade for the white subgroup is 0.978 higher than for the Black subgroup with majority of the gap (0.55) due to direct discrimination.
- **Counterfactual Equalized Odds:** Not applicable since this is a regression task
- **Counterfactual Fairness:** The model is counterfactually fair, illustrating that fairness can differ at the individual vs. group level

DAG & Variables used for Fitting `CausalFairnessDecomposition`

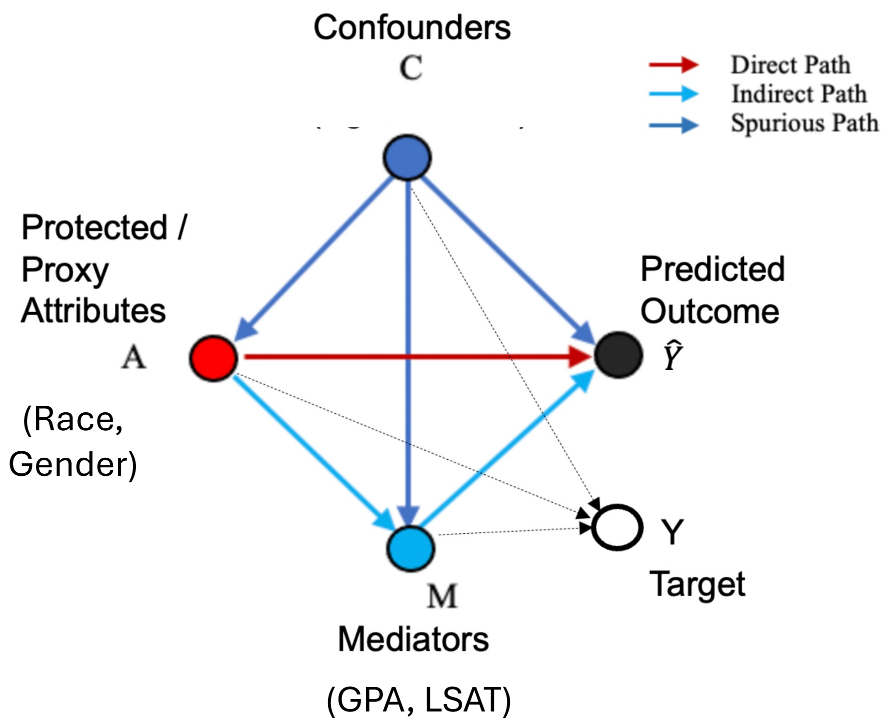


Figure 5: LSAC Standard Fairness Model (Kusner, M.J. et al., 2017.)

Application To Benchmark Datasets

Intersectional Analysis

- **Adult Income:** Black women are 22.1% less likely than white men to have P (Income > $50k) —2% more than the non-intersectional gender gap— with direct effect being the largest contributor. The model is also more counterfactually unfair: changing a Black woman's identity to a white man increases P (Income > $50k) by 6% (vs. 2% non-intersectionally)
- **COMPAS:** The mean difference in P (Recidivism) between Black men and white women (60%) exceeds the non-intersectional racial gap, with direct discrimination as the main driver. Counterfactual unfairness also rises by~11%: changing a Black man's identity to a white woman lowers predicted recidivism by 64% (vs. 55%)
- **LSAC**: The mean difference between Black women and white men is slightly higher than the non-intersectional comparison (0.99 vs. 0.978), mainly due to direct discrimination. As in the non-intersectional case, the model remains counterfactually fair.

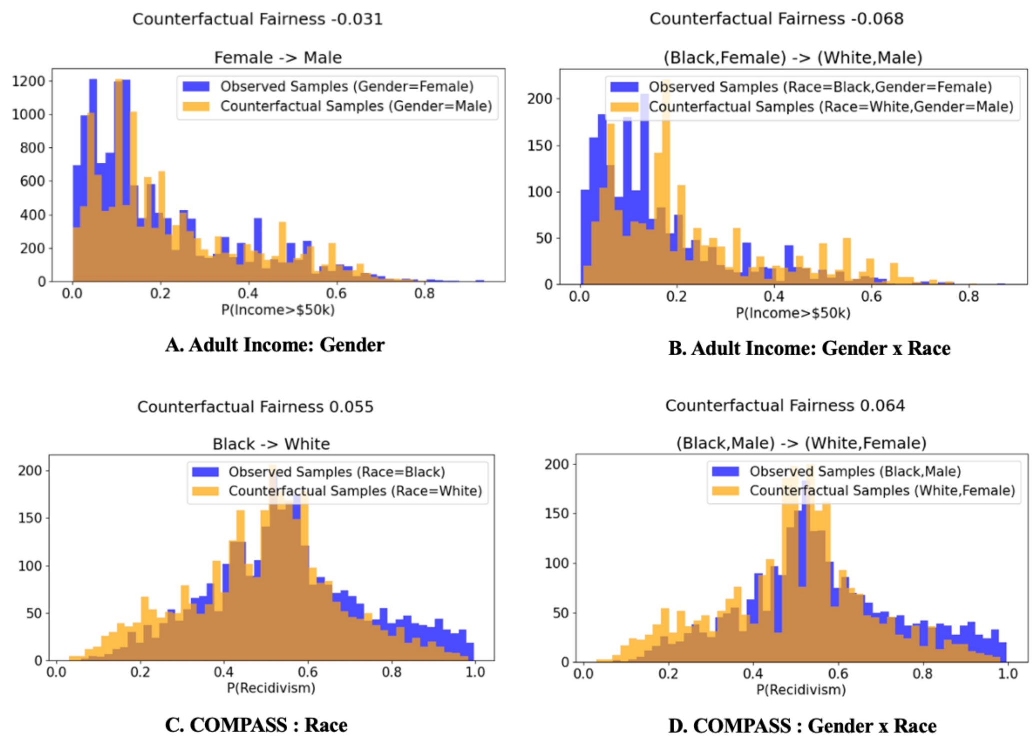


Figure 6: Counterfacual Fairness

Limitations of CausalFairnessInAction

- **Lack of identifiability can limit analysis**
  - Ex: in the Adult Income dataset, identifiability issues prevented the causal decomposition of equalized odds
- **Lack of methods for falsifying DAGs** in the presence of competing causal models can lead to disagreements about result validity
- **Defining a hypothetical intervention on protected attributes** remains a fraught process

Future Work

- Use package results to guide fairness interventions:
  - Feature selection
  - Sample-level reweighting
  - Multi-world regularization

- Integrate bias-reduction algorithms into package

Conclusion

We introduced ***CausalFairnessInAction*** - the first open-source generalizable implementation for calculating key causal fairness metrics

- Applied it to 3 fairness benchmarking datasets

Demonstrated how CausalFairnessInAction provides practitioners with the actionable insight

- Ex : at the very least the Adult Income model must eliminate at least 16.5% difference in statistical parity, while the COMPAS model needs to address 15.4% disparity in statistical parity and 29.7-26.5% in error rates (all of which can be attributed to direct discrimination)

Thank you!

Explore the library, replicate these findings, and apply causal analysis to your own models.

github.com/amazon-science/causal-fairness-in-action

**Forthcoming*

Appendix

CausalFairnessInAction: An Open Source Python Library for Causal Fairness Analysis

Kriti Mahajan, Amazon, kritimhj@amazon.com
Forthcoming at:
https://github.com/amazon-science/causal-fairness-in-action

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Contribution

The Problem: As machine learning enters high-stakes domains, assessing fairness becomes vital—but the typically used statistical fairness metrics have a key limitation: They are **associations(conditional probabilities) thus, they** can state **what the observed disparity is *but not why* it exists.**

Causal Fairness metrics solve this by using Structural Causal Models (SCMs) to uncover generating mechanisms, but have limited adoption due to **technical** & **computational complexity.**

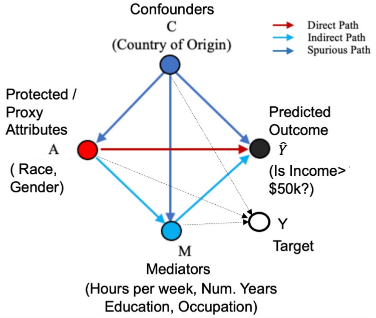
The Solution: CausalFairnessInAction, the first open-source Python package for computing diverse causal fairness metrics, enabling actionable audits by decomposing statistical disparities into causal components.

- **Practical:** applicable across classification and regression tasks ; designed to work with minimal identifiability constraints; doesn't require fully specified SCMs.
- **Comprehensive:** Computes metrics at both **group** & **individual** levels; supports **intersectional analysis**.
- **Efficient:** Optimized for scalability using Gaussian Mixture Models, parallelization to reduce latency

Methodology & Framework

The `CausalFairnessDecomposition` class is built on the standard fairness model [1]:

Figure 1:Standard Fairness Model (Ex: Adult Income Dataset)



Three Implemented Metrics and Methods

`analyse_mean_difference` ➔ **Implements Counterfactual Effects[1]**

Query : What would the disadvantaged (advantaged) **group's acceptance rate** be if they had the identity (A), mediating characteristics (M), or confounding characteristics (C) of the advantaged (disadvantaged) group?

Supported Decompositions: Direct, Indirect, Spurious

`analyse_equalized_odds` ➔ **Implements Counterfactual Equalized Odds[2]**

Query : What would the disadvantaged (advantaged) **group's error rate** be if they had A, M or C of the advantaged (disadvantaged) group?

Supported Decompositions: Direct, Spurious

`analyse_counterfactual_fairness` ➔ **Implements Counterfactual Fairness[3]**

Query : What would the disadvantaged (advantaged) **individual's predicted Y** be if they had the A, M, and C of the advantaged (disadvantaged) group?

Supported Decompositions: N/A

Table 1: Pseudo-Algorithms for Causal Fairness Metrics

`analyse_mean_difference`	`analyse_equalized_odds`	`analyse_counterfactual_fairness`
Inputs: D, A, M, C, a_0, a_1, y	**Inputs:** $D, A, C, a_0, a_1, y, \hat{f}$	**Inputs:** A, M, C, a_0, a_1, DAG
1. For each $(m, c) \in D$: - Compute: $\mathbb{E}(Y = y \mid a_0, m, c)$ - Compute: $\mathbb{E}(Y = y \mid a_1, m, c)$ **2.** Estimate via GMM: $P(m \mid a_0, c), P(m \mid a_1, c)$ $P(c \mid a_0), P(c \mid a_1)$ **3.** Combine expectations and probabilities to compute the counterfactual effects	**1.** For each $c_j \in D$: - Predict: $\hat{f}(c_j, a_0), \hat{f}(c_j, a_1)$ - Obtain: $P(\hat{y}_{a_0,c_j}), P(\hat{y}_{a_1,c_j})$ **2.** Estimate via GMM: $P(c \mid a_0), P(c \mid a_1)$ **3.** Combine predictions and probabilities to compute the Cft-EO	**1.** Fit SCM using DAG and dataset D **2.** For each individual $i \in D$: - Get A_{obs} (observed) and A_{cf} (counterfactual) - Sample from SCM under: $do(A = A_{obs}) \Rightarrow D_{obs}$ $do(A = A_{cf}) \Rightarrow D_{cf}$ - Predict: $\hat{f}(D_{obs}), \hat{f}(D_{cf})$ - Check: $Y_{obs} \neq Y_{cf}$

Code block 1: Example API Call Applied To The Adult Income Dataset

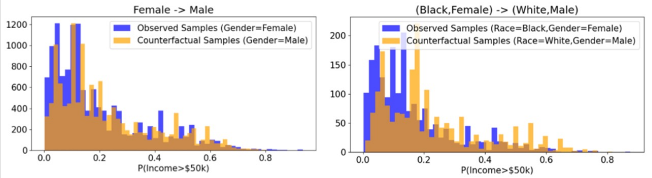
```python
cfd = CausalFairnessDecomposition(**{"X": X_train,
                                     "y_true" : y_train.values,
                                     "y_pred": X_train["prediction"],
                                     "model": trained_logistic_regression_classifier,
                                     "protected_attr": ["Sex_Female"],
                                     "mediators": ([x for x in X_train.columns if "Occupation_" in x]
                                                   + ["EducationNum"]
                                                   + ["HoursPerWeek"]),
                                     "confounders": [x for x in X_train.columns if "Country_" in x],
                                     "yi": 1,
                                     "advantage_group": 0,
                                     "disadvantage_group": 1,
                                     "continuous": False})
mean_diff_decomposition = cfd.analyse_mean_difference(how='decompose')
fig_md, ax_md = plot_mean_diff_waterfall(mean_diff_decomposition,mean_difference)
```

Application To Benchmark Datasets

We benchmarked the library on 3 datasets : **Adult Income**, **COMPAS**, and **LSAC**

- **Direct discrimination is the primary contributor** to mean difference and equalized odds across all 3 datasets
- The classifier for Adult Income, COMPAS is not counterfactually fair but is counterfactually fair for LSAC i.e. **group fairness can differ from individual fairness**
- **Intersectional Analysis (Race x Sex) worsens direct discrimination** across all three datasets

Figure 2: Counterfactual Fairness Plots



A. Adult Income: Gender

B. Adult Income: Gender x Race

Limitations

- **Lack of identifiability can limit analysis:** Ex - in the Adult Income dataset, identifiability issues prevent the causal decomposition of equalized odds

Conclusion & Future Work

- **Actionable:** Provides specific targets for bias mitigation (e.g., fixing the 16.5% direct effect in Adult Income)
- **Future:** Extending the package to include remediation algorithms and sensitivity analysis.

References

1 Plecko, D. &; Bareinboim, E.,2024 . "Causal fairness analysis.". In: Foundations and Trends® in Machine Learning: Vol. 17, No. 3, pp 1–238

2 Zhang, J. &; Bareinboim, E, "Equality of opportunity in classification: A Causal approach. ." In: Advances in Neural Information Processing Systems.

3. Kusner, M.J. et al., 2017." Counterfactual fairness" In: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems

Table 2: CausalFairnessInAction Benchmarking

Dataset	Protected Attribute	Mean Difference	FNR	FPR	$\text{DE}_a^{\text{sym}}(y\|a)$	$\text{IE}_a^{\text{sym}}(y\|a)$	$\text{SE}_{a_0,a_1}(y\|a)$	ER^d	ER^i	ER^s	Counterfactual Fairness
Adult Income	Gender	0.203	0.410	-0.104	0.165	0.039	0.000	0.000	0.000	0.000	-0.031
Adult Income	Intersectional	0.221	0.445	-0.115	0.152	0.069	0.000	0.000	0.000	0.000	-0.068
COMPAS	Race (Black)	0.326	-0.310 (-42)	-0.253 (-0.41)	0.154	0.071	0.101	FPR: -0.297, FNR: -0.265	0	FPR: 0.113, FNR: 0.162	0.055
COMPAS	Intersectional	0.620	-0.620	-0.518	0.513	0.081	0.027	-	-	-	0.640
LSAC	Race (Black)	0.978	-	-	0.554	0.429	0.000	-	-	-	0.001
LSAC	Intersectional	0.990	-	-	0.531	0.458	0.000	-	-	-	-0.007