Fairness Aware Counterfactuals for Subgroups Supplemental Material

Anonymous Author(s)
Affiliation
Address
email

This is the appendix to the main paper, describing in detail the experimental setting (Section A), presenting the datasets (Section B), providing additional results and discussion (Section C), and quantitatively comparing the various fairness notions (Section D).

A Experimental Setting

Models To conduct our experiments, we have used the Logistic Regression[1] classification model, where we use the default implementation of the python package `scikit-learn`[2]. This model corresponds to the black box one that our framework audits in terms of fairness of recourse.

Train-Test Split For our experiments, all datasets are split into training and test sets with proportions 70% and 30%, respectively. Both shuffling of the data and stratification based on the labels were employed. Our results can be reproduced using the random seed value 131313 in the data split function (`train_test_split`[3] from the python package `scikit-learn`). FACTS is deployed solely on the test set.

Frequent Itemset Mining The set of subgroups and the set of actions are generated by executing the fp-growth[4] algorithm for frequent itemset mining. We used the implementation in the Python package `mlxtend` [5]. We deploy fp-growth with support threshold **1%**, i.e., we require the return of subgroups and actions with at least 1% frequency in the respective populations. Recall that subgroups are derived from the affected populations D_0 and D_1 and actions are derived from the unaffected population.

Effectiveness and Budgets As we have stated in Section 2 our main paper , the metrics *Equal Choice for Recourse* and *Equal Cost of Effectiveness* require the definition of a target effectiveness level ϕ, while the metric *Equal Effectiveness within Budget* requires the definition of a target cost level (or budget) c.

Regarding the metrics that require the definition of an effectiveness level ϕ, we used two different values arbitrarily, i.e., a relatively low effectiveness level of $\phi = 30\%$ and a relatively high effectiveness level of $\phi = 70\%$.

For the estimation of budget-level values c we followed a more elaborate procedure. Specifically,

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[2] `https://scikit-learn.org/stable/index.html`

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

[4] `https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/`

[5] `https://github.com/rasbt/mlxtend`

1. Compute the *Equal Cost of Effectiveness* (micro definition) with a target effectiveness level of $\phi = 50\%$ to calculate, for all subgroups G, the minimum cost required to flip the prediction for at least 50% of both G_0 and G_1.
2. Gather all such minimum costs of step 1 in an array.
3. Choose budget values as percentiles of this set of cost values. We have chosen the **30%, 60%** and **90%** percentiles arbitrarily.

Cost Functions Our implementation allows the user to define any cost function based on their domain knowledge and requirements. For evaluation and demonstration purposes, we implement an indicative set of cost functions, according to which, the cost of a change of a feature value v to the value v' is defined as follows:

1. **Numerical features:** $|norm(v) - norm(v')|$, where $norm$ is a function that normalizes values to $[0, 1]$.
2. **Categorical features:** 1 if $v \neq v'$, and 0 otherwise.
3. **Ordinal features:** $|pos(v) - pos(v')|$, where pos is a function that provides the order for each value.

Additionally to the above costs, the user is able to define a feature-specific weight that indicates the difficulty to change the given feature through an action. Thus, for each dataset, the cost of actions can be simply determined by specifying the numerical, categorical, and ordinal features, as well as the weights for each feature.

Feasibility Apart from the cost of actions, we also take care of some obvious unfeasible actions such as that the age and education features can not be reduced and actions should not lead to unknown or missing values.

Compute resources Experiments were run on commodity hardware (AMD Ryzen 5 5600H processor, 8GB RAM). On the software side, all experiments were run in an isolated conda environment using Python 3.9.16.

B Datasets Description

We have used four datasets in our experimental evaluation; the main paper presented results only on the first. For each dataset, we provide details about the preprocessing procedure, specify feature types, and list the cost feature weights applied.

B.1 Adult

We have generated CSCs in the Adult dataset[6] using two different features as protected attributes, i.e., ‘sex’, and ‘race’. The assessment of bias for each protected attribute is done separately. The results for ‘sex’ as the protected attribute are presented in the main paper. Before we present our results for race as the protected attribute, we briefly discuss the preprocessing procedures and feature weights used for the adult dataset.

Preprocessing We removed the features ‘fnlwgt’ and ‘education’ and any rows with unknown values. The ‘hours-per-week’ and ‘age’ features have been discretized into 5 bins each.

Features All features have been treated as categorical, except for ‘capital-gain’ and ‘capital-loss’, which are numeric, and ‘education-num’ and ‘hours-per-week’, which we treat as ordinal. The feature weights that we used for the cost function are presented in Table 2. We need to remind here that this comprises only an indicative weight assignment to serve our experimentation; the weight below try to capture the notion of how feasible/actionable it is to perform a change to a specific feature.

[6] https://raw.githubusercontent.com/columbia/fairtest/master/data/adult/adult.csv

Table 2: Cost Feature Weights for Adult

feature name	weight value	feature name	weight value
native-country	4	Workclass	2
marital-status	5	hours-per-week	2
relationship	5	capital-gain	1
age	10	capital-loss	1
occupation	4	education-num	3

B.2 COMPAS

We have generated CSCs in the COMPAS dataset[7] for race as the protected attribute. Apart from our results, we provide some brief information regarding preprocessing procedures and the cost feature weights for the COMPAS dataset.

Preprocessing We discard the features 'age' and 'c_charge_desc'. The 'priors_count' feature has been discretized into 5 bins: [-0.1,1), [1, 5) [5, 10) [10, 15) and [15, 38), while trying to keep the frequencies of each bin approximately equal (the distribution of values is highly asymmetric so this is not possible with the direct use of e.g., `pandas.qcut`[8]).

Features We treat the features 'juv_fel_count', 'juv_misd_count', 'juv_other_count' as numerical and the rest as categorical. The feature weights used for the cost function are shown in Table 3.

Table 3: Cost Feature Weights for COMPAS

feature name	weight value
age_cat	10
juv_fel_count	1
juv_fel_count	1
juv_other_count	1
priors_count	1
c_charge_degree	1

B.3 SSL

We have generated CSCs in the SSL dataset[9] for race as the protected attribute. Before we move to our results, we discuss briefly preprocessing procedures and feature weights applied in the SSL dataset.

Preprocessing We remove all rows with missing values ('U' or 'X') from the dataset. We also discretize the feature 'PREDICTOR RAT TREND IN CRIMINAL ACTIVITY' into 6 bins. Finally, since the target labels are values between 0 and 500, we 'binarize' them by assuming values above 344 to be positively impacted and below 345 negatively impacted (following the principles used in [10]).

Features In this dataset, we treat all features as numerical (apart from the protected race feature). The feature weights used for the cost function are presented in Table 4.

B.4 Ad Campaign

We have generated CSCs in the Ad Campaign dataset[11] for gender as the protected attribute.

[7] https://aif360.readthedocs.io/en/latest/modules/generated/aif360.sklearn.datasets.fetch_compas.html
[8] `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.qcut.html`
[9] https://raw.githubusercontent.com/samuel-yeom/fliptest/master/exact-ot/chicago-ssl-clean.csv
[10] https://arxiv.org/abs/1906.09218
[11] https://developer.ibm.com/exchanges/data/all/bias-in-advertising/

Table 4: Cost Feature Weights for SSL

feature name	**weight value**
PREDICTOR RAT AGE AT LATEST ARREST	10
PREDICTOR RAT VICTIM SHOOTING INCIDENTS	1
PREDICTOR RAT VICTIM BATTERY OR ASSAULT	1
PREDICTOR RAT ARRESTS VIOLENT OFFENSES	1
PREDICTOR RAT GANG AFFILIATION	1
PREDICTOR RAT NARCOTIC ARRESTS	1
PREDICTOR RAT TREND IN CRIMINAL ACTIVITY	1
PREDICTOR RAT UUW ARRESTS	1

Preprocessing We decided not to remove missing values, since they represent the vast majority of values for all features. However, we did not allow actions that lead to missing values in the CSCs representation.

Features In this dataset, we treat all features, apart from the protected one, as categorical. The feature weights used for the cost function are shown in Table 5.

Table 5: Cost Feature Weights for Ad Campaign

feature name	**weight value**
religion	5
politics	2
parents	3
age	10
income	3
area	2
college_educated	3
homeowner	1

C Additional Results

This section repeats the experiment described in the main paper, concerning the Adult dataset with 'gender' as the protected attribute (Section 4), to three other cases. Specifically, we provide three subgroups that were ranked first in terms of unfairness according to a metric, highlight why they were marked as unfair by our framework, and summarize their unfairness scores according to rest of the metrics.

C.1 Results for Adult with race as the protected attribute

We showcase three prevalent subgroups for which the rankings assigned by different fairness definitions truly yield different kinds of information. This is showcased in Table 6. We once again note that the results presented here are for 'race' as a protected attribute, while the corresponding results for 'gender' are presented in Section 4 of the main paper.

Table 6: Example of three unfair subgroups in Adult (protected attribute race)

	Subgroup 1			**Subgroup 2**			**Subgroup 3**		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	Fair	Fair	0.0	3047.0	Non-White	0.115	1682.0	Non-White	0.162
Equal Choice for Recourse ($\phi = 0.3$)	1	Non-White	10.0	10.0	Non-White	1.0	Fair	Fair	0.0
Equal Choice for Recourse ($\phi = 0.7$)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1.15)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 10.0)	303.0	Non-White	0.242	2201.0	Non-White	0.115	4035.0	Non-White	0.071
Equal Effectiveness within Budget (c = 21.0)	Fair	Fair	0.0	2978.0	Non-White	0.115	1663.0	Non-White	0.162
Equal Cost of Effectiveness ($\phi = 0.3$)	18.0	Non-White	0.15	1	Non-White	inf	Fair	Fair	0.0
Equal Cost of Effectiveness ($\phi = 0.7$)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	909.0	Non-White	0.242	4597.0	Non-White	0.115	2644.0	Non-White	0.162
Equal (Conditional) Mean Recourse	5897.0	White	0.021	5309.0	White	0.047	1	Non-White	inf

In Figure 4 we present the Comparative Subgroup Counterfactual representation for the subgroups of Table 6 that corresponds to the fairness metric for which each subgroup presents the minimum rank.

These results are in line with the findings reported in the main paper (Section 4), on the same dataset (Adult), but on a different protected attribute (race instead of gender). Subgroups that are ranked first (highly unfair) with respect to a specific definition, are ranked much lower or even considered as fair according to most of the remaining definitions. This serves as an indication for the utility of the different fairness definitions, which is further strengthen by the diversity of the respective CSCs of Table 6. For example, the Subgroup 1 CSC (ranked first *Equal Choice for Recourse ($\phi = 0.3$)*), demonstrates unfairness by contradicting a plethora of actions for the "White" protected subgroup, as opposed to much less actions for the the "Non-White" protected subgroup. For Subgroup 2, a much more concise representation is provided, tied to the respective definition (*Equal Cost of Effectiveness ($\phi = 0.3$)*): no recourses are identified for the desired percentage of the "Non-White" unfavored population, as opposed to the "White" unfavored population.

C.2 Results for COMPAS

We present some ranking statistics for three interesting subgroups for all fairness definitions (Table 7). The Comparative Subgroup Counterfactuals for the same three subgroups are shown in Figure 5.

Table 7: Example of three unfair subgroups in COMPAS

	Subgroup 1			**Subgroup 2**			**Subgroup 3**		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	Fair	Fair	0.0	116.0	African-American	0.151	209.0	African-American	0.071
Equal Choice for Recourse ($\phi = 0.3$)	Fair	Fair	0.0	3.0	African-American	1.0	Fair	Fair	0.0
Equal Choice for Recourse ($\phi = 0.7$)	**1**	African-American	3.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	66.0	African-American	0.167	79.0	African-American	0.151	185.0	African-American	0.071
Equal Effectiveness within Budget (c = 10)	84.0	African-American	0.167	108.0	African-American	0.151	220.0	African-American	0.071
Equal Cost of Effectiveness ($\phi = 0.3$)	Fair	Fair	0.0	**1**	African-American	inf	Fair	Fair	0.0
Equal Cost of Effectiveness ($\phi = 0.7$)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	3.0	African-American	0.5	214.0	African-American	0.151	376.0	African-American	0.071
Equal (Conditional) Mean Recourse	59.0	African-American	1.667	Fair	Fair	0.0	**1**	African-American	inf

C.3 Results for SSL

In Table 8 we present a summary of the ranking statistics for three interesting subgroups. and their respective Comparative Subgroup Counterfactuals in Figure 6.

Table 8: Example of three unfair subgroups in SSL

	Subgroup 1			**Subgroup 2**			**Subgroup 3**		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	1630.0	Black	0.076	70.0	Black	0.663	979.0	Black	0.151
Equal Choice for Recourse ($\phi = 0.3$)	Fair	Fair	0.0	12.0	Black	1.0	12.0	Black	1.0
Equal Choice for Recourse ($\phi = 0.7$)	13.0	Black	3.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	Fair	Fair	0.0	195.0	Black	0.663	1692.0	White	0.138
Equal Effectiveness within Budget (c = 2)	2427.0	Black	0.111	126.0	Black	0.663	3686.0	White	0.043
Equal Effectiveness within Budget (c = 10)	2557.0	Black	0.076	73.0	Black	0.663	1496.0	Black	0.151
Equal Cost of Effectiveness ($\phi = 0.3$)	Fair	Fair	0.0	**1**	Black	inf	**1**	Black	inf
Equal Cost of Effectiveness ($\phi = 0.7$)	**1**	Black	inf	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	3393.0	Black	0.111	443.0	Black	0.663	2685.0	Black	0.151
Equal (Conditional) Mean Recourse	3486.0	Black	0.053	**1**	Black	inf	1374.0	White	0.95

C.4 Results for Ad Campaign

In Table 9 we present, as we did for the other datasets, the ranking results for 3 interesting subgroups, while in Figure 7, we show the respective Comparative Subgroup Counterfactuals for these subgroups.

Subgroup 1
```
If workclass=Private, age=(34.0, 41.0], capital-gain=0, capital-loss=0, marital-status=Never-married, native-
    country=United-States, relationship=Not-in-family:
  Protected Subgroup = 'Non-White', 1.09% covered
    Make Workclas=Federal-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 36.84%.
    Make capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with effectiveness
      100.00%.
    Make age=(41.0, 50.0], capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 100.00%.
  Protected Subgroup = 'White', 1.94% covered
    Make age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married with effectiveness
      45.14%.
    Make marital-status=Married-civ-spouse, relationship=Married with effectiveness 40.00%.
    Make age=(50.0, 90.0], marital-status= Married-civ-spouse, relationship=Married with effectiveness
      42.86%.
    Make Workclass=Local-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 44.00%.
    Make Workclass=Local-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 44.00%.
    Make Workclas=Self-emp-inc, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 52.57%.
    Make Workclass=Self-emp-inc, age=(50.0, 90.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 49.71%.
    Make Workclass=Local-gov, marital-status=Married-civ-spouse, relationship=Married with effectiveness
      36.00%.
    Make Workclass=Self-emp-inc, marital-status=Married-civ-spouse, relationship=Married with
      effectiveness 45.14%.
    Make Workclass= Federal-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 62.29%.
    Make Workclass=State-gov, age=(41.0, 50.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 40.57%.
    Make capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married with effectiveness
      99.43%.
    Make Workclass=Local-gov, age=(50.0, 90.0], marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 40.00%.
    Make age=(41.0, 50.0], capital-gain=15024, marital-status=Married-civ-spouse, relationship=Married
      with effectiveness 100.00%.
  Bias against 'Non-White' due to Equal Choice for Recourse (threshold = 0.3). Unfairness score = 10.
```

Subgroup 2
```
If hours-per-week = FullTime, native-country = United-States, occupation = Adm-clerical, relationship =
    Married:
  Protected Subgroup = 'Non-White', 1.66% covered
    No recourses for this subgroup.
  Protected Subgroup = 'White', 1.66% covered
    Make hours-per-week = BrainDrain, occupation = Exec-managerial with effectiveness 70.00%.
  Bias against 'Non-White' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Subgroup 3
```
If hours-per-week = PartTime, marital-status = Divorced, native-country = United-States:
  Protected Subgroup = 'Non-White', 1.15% covered
    Make marital-status=Married-civ-spouse with effectiveness 0.00%.
    Make hours-per-week=MidTime, marital-statu=Married-civ-spouse with effectiveness 0.00%.
    Make hours-per-week=FullTime, marital-status=Married-civ-spouse with effectiveness 0.00%.
    Make hours-per-week=OverTime, marital-status=Married-civ-spouse with effectiveness 0.00%.
    Make hours-per-week=OverTime, marital-status=Never-married with effectiveness 0.00%.
    Make hours-per-week=BrainDrain, marital-status=Married-civ-spouse with effectiveness 0.00%.
  Protected Subgroup = 'White', 1.66% covered
    Make marital-status=Married-civ-spouse with effectiveness 1.01%.
    Make hours-per-week=MidTime, marital-statu=Married-civ-spouse with effectiveness 1.01%.
    Make hours-per-week=FullTime, marital-status=Married-civ-spouse with effectiveness 7.07%.
    Make hours-per-week=OverTime, marital-status=Married-civ-spouse with effectiveness 7.07%.
    Make hours-per-week=OverTime, marital-status=Never-married with effectiveness 15.15%.
    Make hours-per-week=BrainDrain, marital-status=Married-civ-spouse with effectiveness 16.16%.
  Bias against 'Non-White' due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure 4: Example of three Comparative Subgroup Counterfactuals in Adult (protected attribute race); ref. Table 6

Table 9: Example of three unfair subgroups in Ad Campaign

	Subgroup 1			Subgroup 2			Subgroup 3		
	rank	bias against	unfairness score	rank	bias against	unfairness score	rank	bias against	unfairness score
Equal Effectiveness	319.0	Female	0.286	Fair	Fair	0.0	467.0	Male	0.099
Equal Choice for Recourse ($\phi = 0.3$)	5.0	Female	1.0	2.0	Female	4.0	Fair	Fair	0.0
Equal Choice for Recourse ($\phi = 0.7$)	Fair	Fair	0.0	**1**	Female	4.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 1)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Equal Effectiveness within Budget (c = 5)	Fair	Fair	0.0	Fair	Fair	0.0	345.0	Male	0.099
Equal Cost of Effectiveness ($\phi = 0.3$)	**1**	Female	inf	Fair	Fair	0.0	Fair	Fair	0.0
Equal Cost of Effectiveness ($\phi = 0.7$)	Fair	Fair	0.0	Fair	Fair	0.0	Fair	Fair	0.0
Fair Effectiveness-Cost Trade-Off	331.0	Female	0.286	Fair	Male	0.0	547.0	Male	0.099
Equal (Conditional) Mean Recourse	Fair	Fair	0.0	Fair	Fair	0.0	**1**	Male	inf

Subgroup 1

```
If age_cat = 25 - 45, c_charge_degree = M, juv_misd_count = 0, priors_count = (10.0, 15.0]:
    Protected Subgroup = 'Caucasian', 1.03% covered
        Make c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make priors_count=(1.0, 5.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(1.0, 5.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(1.0, 5.0] with effectiveness 100.00%.
    Protected Subgroup = 'African-American', 1.16% covered
        Make priors_count=(-0.1, 1.0] with effectiveness 83.33%.
        Make age_cat=Greater than 45, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, c_charge_degree=F, priors_count=(-0.1, 1.0] with effectiveness 100.00%.
        Make age_cat=Greater than 45, priors_count=(1.0, 5.0] with effectiveness 83.33%.
    Bias against African-American due to Equal Choice for Recourse (threshold = 0.7). Unfairness score = 3.
```

Subgroup 2

```
If c_charge_degree = M, juv_other_count = 1:
    Protected Subgroup = 'Caucasian', 3.59% covered
        Make juv_other_count = 0 with effectiveness 42.86%.
    Protected Subgroup = 'African-American', 3.48% covered
        No recourses for this subgroup.
    Bias against African-American due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf
        .
```

Subgroup 3

```
If age_cat = Greater than 45, c_charge_degree = F, juv_fel_count = 0, juv_misd_count = 0, juv_other_count = 0,
      sex = Male:
    Protected Subgroup = 'Caucasian', 7.18% covered
        Make c_charge_degree=M with effectiveness 7.14%.
    Protected Subgroup = 'African-American', 6.00% covered
        Make c_charge_degree=M with effectiveness 0.00%.
    Bias against African-American due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure 5: Example of three Comparative Subgroup Counterfactuals in COMPAS; ref. Table 7

```
Subgroup 1
If PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 1, PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT VICTIM
    BATTERY OR ASSAULT = 1:
  Protected Subgroup = 'Black', 1.04% covered
          No recourses for this subgroup.
  Protected Subgroup = 'White', 1.00% covered
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT VICTIM
        BATTERY OR ASSAULT = 0 with effectiveness 72.73%
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT VICTIM
        BATTERY OR ASSAULT = 0 with effectiveness 72.73%
    Make PREDICTOR RAT ARRESTS VIOLENT OFFENSES = 0, PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT VICTIM
        BATTERY OR ASSAULT = 0 with effectiveness 72.73%
  Bias against 'Black' due to Equal Cost of Effectiveness (threshold = 0.7). Unfairness score = inf.
```

```
Subgroup 2
If PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1], PREDICTOR RAT
    UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0:
  Protected Subgroup = 'Black', 2.51% covered
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.3, -0.2],
        PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1],
        PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
  Protected Subgroup = 'White', 2.87% covered
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        57.14%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.3, -0.2],
        PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.1, 0.1], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-0.2, -0.1],
        PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.1, 0.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (-8.200999999999999,
        -0.3], PREDICTOR RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness
        0.00%
    Make PREDICTOR RAT NARCOTIC ARRESTS = 1, PREDICTOR RAT TREND IN CRIMINAL ACTIVITY = (0.3, 7.3], PREDICTOR
        RAT UUW ARRESTS = 0, PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with effectiveness 0.00%
  Bias against 'Black' due to Equal(Conditional) Mean Recourse. Unfairness score = inf.
```

```
Subgroup 3
If PREDICTOR RAT GANG AFFILIATION = 1, PREDICTOR RAT NARCOTIC ARRESTS = 2, PREDICTOR RAT TREND IN CRIMINAL
    ACTIVITY = (-8.200999999999999, -0.3], PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0:
  Protected Subgroup = 'Black', 1.18% covered
          No recourses for this subgroup.
  Protected Subgroup = 'White', 1.00% covered
    Make PREDICTOR RAT GANG AFFILIATION = 0, PREDICTOR RAT NARCOTIC ARRESTS = 0, PREDICTOR RAT TREND IN
        CRIMINAL ACTIVITY = (-8.200999999999999, -0.3], PREDICTOR RAT VICTIM SHOOTING INCIDENTS = 0 with
        effectiveness 31.82%
  Bias against 'Black' due to Equal Cost of Effectiveness (threshold = 0.3). Unfairness score = inf.
```

Figure 6: Example of three Comparative Subgroup Counterfactuals in SSL; ref. Table 8

```
Subgroup 1
If age = 45-54, area = Unknown, parents = 1:
    Protected Subgroup = Male', 1.22% covered
        Make age=55-64, area=Rural with effectiveness 30.77%.
    Protected Subgroup = 'Female', 1.13% covered
        No recourses for this subgroup.
    Bias against Female due to Equal Cost of Effectiveness (threshold=0.3). Unfairness score = inf.
```

```
Subgroup 2
If age = 55-64, area = Unknown, homeowner = 1, income = Unknown, parents = 0, politics = Unknown, religion =
    Unknown:
    Protected Subgroup = 'Male', 2.53% covered
        Make homeowner=0, parents=1 with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Other with effectiveness 100.00%.
        Make area=Urban, parents=1, religion=Christianity with effectiveness 93.91%.
        Make area=Urban, parents=1, religion=Other with effectiveness 93.91%.
        Make homeowner=0, income=<100K, parents=1 with effectiveness 100.00%.
        Make area=Rural, parents=1, religion=Other with effectiveness 100.00%.
        Make area=Rural, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Other with effectiveness 100.00%.
    Protected Subgroup = 'Female', 2.33% covered
        Make homeowner=0, parents=1 with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, parents=1, religion=Other with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1 with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Christianity with effectiveness 100.00%.
        Make homeowner=0, income=<100K, parents=1, religion=Other with effectiveness 100.00%.
    Bias against Female due to Equal Choice for Recourse (threshold=0.7). Unfairness score = 4.
```

```
Subgroup 3
If ages = 55-64, income = <100K, religion = Unknown:
    Protected Subgroup = 'Male', 1.02% covered
        Make religion=Christianity with effectiveness 0.00%.
        Make religion=Other with effectiveness 0.00%.
    Protected Subgroup = 'Female', 1.08% covered
        Make religion=Christianity with effectiveness 9.86%.
        Make religion=Other with effectiveness 9.86%.
    Bias against Male due to Equal Conditional Mean Recourse. Unfairness score = inf.
```

Figure 7: Example of three Comparative Subgroup Counterfactuals in Ad Campaign; ref. Table 9

D Comparison of Fairness Metrics

The goal of this section is to answer the question: "How different are the fairness of recourse metrics". To answer it, we consider all subgroups and compare how they rank in terms of unfairness according to 12 distinct metrics. The results justify our claim in the main paper that the fairness metrics capture different aspects of recourse unfairness. For each dataset and protected attribute, we provide: (a) the ranking analysis table, and (b) the aggregated rankings table.

The first column of the *ranking analysis* table shows the number of the most unfair subgroups per metric, i.e., how many ties are in rank 1. Depending on the unit of the unfairness score being compared between the protected subgroups (namely: cost, effectiveness, or number of actions), the number of ties can vary greatly. Therefore, we expect to have virtually no ties when comparing effectiveness percentages and to have many ties when comparing costs. The second and third columns show the number of subgroups where we observe bias in one direction (e.g., against males) and the opposite (e.g., against females) among the top 10% most unfair subgroups.

The *aggregated rankings* table is used as evidence that different fairness metrics capture different types of recourse unfairness. Each row concerns the subgroups that are the most unfair (i.e., tied at rank 1) according to each fairness metric. The values in the row indicate the average percentile ranks of these subgroups (i.e., what percentage of subgroups are more unfair) when ranked according to the other fairness metrics, shown as columns. Concretely, the value v of each cell i, j of this table is computed as follows:

1. We collect all subgroups of the fairness metric appearing in row i that are ranked first (the most biased) due to this metric.
2. We compute the average ranking a of these subgroups in the fairness metric appearing in column j.
3. We divide a with the largest ranking tier of the fairness metric of column j to arrive at v.

Each non-diagonal value of this table represents the relative ranking based on the specific metric of the column for all the subgroups that are ranked first in the metric of the respective row (all diagonal values of this table are left empty). A relative ranking of v in a specific metric m means that the most unfair subgroups of another metric are ranked lower on average (thus are fairer) for metric m.

D.1 Comparison on Adult for protected attribute gender

The number of affected individuals in the test set for the adult dataset is 10,205. We first split the affected individuals on the set of affected males D_1 and the set of affected females D_0. The number of subgroups formed by running fp-growth with support threshold 1% on D_1 and on D_0 and computing their intersection is 12,880. Our fairness metrics will evaluate and rank these subgroups based on the actions applied.

Tables 10 and 11 present the ranking analysis and the aggregated rankings respectively on the gender attribute, on the Adult dataset. Next, we briefly discuss the findings from these two tables; similar findings stand for the respective tables of the other datasets, thus we omit the respective discussion.

It is evident from Table 10 that the different ways to produce ranking scores by different definitions can lead to considerable differences in ties, i.e., the number of subgroups receiving the same rank (here only rank 1 is depicted). The "Top 10%" columns demonstrate interesting statistics on the protected subgroup for which bias is identified: while it is expected that mostly bias against "Female" will be identified, subgroups with reverse bias (bias against "Male") are identified, indicating robustness to gerrymandering, as hinted in Section 4 of the main paper.

Table 11 is produced to provide stronger evidence on the unique utility of the various presented definitions (see footnote 1 of the main paper: *"Additional examples, as well as statistics measuring this pattern on a significantly larger sample, are included in the supplementary material to further support this finding."*). In particular, in this table, for all subgroups that are ranked first in a definition, we calculate their average relative (normalized in $[0, 1]$) ranking in the remaining definitions. Given this, a value close to 1 means very low average rank and a value close to 0 means very high rank. Consequently, values away from 0 indicate the uniqueness and non-triviality of the different definitions and this becomes evident from the majority of the values of the table.

Table 10: Ranking Analysis in Adult (protected attribute gender)

	# Most Unfair Subgroups	# Subgroups w. Bias against Males (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Females (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	1673	56	206
(Equal Cost of Effectiveness(Macro), 0.7)	301	26	37
(Equal Choice for Recourse, 0.3)	2	54	286
(Equal Choice for Recourse, 0.7)	6	31	50
Equal Effectiveness	1	39	1040
(Equal Effectiveness within Budget, 5.0	1	41	616
(Equal Effectiveness within Budget, 10.0)	1	6	904
(Equal Effectiveness within Budget, 18.0)	1	22	964
(Equal Cost of Effectiveness(Micro), 0.3)	1523	10	226
(Equal Cost of Effectiveness(Micro), 0.7)	290	38	27
Equal(Conditional Mean Recourse)	764	540	565
(Fair Effectiveness-Cost Trade-Off, value)	1	61	1156

Table 11: Aggregated Rankings in Adult (protected attribute gender)

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 5.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Effectiveness within Budget, 18.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.836	1.0	0.214	0.509	0.342	0.285	0.3	1.0	0.441	0.237
(Equal Cost of Effectiveness(Macro), 0.7)	0.634	-	0.864	0.686	0.358	0.602	0.464	0.407	0.738	0.293	0.481	0.307
(Equal Choice for Recourse, 0.3)	0.018	1.0	-	1.0	0.001	0.006	0.001	0.001	0.017	1.0	0.105	0.001
(Equal Choice for Recourse, 0.7)	1.0	0.364	0.857	-	0.814	0.528	0.813	0.81	1.0	0.882	0.451	0.34
Equal Effectiveness	0.018	1.0	0.214	1.0	-	0.003	0.0	0.0	0.017	1.0	0.058	0.0
(Equal Effectiveness within Budget, 5.0	0.018	1.0	0.857	1.0	0.006	-	0.004	0.006	0.017	1.0	1.0	0.006
(Equal Effectiveness within Budget, 10.0)	0.018	1.0	0.214	1.0	0.0	0.002	-	0.0	0.017	1.0	0.047	0.0
(Equal Effectiveness within Budget, 18.0)	0.018	1.0	0.214	1.0	0.0	0.003	0.0	-	0.017	1.0	0.058	0.0
(Equal Cost of Effectiveness(Micro), 0.3)	0.238	1.0	0.857	1.0	0.136	0.452	0.263	0.215	-	1.0	0.462	0.155
(Equal Cost of Effectiveness(Micro), 0.7)	0.611	0.279	0.864	0.771	0.336	0.621	0.449	0.402	0.7	-	0.465	0.295
Equal(Conditional) Mean Recourse	0.996	1.0	1.0	1.0	0.723	0.946	0.875	0.777	0.997	1.0	-	0.83
(Fair Effectiveness-Cost Trade-Off, value)	0.018	1.0	0.214	1.0	0.0	0.002	0.0	0.0	0.017	1.0	0.047	-

D.2 Comparison on Adult for protected attribute race

The number of affected individuals in the test set for the adult dataset is 10,205. We first split the affected individuals on the set of affected whites D_1 and the set of affected non-whites D_0. The number of subgroups formed by running fp-growth with support threshold 1% on D_1 and on D_0 and computing their intersection is 16,621. Our fairness metrics will evaluate and rank these subgroups based on the actions applied.

Table 12: Ranking Analysis in Adult (protected attribute race)

	# Most Unfair Subgroups	# Subgroups w. Bias against Whites (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Non-Whites (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	1731	0	295
(Equal Cost of Effectiveness(Macro), 0.7)	325	7	51
(Equal Choice for Recourse, 0.3)	1	2	391
(Equal Choice for Recourse, 0.7)	2	10	60
Equal Effectiveness	1	6	1433
(Equal Effectiveness within Budget, 1.15	1	50	24
(Equal Effectiveness within Budget, 10.0	1	3	1251
(Equal Effectiveness within Budget, 21.0)	1	0	1423
(Equal Cost of Effectiveness(Micro), 0.3)	1720	0	294
(Equal Cost of Effectiveness(Micro), 0.7)	325	7	51
Equal(Conditional Mean Recourse)	2545	53	1316
(Fair Effectiveness-Cost Trade-Off, value)	2	0	0

D.3 Comparison on COMPAS

The number of affected individuals in the test set for the COMPAS dataset is 745. We first split the affected individuals on the set of affected caucasians D_1 and the set of affected african-americans D_0. The number of subgroups formed by running fp-growth with support threshold 1% on D_1 and on D_0 and computing their intersection is 995. Our fairness metrics will evaluate and rank these subgroups based on the actions applied.

D.4 Comparison on SSL

The number of affected individuals in the test set for the SSL dataset is 11,343. We first split the affected individuals on the set of affected blacks D_1 and the set of affected whites D_0 based on the race attribute (appears with the name RACE CODE CD in the dataset). The number of subgroups

Table 13: Aggregated Rankings in Adult (protected attribute race)

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.15)	(Equal Effectiveness within Budget, 10.0)	(Equal Effectiveness within Budget, 21.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.845	1.0	0.162	0.996	0.283	0.177	0.026	1.0	0.448	0.194
(Equal Cost of Effectiveness(Macro), 0.7)	0.7	-	0.9	0.829	0.147	0.973	0.315	0.169	0.698	0.05	0.421	0.12
(Equal Choice for Recourse, 0.3)	0.419	1.0	-	1.0	1.0	1.0	0.03	1.0	0.419	1.0	0.782	0.073
(Equal Choice for Recourse, 0.7)	1.0	0.095	0.909	-	0.644	1.0	0.003	0.328	1.0	0.1	0.041	0.011
Equal Effectiveness	0.023	1.0	0.909	1.0	-	1.0	0.01	0.0	0.023	1.0	0.0	0.0
(Equal Effectiveness within Budget, 1.15	1.0	0.048	1.0	0.857	0.069	-	0.047	0.07	1.0	0.05	1.0	0.102
(Equal Effectiveness within Budget, 10.0	0.395	0.048	0.818	0.571	0.001	1.0	-	0.001	0.395	0.05	0.611	0.002
(Equal Effectiveness within Budget, 21.0)	0.023	1.0	0.909	1.0	0.0	1.0	0.01	-	0.023	1.0	0.0	0.0
(Equal Cost of Effectiveness(Micro), 0.3)	0.023	1.0	0.845	1.0	0.162	0.996	0.284	0.177	-	1.0	0.449	0.195
(Equal Cost of Effectiveness(Micro), 0.7)	0.7	0.048	0.9	0.829	0.147	0.973	0.315	0.169	0.698	-	0.421	0.12
Equal(Conditional) Mean Recourse	0.979	1.0	1.0	1.0	0.628	1.0	0.778	0.633	0.979	1.0	-	0.721
(Fair Effectiveness-Cost Trade-Off, value)	0.023	1.0	0.818	1.0	0.001	1.0	0.012	0.001	0.023	1.0	0.003	-

Table 14: Ranking Analysis in COMPAS

	# Most Unfair Subgroups	# Subgroups w. Bias against Caucasians (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against African-Americans (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	51	0	11
(Equal Cost of Effectiveness(Macro), 0.7)	46	0	6
(Equal Choice for Recourse, 0.3)	13	12	8
(Equal Choice for Recourse, 0.7)	15	8	6
Equal Effectiveness	1	14	37
(Equal Effectiveness within Budget, 1.0)	4	16	30
(Equal Effectiveness within Budget, 10.0)	1	20	39
(Equal Cost of Effectiveness(Micro), 0.3)	51	0	11
(Equal Cost of Effectiveness(Micro), 0.7)	46	0	6
Equal(Conditional Mean Recourse)	37	19	24
(Fair Effectiveness-Cost Trade-Off, value)	5	18	62

formed by running fp-growth with support threshold 1% on D_1 and on D_0 and computing their intersection is 6,551. Our fairness metrics will evaluate and rank these subgroups based on the actions applied.

D.5 Comparison on Ad Campaign

The number of affected individuals in the test set for the Ad campaign dataset is 273,773. We first split the affected individuals on the set of affected males D_1 and the set of affected females D_0 based on the gender attribute. The number of subgroups formed by running fp-growth with support threshold 1% on D_1 and on D_0 and computing their intersection is 1,432. Our fairness metrics will evaluate and rank these subgroups based on the actions applied.

Table 15: Aggregated Rankings in COMPAS

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	1.0	0.65	1.0	0.169	0.801	0.398	0.2	1.0	0.797	0.226
(Equal Cost of Effectiveness(Macro), 0.7)	0.96	-	0.925	0.625	0.127	0.518	0.236	0.96	0.2	0.52	0.149
(Equal Choice for Recourse, 0.3)	0.32	0.76	-	0.775	0.082	1.0	0.178	0.32	0.76	0.297	0.116
(Equal Choice for Recourse, 0.7)	0.9	0.46	0.8	-	0.424	0.484	0.057	0.9	0.46	0.259	0.045
Equal Effectiveness	0.2	1.0	0.75	1.0	-	1.0	0.003	0.2	1.0	0.003	0.002
(Equal Effectiveness within Budget, 1.0)	0.8	1.0	0.75	0.75	1.0	-	1.0	0.8	1.0	0.413	0.002
(Equal Effectiveness within Budget, 10.0)	0.2	1.0	0.75	1.0	0.003	1.0	-	0.2	1.0	0.003	0.002
(Equal Cost of Effectiveness(Micro), 0.3)	0.2	1.0	0.65	1.0	0.169	0.801	0.398	-	1.0	0.797	0.226
(Equal Cost of Effectiveness(Micro), 0.7)	0.96	0.2	0.925	0.625	0.127	0.518	0.236	0.96	-	0.52	0.149
Equal(Conditional) Mean Recourse	0.98	1.0	1.0	1.0	0.507	0.772	0.512	0.98	1.0	-	0.627
(Fair Effectiveness-Cost Trade-Off, value)	0.68	1.0	0.75	0.8	0.801	0.202	0.801	0.68	1.0	0.331	-

Table 16: Ranking Analysis in SSL

	# Most Unfair Subgroups	# Subgroups w. Bias against Whites (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Blacks (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	371	10	107
(Equal Cost of Effectiveness(Macro), 0.7)	627	26	124
(Equal Choice for Recourse, 0.3)	1	108	184
(Equal Choice for Recourse, 0.7)	16	78	229
Equal Effectiveness	1	15	389
(Equal Effectiveness within Budget, 1.0)	18	18	436
(Equal Effectiveness within Budget, 2.0)	2	19	532
(Equal Effectiveness within Budget, 10.0)	1	15	548
(Equal Cost of Effectiveness(Micro), 0.3)	458	5	135
(Equal Cost of Effectiveness(Micro), 0.7)	671	23	130
Equal(Conditional Mean Recourse)	100	41	434
(Fair Effectiveness-Cost Trade-Off, value)	80	76	544

Table 17: Aggregated Rankings in SSL

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 2.0)	(Equal Effectiveness within Budget, 10.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	0.883	0.854	0.988	0.216	0.401	0.285	0.238	0.3	0.843	0.678	0.338
(Equal Cost of Effectiveness(Macro), 0.7)	0.929	-	0.877	0.725	0.239	0.421	0.332	0.264	0.871	0.314	0.829	0.342
(Equal Choice for Recourse, 0.3)	0.143	1.0	-	1.0	0.328	0.704	0.464	0.368	0.143	1.0	0.727	0.601
(Equal Choice for Recourse, 0.7)	1.0	0.167	0.769	-	0.083	0.177	0.127	0.086	1.0	0.143	0.926	0.135
Equal Effectiveness	0.143	0.167	0.923	0.938	-	0.002	0.0	0.0	0.143	0.143	0.0	0.003
(Equal Effectiveness within Budget, 1.0)	0.857	0.833	0.854	0.881	0.89	-	0.923	0.876	0.857	0.857	0.327	0.0
(Equal Effectiveness within Budget, 2.0)	0.286	0.333	0.923	0.938	0.5	0.002	-	0.5	0.286	0.286	0.0	0.003
(Equal Effectiveness within Budget, 10.0)	0.143	0.167	0.923	0.938	0.0	0.002	0.0	-	0.143	0.143	0.0	0.003
(Equal Cost of Effectiveness(Micro), 0.3)	0.443	0.833	0.877	0.969	0.143	0.312	0.198	0.154	-	0.843	0.729	0.268
(Equal Cost of Effectiveness(Micro), 0.7)	0.9	0.383	0.892	0.788	0.203	0.406	0.299	0.225	0.886	-	0.816	0.327
Equal(Conditional) Mean Recourse	0.6	0.733	0.946	0.969	0.244	0.464	0.395	0.378	0.514	0.729	-	0.396
(Fair Effectiveness-Cost Trade-Off, value)	0.971	0.967	0.838	0.869	0.967	0.774	0.977	0.96	0.971	0.971	0.837	-

Table 18: Ranking Analysis in Ad Campaign

	# Most Unfair Subgroups	# Subgroups w. Bias against Males (in Top 10% Unfair Subgroups)	# Subgroups w. Bias against Females (in Top 10% Unfair Subgroups)
(Equal Cost of Effectiveness(Macro), 0.3)	427	0	44
(Equal Cost of Effectiveness(Macro), 0.7)	264	0	26
(Equal Choice for Recourse, 0.3)	2	1 0	66
(Equal Choice for Recourse, 0.7)	384	0	39
Equal Effectiveness	15	0	123
(Equal Effectiveness within Budget, 1.0)	1	0	42
(Equal Effectiveness within Budget, 5.0)	10	0	114
(Equal Cost of Effectiveness(Micro), 0.3)	427	0	44
(Equal Cost of Effectiveness(Micro), 0.7)	264	0	26
Equal(Conditional Mean Recourse)	108	9	74
(Fair Effectiveness-Cost Trade-Off, value)	15	0	128

Table 19: Aggregated Rankings in Ad Campaign

	(Equal Cost of Effectiveness (Macro), 0.3)	(Equal Cost of Effectiveness (Macro), 0.7)	(Equal Choice for Recourse, 0.3)	(Equal Choice for Recourse, 0.7)	Equal Effectiveness	(Equal Effectiveness within Budget, 1.0)	(Equal Effectiveness within Budget, 5.0)	(Equal Cost of Effectiveness (Micro), 0.3)	(Equal Cost of Effectiveness (Micro), 0.7)	Equal(Conditional Mean Recourse)	(Fair Effectiveness-Cost Trade-Off, value)
(Equal Cost of Effectiveness(Macro), 0.3)	-	0.7	0.483	0.6	0.167	1.0	0.276	0.25	0.7	0.487	0.154
(Equal Cost of Effectiveness(Macro), 0.7)	0.25	-	0.35	0.333	0.082	1.0	0.21	0.25	0.5	0.506	0.079
(Equal Choice for Recourse, 0.3)	0.25	1.0	-	1.0	0.73	1.0	1.0	0.25	1.0	0.037	0.338
(Equal Choice for Recourse, 0.7)	0.5	0.65	0.333	-	0.296	0.851	0.385	0.5	0.65	0.566	0.273
Equal Effectiveness	0.25	0.5	0.333	0.333	-	1.0	0.205	0.25	0.5	0.002	0.001
(Equal Effectiveness within Budget, 1.0)	1.0	1.0	1.0	1.0	0.714	-	0.671	1.0	1.0	0.305	0.395
(Equal Effectiveness within Budget, 5.0)	0.25	0.5	0.333	0.333	0.001	1.0	-	0.25	0.5	0.002	0.001
(Equal Cost of Effectiveness(Micro), 0.3)	0.25	0.7	0.483	0.6	0.167	1.0	0.276	-	0.7	0.487	0.154
(Equal Cost of Effectiveness(Micro), 0.7)	0.25	0.5	0.35	0.333	0.082	1.0	0.21	0.25	-	0.506	0.079
Equal(Conditional) Mean Recourse	0.525	0.75	0.65	0.7	0.25	1.0	0.408	0.525	0.75	-	0.267
(Fair Effectiveness-Cost Trade-Off, value)	0.25	0.5	0.333	0.333	0.001	1.0	0.205	0.25	0.5	0.002	-