UNIVERSITY of IOANNINA

Universität der Bundeswehr München

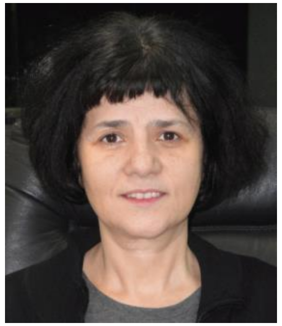
CODE FI

ARCHIMEDES

Tampere University

# Fairness and Explainability in AI

## Models, Measures, and Mitigation Strategies



Evaggelia Pitoura

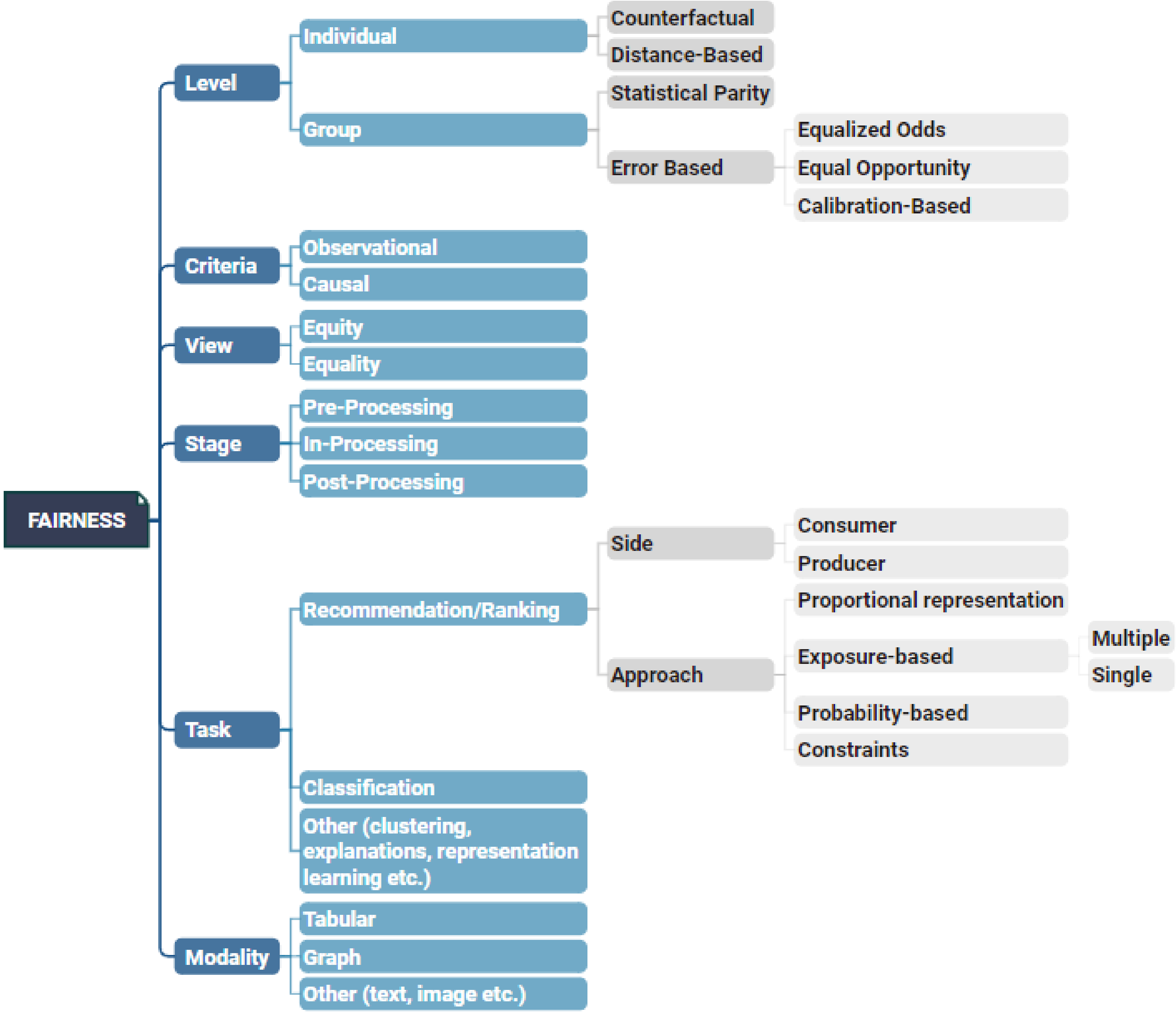
Panayiotis Tsaparas

Eirini Ntoutsi

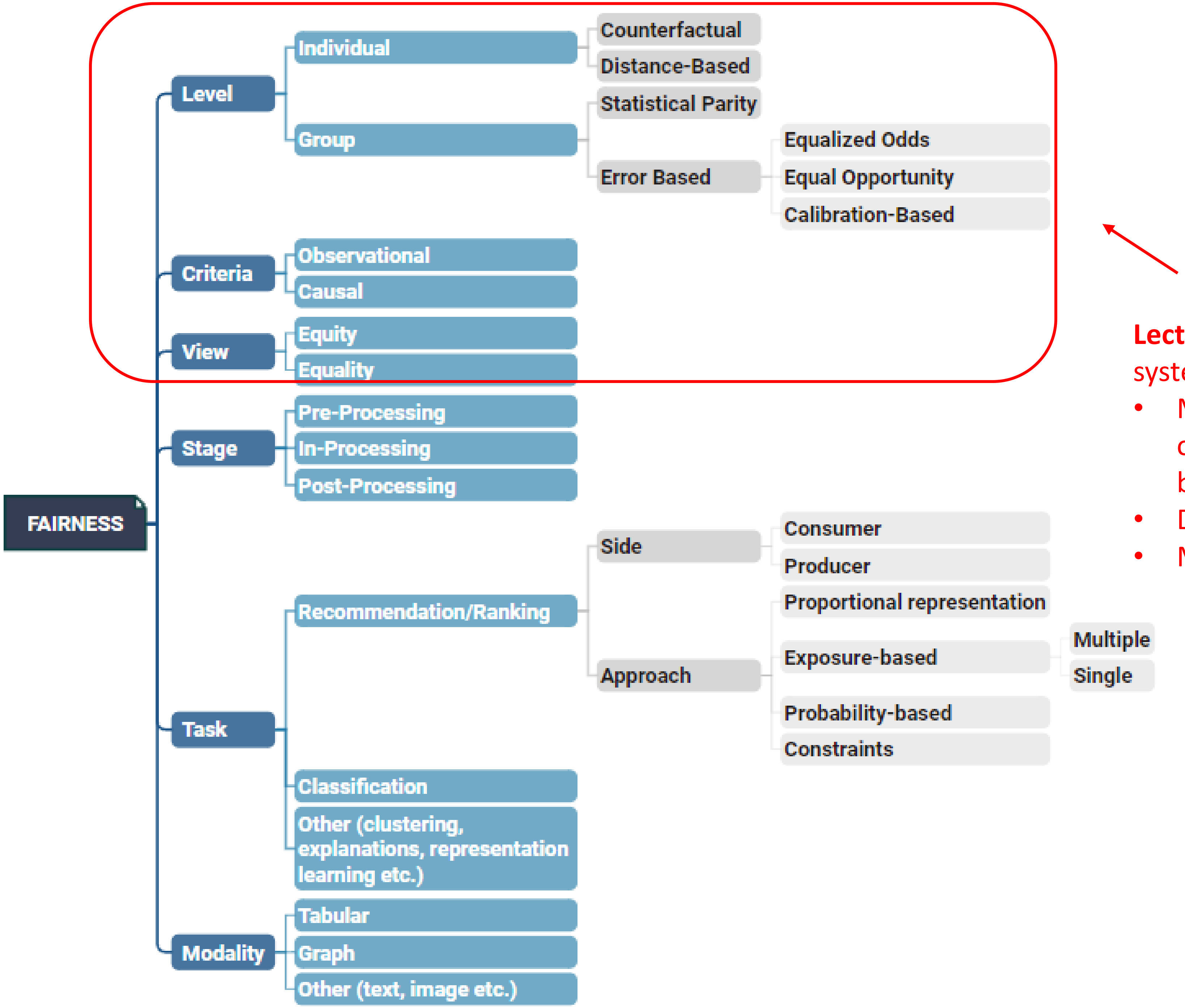
Kostas Stefanidis

ESSAI 2024, Athens (July 15 – July 19, 2024)

# Algorithmic Fairness

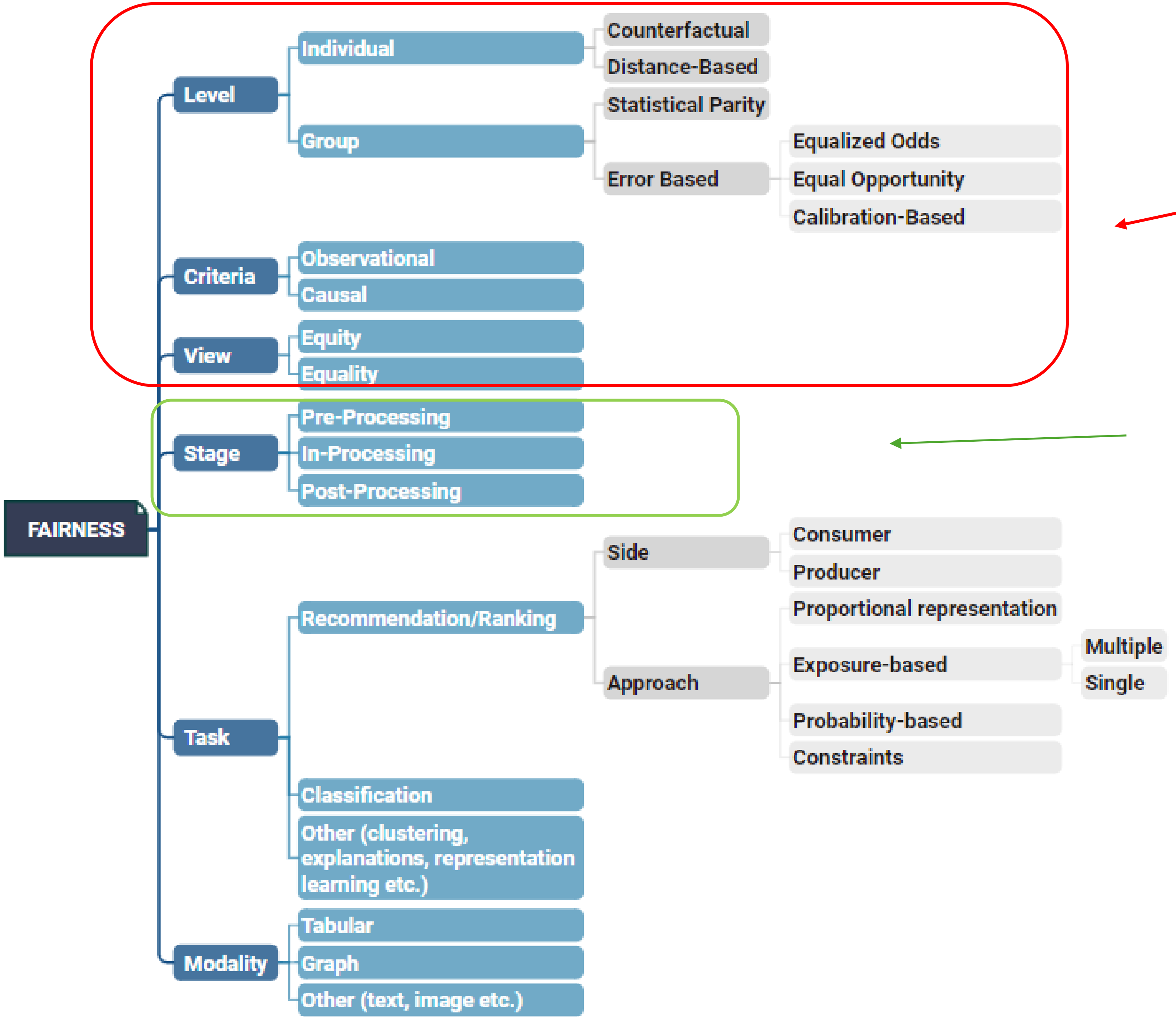


- FAIRNESS
  - Level
    - Individual
      - Counterfactual
      - Distance-Based
    - Group
      - Statistical Parity
      - Error Based
        - Equalized Odds
        - Equal Opportunity
        - Calibration-Based
  - Criteria
    - Observational
    - Causal
  - View
    - Equity
    - Equality
  - Stage
    - Pre-Processing
    - In-Processing
    - Post-Processing
  - Task
    - Recommendation/Ranking
      - Side
        - Consumer
        - Producer
      - Approach
        - Proportional representation
        - Exposure-based
          - Multiple
          - Single
        - Probability-based
        - Constraints
    - Classification
    - Other (clustering, explanations, representation learning etc.)
  - Modality
    - Tabular
    - Graph
    - Other (text, image etc.)



**Lecture 1** - Bias and discrimination in AI systems

- Motivation and application examples of algorithms exhibiting biased behaviour
- Different types of bias and their cause
- Models of fairness

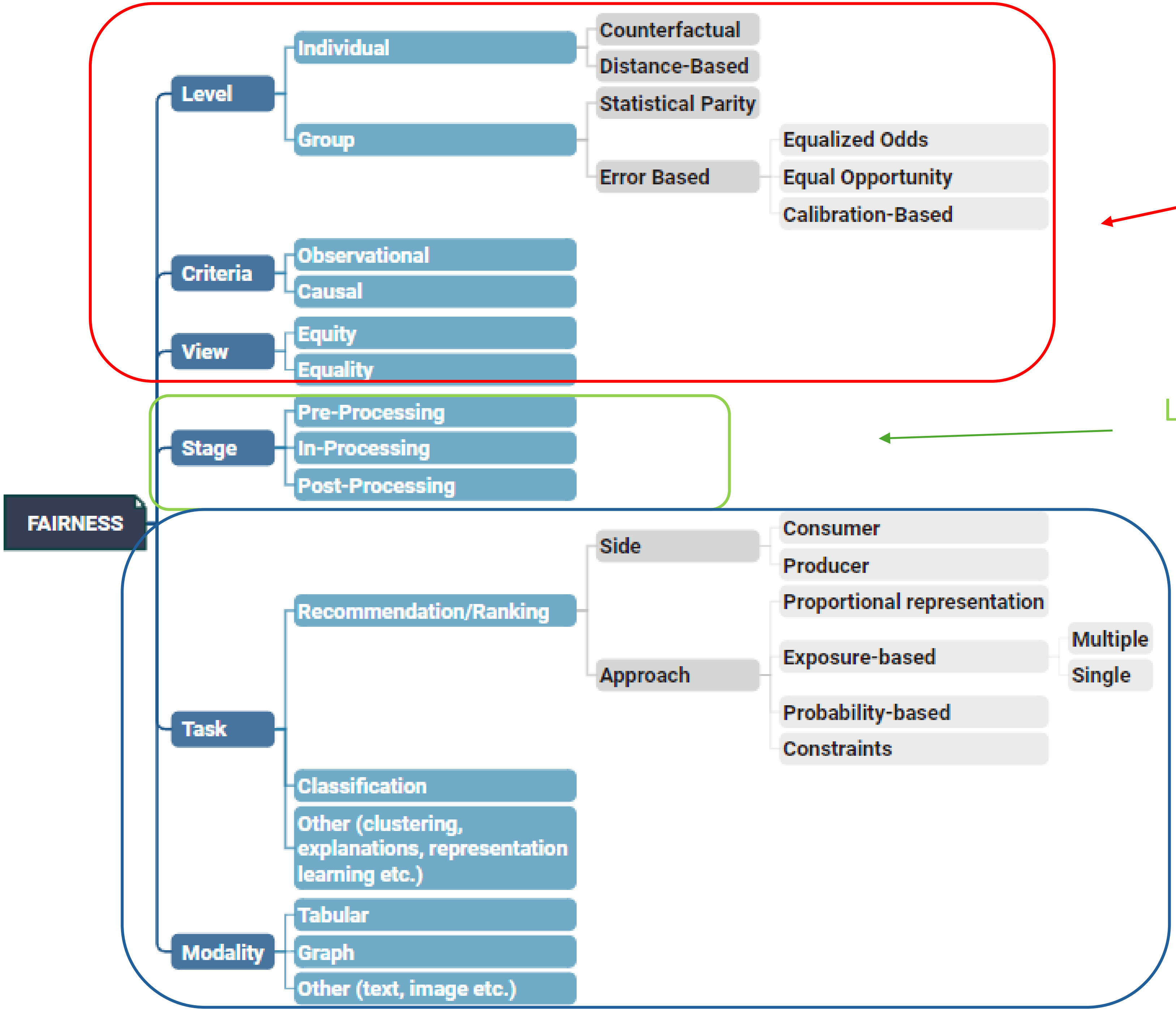


- FAIRNESS
  - Level
    - Individual
      - Counterfactual
      - Distance-Based
    - Group
      - Statistical Parity
      - Error Based
        - Equalized Odds
        - Equal Opportunity
        - Calibration-Based
  - Criteria
    - Observational
    - Causal
  - View
    - Equity
    - Equality
  - Stage
    - Pre-Processing
    - In-Processing
    - Post-Processing
  - Task
    - Recommendation/Ranking
      - Side
        - Consumer
        - Producer
      - Approach
        - Proportional representation
        - Exposure-based
          - Multiple
          - Single
        - Probability-based
        - Constraints
    - Classification
    - Other (clustering, explanations, representation learning etc.)
  - Modality
    - Tabular
    - Graph
    - Other (text, image etc.)

Lecture 1 (motivation, biases, formalization of the problem)

**Lecture 2 Bias mitigation**

- Pre-, In- and Post-processing approaches to fairness-aware learning
- End-to-end approaches to fairness-aware learning



- FAIRNESS
  - Level
    - Individual
      - Counterfactual
      - Distance-Based
    - Group
      - Statistical Parity
      - Error Based
        - Equalized Odds
        - Equal Opportunity
        - Calibration-Based
  - Criteria
    - Observational
    - Causal
  - View
    - Equity
    - Equality
  - Stage
    - Pre-Processing
    - In-Processing
    - Post-Processing
  - Task
    - Recommendation/Ranking
      - Side
        - Consumer
        - Producer
      - Approach
        - Proportional representation
        - Exposure-based
          - Multiple
          - Single
        - Probability-based
        - Constraints
    - Classification
    - Other (clustering, explanations, representation learning etc.)
  - Modality
    - Tabular
    - Graph
    - Other (text, image etc.)

Lecture 1 (motivation, biases, formalization of the problem)

Lecture 2 Bias mitigation

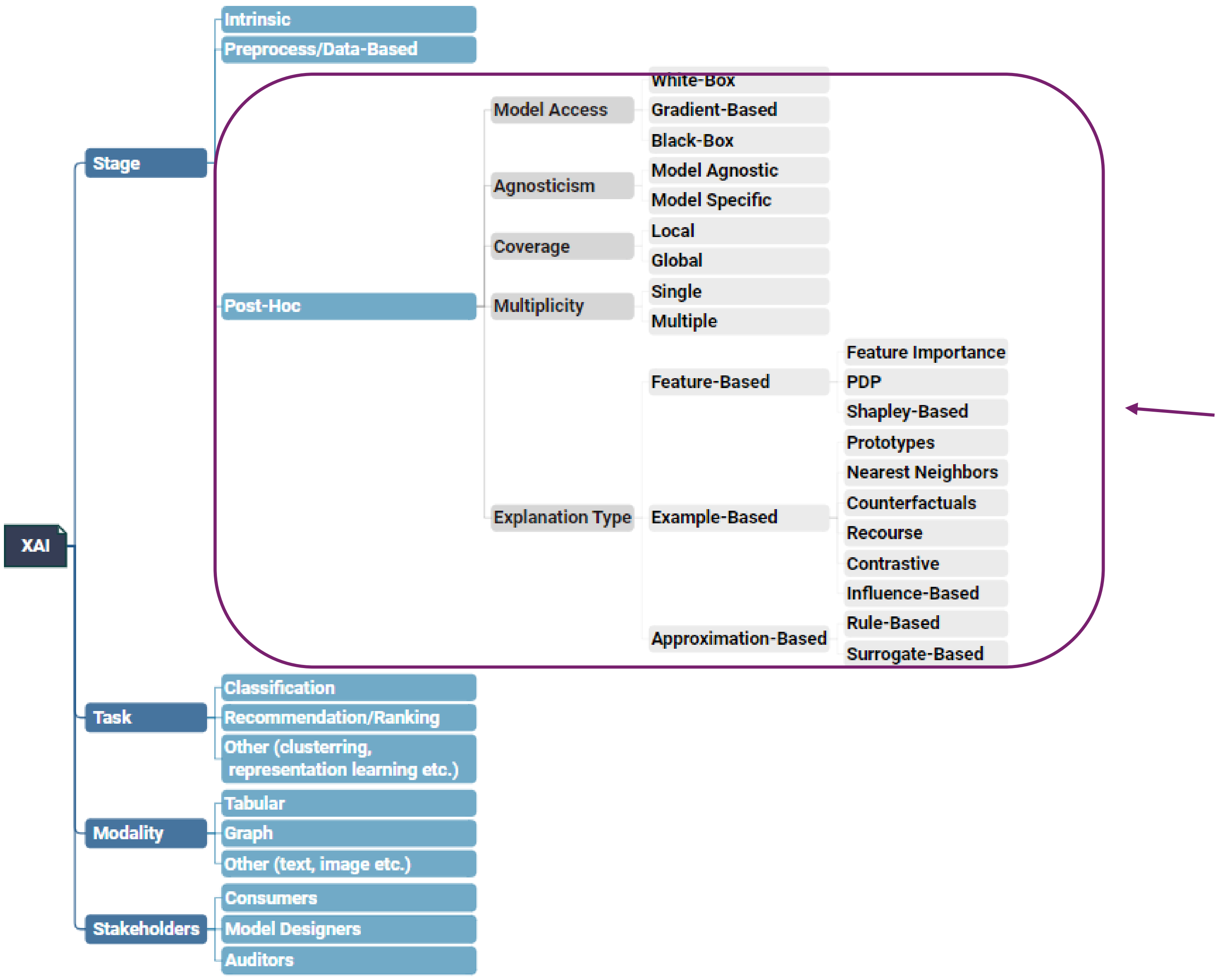
**Lecture 3**

Solutions for mitigating unfairness in concrete contexts

- Fairness in rankings and recommendations, entity resolution, graphs

E. Pitoura, K.Stefanidis, G. Koutrika: Fairness in rankings and recommendations: an overview. VLDB J. 31(3): 431-458 (2022)

# Explanations



**Lecture 4** - Explainable AI: Models and methods

- Introduction to explainable AI (XAI)
- Overview of post-hoc explanations
- LIME, Shapley values, counterfactual explanations

# Course overview

**Lecture 1** - Bias and discrimination in AI systems: Sources of bias, definitions and models of fairness

- Motivation and application examples of algorithms exhibiting biased behaviour
- Different types of bias and their cause
- Definitions of fairness

**Lecture 2.** Bias mitigation

- Pre-, In- and Post-processing approaches to fairness-aware learning
- End-to-end approaches to fairness-aware learning

**Lecture 3.** Solutions for mitigating unfairness in concrete contexts

- Fairness in rankings and recommendations, entity resolution, graphs

**Lecture 4** - Explainable AI: Models and methods

- Introduction to explainable AI (XAI)
- Overview of post-hoc explanations
- LIME, Shapley values, counterfactual explanations

**Lecture 5** - Connections between fairness and explanations

- Using explanations for fairness
    - Counterfactual explanation of unfairness
    - Actionable recourse
    - Shapley-based
- Fairness of explanations

# Fairness and Explainability in AI

## Models, Measures, and Mitigation Strategies

## Lecture 5:
## Connections between Fairness and Explainability

*Images created by deepai.org logo generator*

# Outline

- Can we explore explanations **for** fairness?
    - Counterfactuals
        - Actionable Recourse
    - Shapley Values

- Are explanation methods fair?

# Explanations for fairness

# Explanations for Fairness

**Understand** causes:

Identify causes contributing to biases

Enhance **fairness metrics**:

Propose new metrics to quantify (un)fairness

**Mitigate** unfairness:

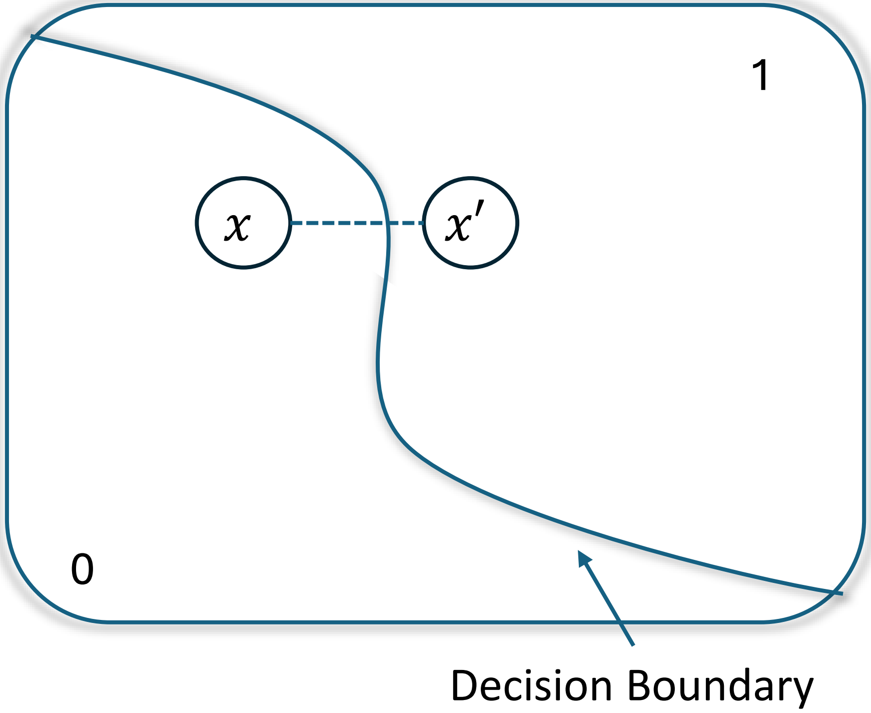
Recommend specific actions to counteract unfairness

# Counterfactual explanations (CFE) (recap)

## Example-based local explanations

Assume a binary classifier $f$

- Let an input instance (factual) $x$, for which we do not get the desired, or expected output
- Why? How should we (minimally) change $x$ to get an instance $x$' that receives the desired, or expected output?



$x$' is the counterfactual of $x$

$$\arg min_{x'} distance(x, x')\ s.t.\ f(x') \neq f(x)$$

Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, Chirag Shah, Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review, https://arxiv.org/abs/2010.10596, Nov 2022

# Counterfactuals: example (recap)

Alice is applying for a loan

- Features: $(Income, CreditScore, Education, Age)$

  $x_{ALICE}$ = (35K,1K,BSc,22)
- She is denied the loan: $f(x) = 0$

(1) Why was the loan denied? and

(2) What can she do differently so that the loan will be approved in the future?

Small changes to the feature vector, such as:

$x'_{ALICE}$ = (45K,1K,BSc,22) increase income by 10K

$x'_{ALICE}$ = (35K,1K,MSc,22) get an MSc

$x'_{ALICE}$ = (45K,1K,MSc,22) or both

CFE explicitly helps with (2) as well – actions to reverse a classifier decision

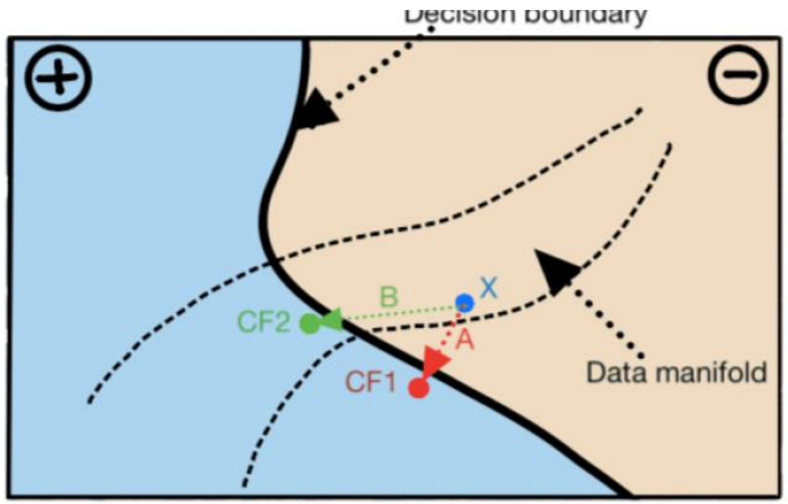
# Counterfactuals: additional constraints (recap)

Validity: A counterfactual is valid if it is classified in the desired class

$$\arg min_{x'}\, distance(x,x')\ s.t.\ f(x') = y'$$

Posed as an optimization problem

$$\arg min_{x'}\, max_{\lambda} \lambda\,(f(x') - y')^2 + distance(x,x')$$



Actionability: Distinguish between features that are mutable (e.g., income, education) and which are not (e.g., height, race, country of origin).

$$\arg min_{x' \in A}\ max_{\lambda} \lambda\,(f(x') - y')^2 + distance(x,x')$$

$A$ restricts to set of mutable features

# Counterfactuals: additional constraints (recap)

Sparsity:

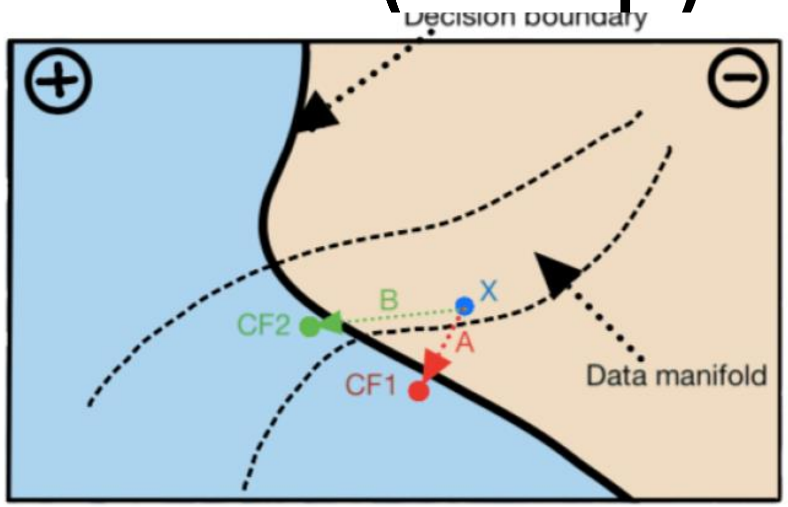
- Trade-off between the number of features changed and the total amount of change
- Ideally change a small number of features
- People find it easier to understand shorter explanations

$$\arg min_{x' \in A} max_{\lambda} \lambda \left(f(x') - y'\right)^2 + distance(x, x') + g(x, x')$$

A penalty term to encourage sparsity, e.g., L0/L1 norm

$x_{ALICE}$ = (35K,1K,BSc,22)
$x'_{ALICE}$ = (45K,1K,MSc,22)
$x'_{ALICE}$ = (55K,1K,BSc,22)

# Counterfactuals: additional constraints (recap)



## Closeness to the Data Manifold

The counterfactual should be *realistic* in the sense that it is near the training data and adheres to observed correlations among the features.

$$\arg min_{x'\in A} max_{\lambda} \lambda \left(f(x') - y'\right)^2 + distance(x, x') + g(x, x') + l(x'; X)$$

## Causality

The counterfactual should maintain any known causal relations between features

For example, getting a new educational degree means increasing age by at least some amount.

# Counterfactuals

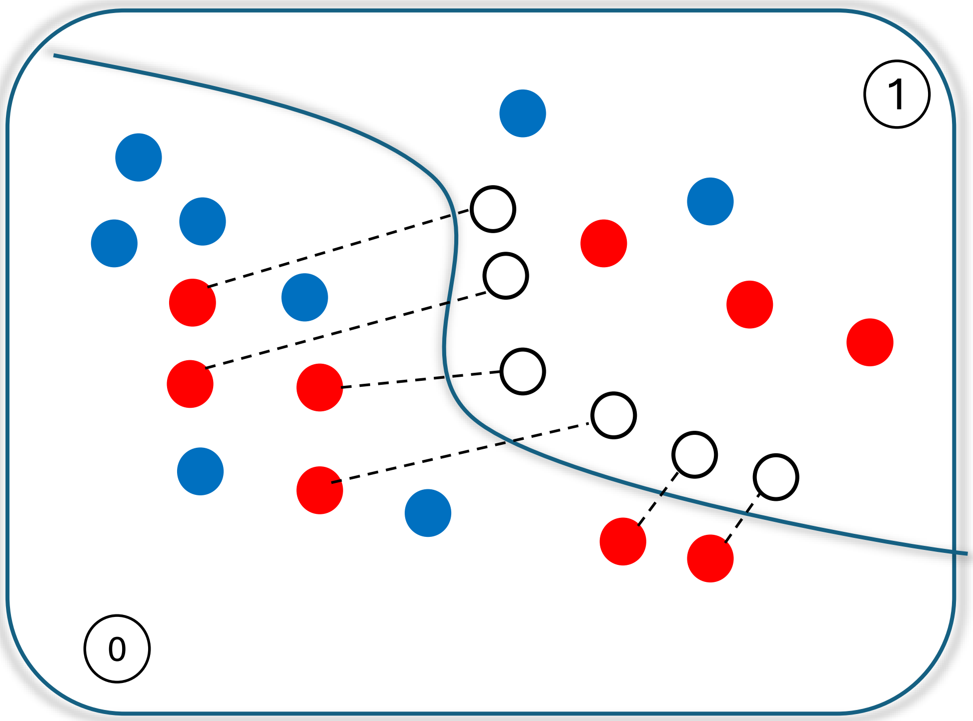
How can we leverage counterfactuals to identify, understand, and mitigate unfairness?

# Counterfactuals for explaining (un)fairness

Assume two groups: the blue $B$ and the red $R$

First Approach: 1-1 mapping

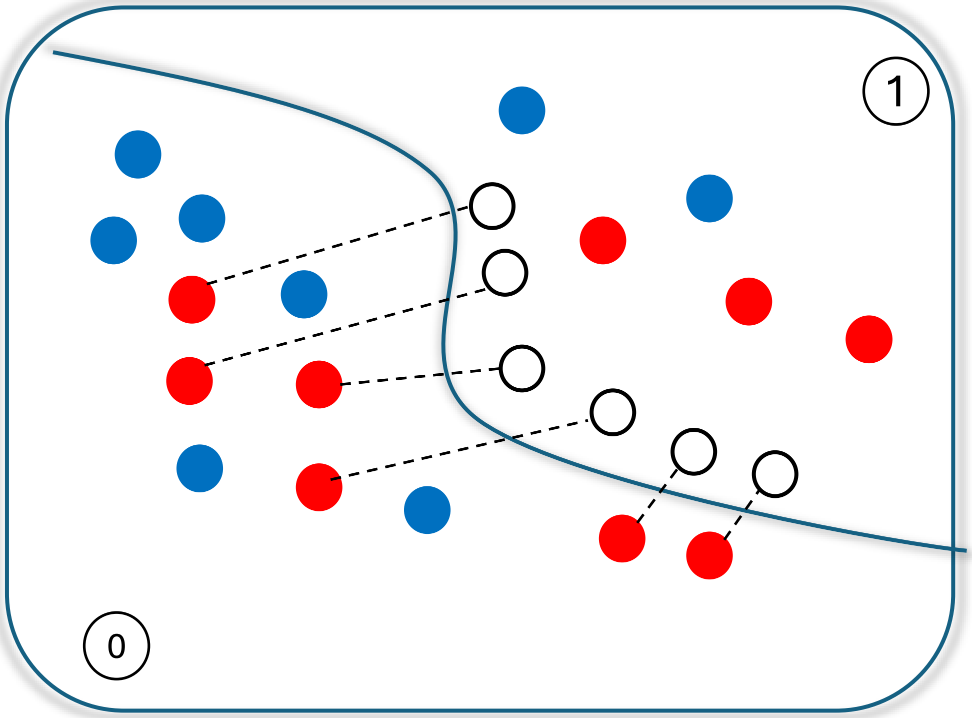
1. Generate counterfactuals for some $x \in R$
2. Generate counterfactuals for some $x \in B$
3. Explain group unfairness by aggregating the two sets of generated counterfactuals



Shubham Sharma, Jette Henderson, Joydeep Ghosh: CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. AIES 2020: 166-172Sofie Goethals, David Martens, Toon Calders: PreCoF: counterfactual explanations for fairness. Mach. Learn. 113(5): 3111-3142 (2024)
Alejandro Kuratomi, Evaggelia Pitoura, Panagiotis Papapetrou, Tony Lindgren, Panayiotis Tsaparas: Measuring the Burden of (Un)fairness Using Counterfactuals. PKDD/ECML Workshops (1) 2022: 402-417

# Counterfactuals for explaining (un)fairness

- How do we capture the *different definitions of fairness*?
- Do we allow *changing the protected attribute* in the counterfactual?
- How to *aggregate the explanations in the two sets* to understand, measure, and mitigate unfairness?



# Counterfactuals for explaining (un)fairness

## Different fairness definitions

Generate counterfactuals for different subsets of the two groups

Why not demographic parity: $P(\hat{Y} = 1 | x \in R) = P(\hat{Y} = 1 | x \in B)$

For the Negatives of each group,
explains why the predicted class $\widehat{Y}$ for each group is not the favorable on

Why not equal opportunity: $P(\hat{Y} = 1 | Y = 1, x \in R) = P(\hat{Y} = 1 | Y = 1, x \in B)$

For the False Negatives of each group,
explains why a positive instance is falsely classified in the negative class

# Counterfactuals for explaining (un)fairness

## Treatment of the protected attribute

Mutable or immutable?

Include them in the classification?

Sofie Goethals, David Martens, Toon Calders: PreCoF: counterfactual explanations for fairness. Mach. Learn. 113(5): 3111-3142 (2024)

# Counterfactuals for explaining (un)fairness

## Treatment of the protected attribute

Mutable or immutable?

- Explicit bias: Search for counterfactual explanations that update only the protected attribute
  - For example, if you have not been a woman, you would have received the loan

- Implicit bias: Remove the protected  attribute from the dataset before training the model

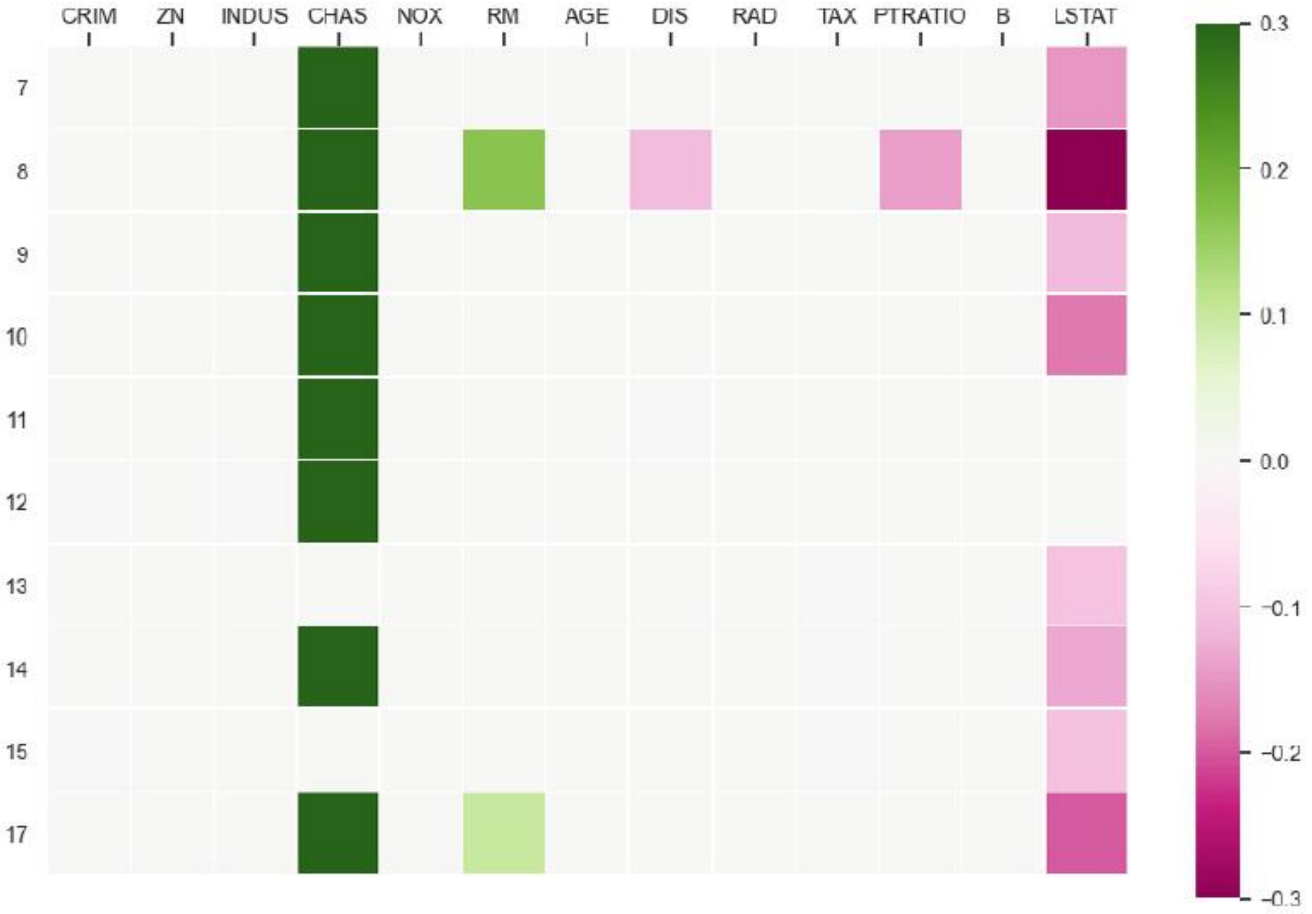
Sofie Goethals, David Martens, Toon Calders: PreCoF: counterfactual explanations for fairness. Mach. Learn. 113(5): 3111-3142 (2024)

# Counterfactuals for explaining (un)fairness

How to interpret the resulting set of counterfactuals

# Counterfactuals for explaining (un)fairness

Visualize the updates (and their volume) for the individual instances in each group



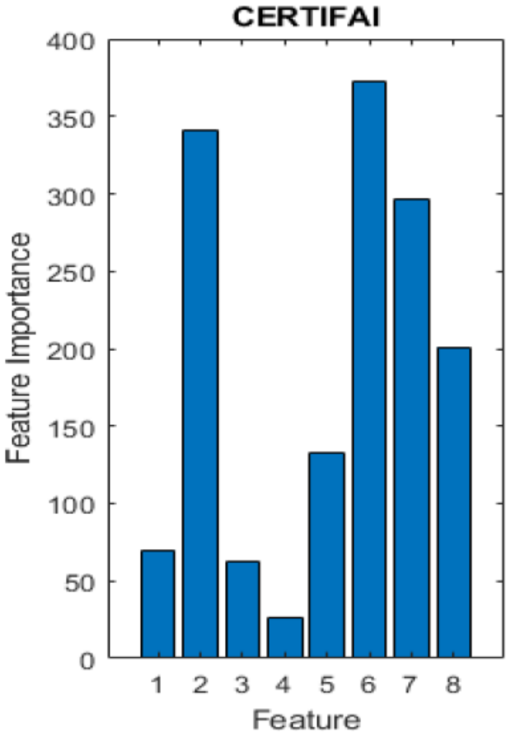
E. Carrizosa, J. Ramirez-Ayerbe, and D. R. Morales, “Mathematical optimization modelling for group counterfactual explanations,” European Journal of Operational Research, 2024.

# Counterfactuals for explaining (un)fairness

Feature attribution by aggregating the appearances of each feature in the counterfactuals

Aggregate the explanations

- calculate for **how many instances** in each group each feature was updated,
- compute the **average value of the update** for each feature



# Counterfactuals for explaining (un)fairness

Adult dataset: Predict whether income exceeds $50K/yr based on census data

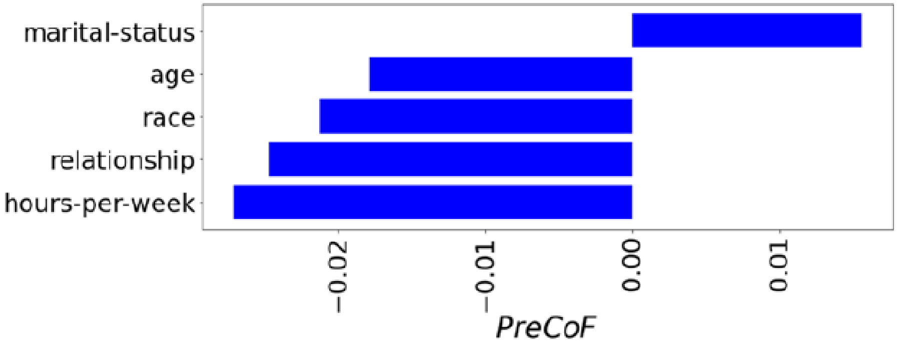
PreCoF: Only one feature is allowed to change

Explicit Bias

Aggregate the explanations by calculating for **how many instances of each group the protected attribute was updated**

*'If you would have **been a man**, you would have been predicted to have a high income'*
13 times
*'If you would have **been a woman**, you would have been predicted to have a high income'*)
1 time

Implicit Bias



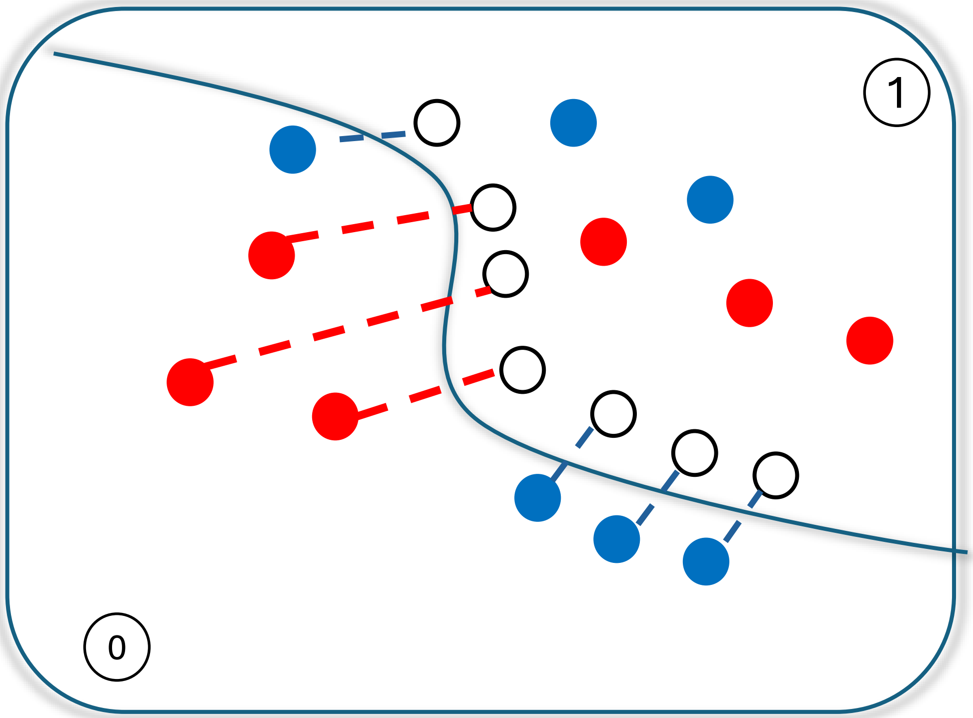
Difference between M and F

# Enhance Fairness Metrics: Burden

Cost (burden) for a group to switch the decision of the model

Simplest formulation:

$$Burden(G) = \frac{1}{|G|} \sum_{x_i \in G} distance(x_i, x_i')$$

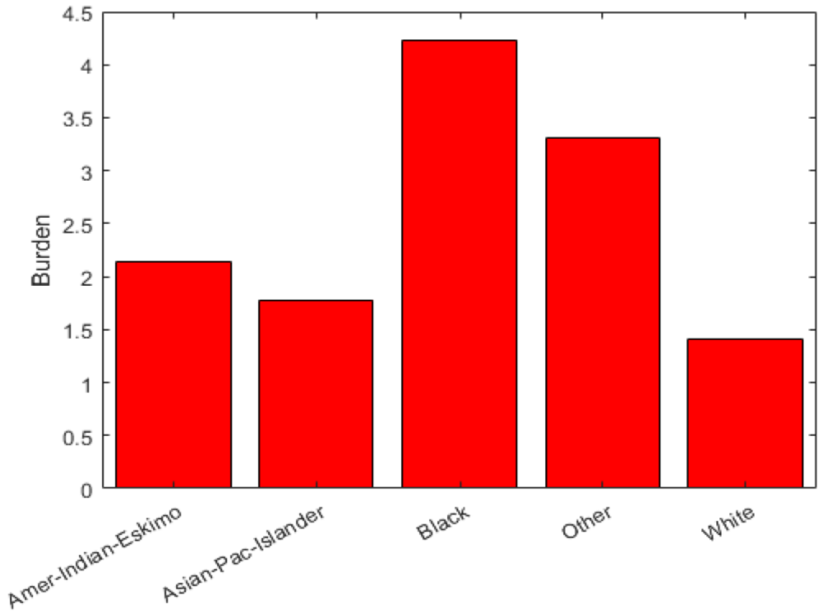


# Enhance Fairness Metrics: Burden

Cost (burden) for a group to switch the decision of the model

$$Burden(G) = \frac{1}{|G|} \sum_{x_i \in G} distance(x_i, x_i')$$

Adult dataset: Predict whether income exceeds $50K/yr based on census data

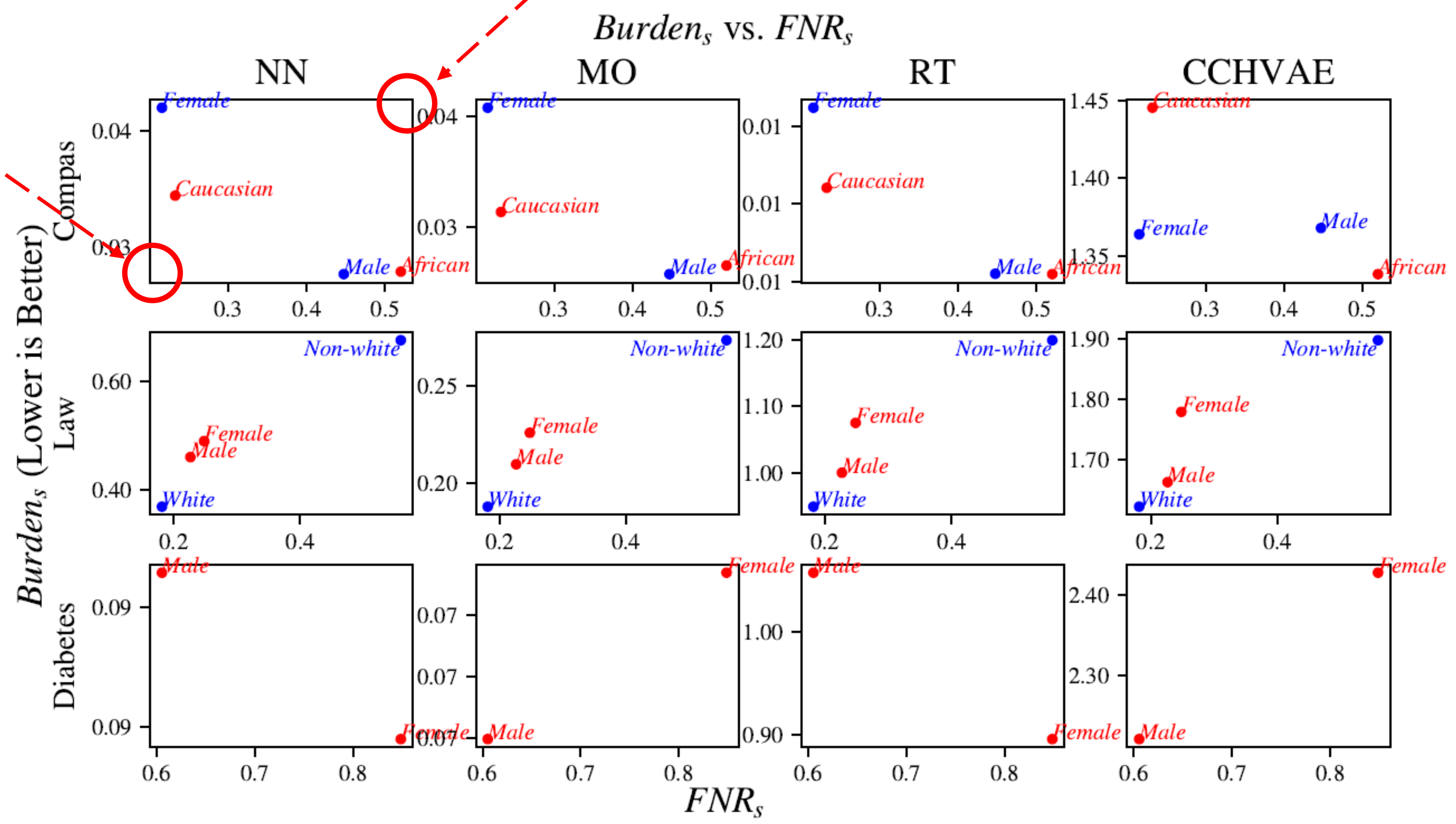


Burden on Black and the Other race is more than the other races. This means that on average, these groups would have to make more changes to achieve a desired prediction as compared to others

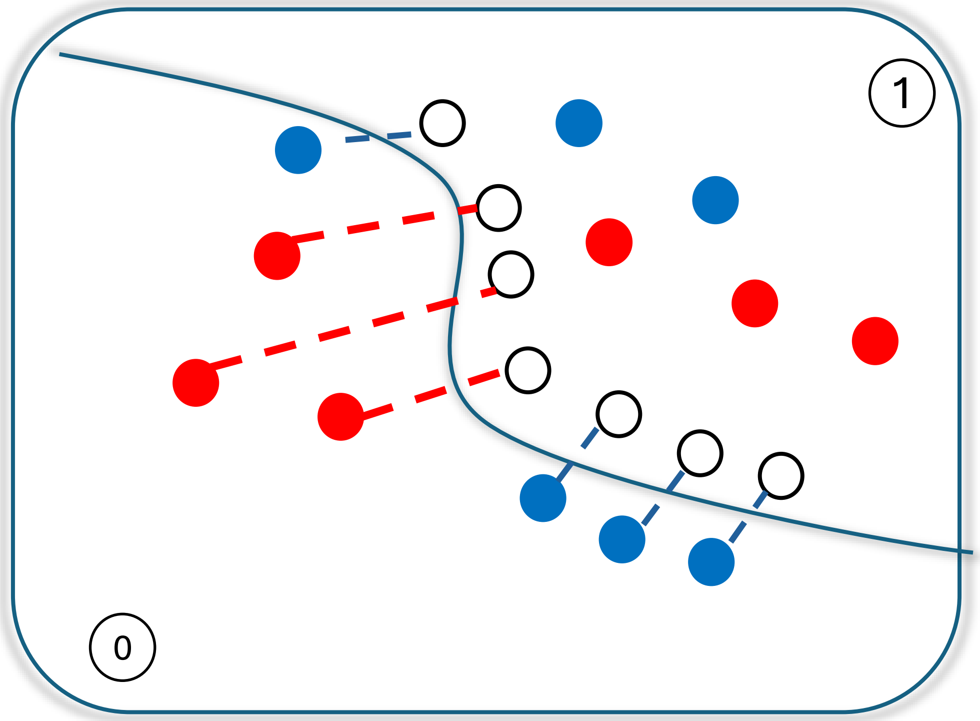
# Enhance Fairness Metrics: Burden

worst

best



# Enhance Fairness Metrics: Burden



## Burden and Robustness

Given two black-box models, M1 and M2,
if  in M1, the counterfactuals across classes are farther away from the input instances on average for M1 than M2,
then M1 would be harder to fool (more robust)

# Counterfactuals for explaining (un)fairness

## Further analysis

- Which categories of the protected group experience bias the most?
  - For example, subcategories with large burden
- Intersectional fairness
- Summary of counterfactuals: clustering, decision trees

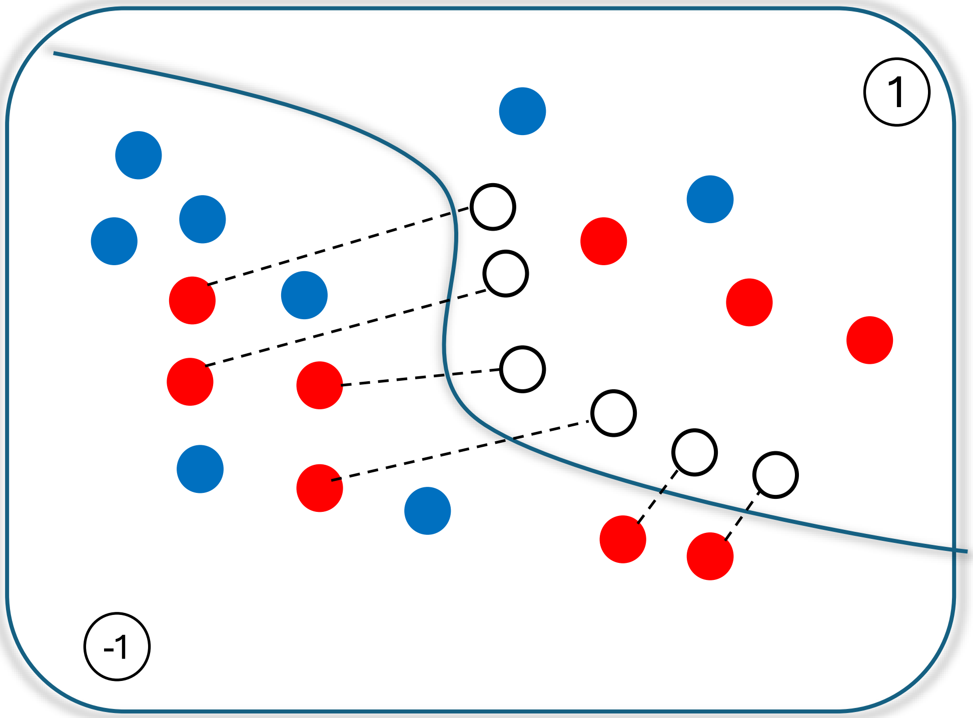
# Counterfactuals for explaining (un)fairness

So far,

1-1 mapping

1 counterfactual for 1 factual

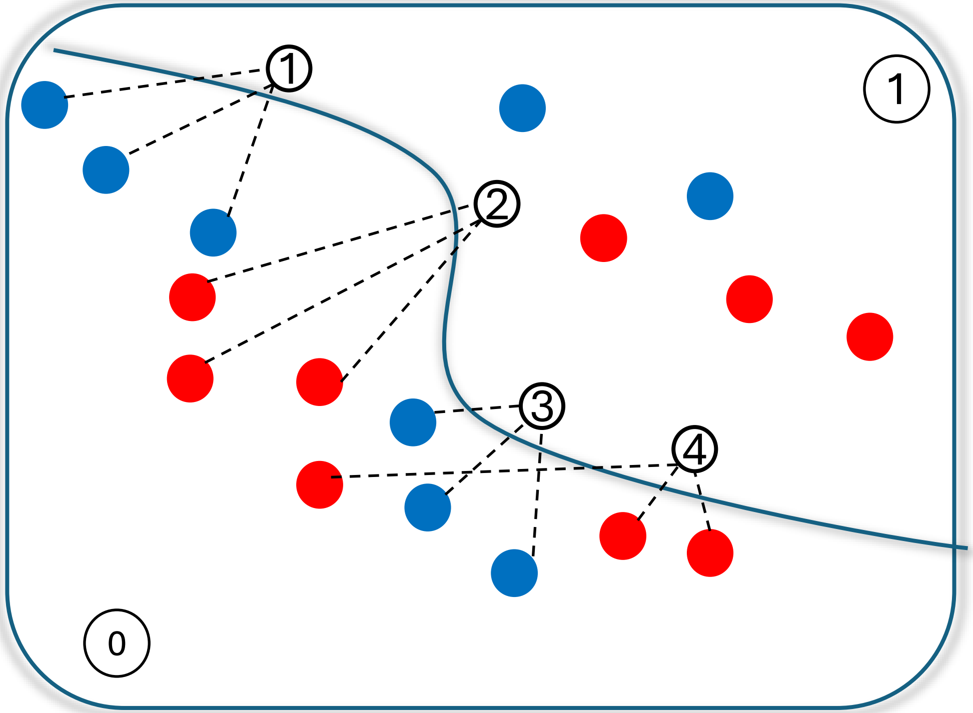
*What about counterfactuals for the group as a whole?*



# Counterfactuals for explaining (un)fairness

N-1 mapping

1 counterfactual for more than one factual



Second Approach: Group counterfactuals

1. Generate group counterfactual $R'$ for $R$
2. Generate group counterfactual $B'$ for $B$
3. Explain group unfairness by comparing $R'$ and $B'$

$$R' = \{2, 4\}$$
$$B' = \{1, 3\}$$

# Group counterfactuals

## Problem definition (high level)

Given a group $G$ of factuals, find a set G’ (i.e., the group counterfactuals) such that

(1) The size of $G'$ is small (interpretability, complexity, size)

(2) The cost of $G'$ is small (e.g., on average the distance of each $x \in G$ to each closest counterfactual)

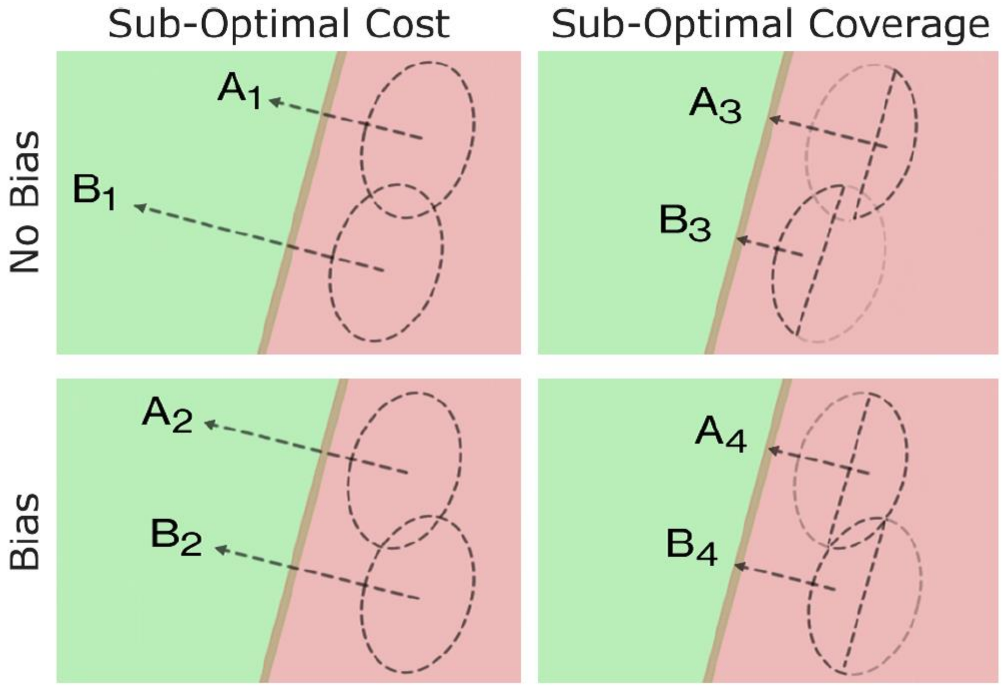
(3) The coverage of $G'$ is large, for each $x \in G$, there is at least one $x' \in G'$, s.t., $f(x') \neq f(x)$

# Why cost and coverage are important for fairness burden

Groups are shown as eclipses
Ai and Bi are the group counterfactual

Group counterfactuals show there is bias (different burdens), where there is not

Group counterfactuals show the is no bias (same burden), where there is



# Approaches to Group Counterfactuals

# Actionable recourse

So far, formulation of the CFE problem as

$$\arg min_{x' \in F}\, distance(x, x')\; s.t.\, f(x') \neq f(x)$$

Alternative formulation:

$$\arg min_{a\, \in\, \mathrm{A(x)}}\; cost(a;\; x')\; s.t.\, f(a(x))\; \neq f(x)$$

where $a$ is an action applied independently to $x$ to get the counterfactual $x'$of $x$

- Also called flipsets
- $A(x)$: set of feasible actions (often given as input)
- $cost$: cost function to choose between feasible actions

*Berk Ustun, Alexander Spangher, Yang Liu: Actionable Recourse in Linear Classification. FAT 2019: 10-19*

# Group counterfactuals based on actions

Based on the actionable recourse CFE formulation

$$\arg min_{a \in \mathrm{A(x)}} \; cost(a;\; x') \; s.t. f(a(x)) \; \neq f(x)$$

Group CFE formulation

Find a set $A$ of actions *a* such $|A|$ is small, $cost(A)$ is small and $coverage(A)$ is large

Action may be expressed, e.g., as conjunctions of predicates to be applied to $x$

Alice example

Features: $(Income, CreditScore, Education, Age)$

$x_{ALICE}$ = (35K,1K,BSc,22)

Actions: Income = 45K, Education = MSc, Income = 45K AND Education = MSc

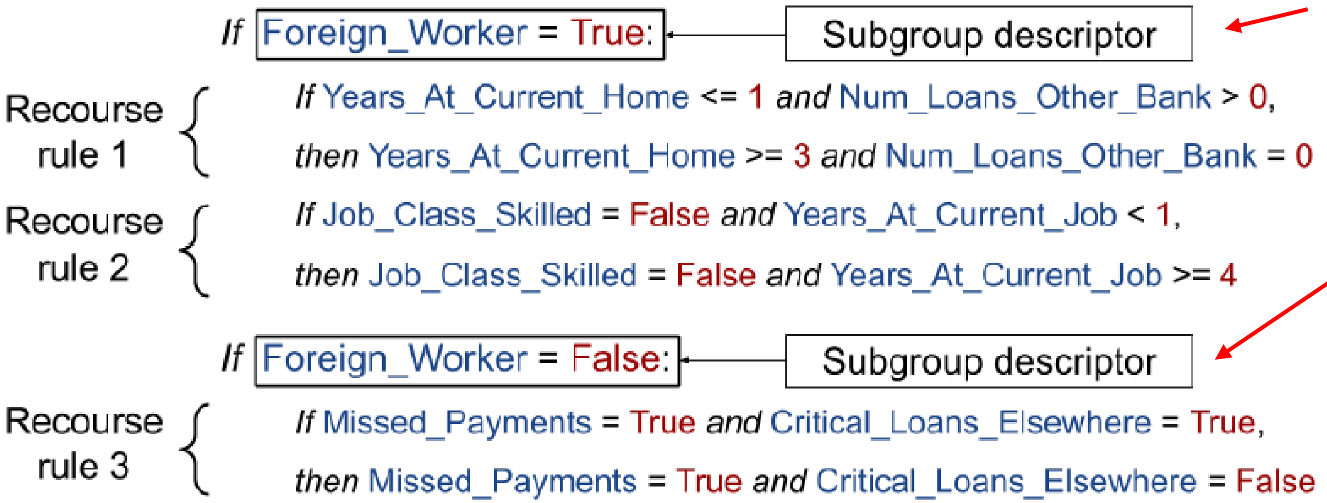
Output as a small set of actions, each action to be applied to a subset of the instances in G

# Actionable Recourse Summaries (AReS)

A hierarchical model: two levels

Recourse rule $r(c, c')$: for the instances that $c$ holds, to get the desirable output, apply action $c'$
$c, c'$ conjunctions of predicates

The inner if-then rules are the recourse rules



*If* Foreign_Worker = True: ← Subgroup descriptor

Recourse rule 1
*If* Years_At_Current_Home <= 1 *and* Num_Loans_Other_Bank > 0,
*then* Years_At_Current_Home >= 3 *and* Num_Loans_Other_Bank = 0

Recourse rule 2
*If* Job_Class_Skilled = False *and* Years_At_Current_Job < 1,
*then* Job_Class_Skilled = False *and* Years_At_Current_Job >= 4

*If* Foreign_Worker = False: ← Subgroup descriptor

Recourse rule 3
*If* Missed_Payments = True *and* Critical_Loans_Elsewhere = True,
*then* Missed_Payments = True *and* Critical_Loans_Elsewhere = False

The outer-if rules describe the subgroups

*Kaivalya Rawal, Himabindu Lakkaraju: Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. NeurIPS 2020*

# AReS: Algorithm

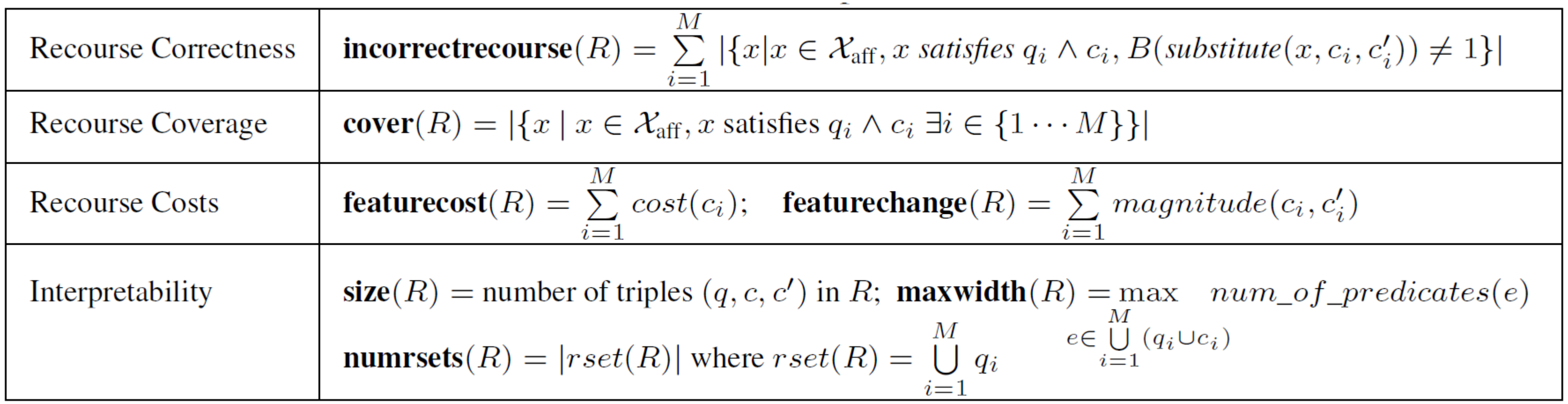
Input:

- Group $X_{aff}$ of instances that received unfavorable predictions
- the black box ML model $B$
- *candidate set of conjunctions of predicates* (e.g., age > 50 and gender = female) from which to pick the subgroup descriptors,
- *candidate set of conjunctions of predicates* from which to pick the recourse rules

If candidate sets are not provided, a frequent itemset mining algorithm is used such as apriori

As an optimization problem with an objective function that can jointly optimize for recourse correctness, coverage, costs, and interpretability

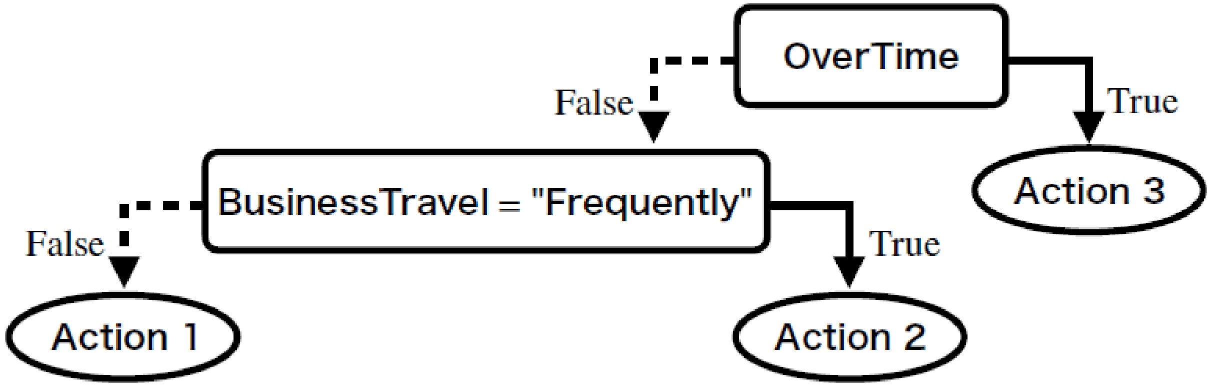
# AReS: Algorithm



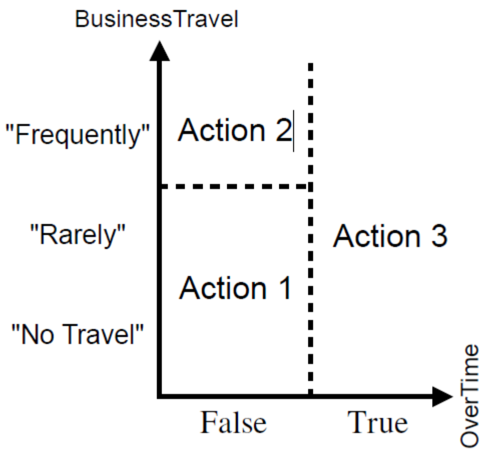
| | |
|---|---|
| Recourse Correctness | $\textbf{incorrectrecourse}(R) = \sum_{i=1}^{M} \lvert\{x \mid x \in \mathcal{X}_{\text{aff}}, x\ \textit{satisfies}\ q_i \wedge c_i, B(\textit{substitute}(x, c_i, c'_i)) \neq 1\}\rvert$ |
| Recourse Coverage | $\textbf{cover}(R) = \lvert\{x \mid x \in \mathcal{X}_{\text{aff}}, x \text{ satisfies } q_i \wedge c_i\ \exists i \in \{1 \cdots M\}\}\rvert$ |
| Recourse Costs | $\textbf{featurecost}(R) = \sum_{i=1}^{M} cost(c_i);\quad \textbf{featurechange}(R) = \sum_{i=1}^{M} magnitude(c_i, c'_i)$ |
| Interpretability | $\textbf{size}(R) =$ number of triples $(q, c, c')$ in $R$; $\textbf{maxwidth}(R) = \max_{e \in \bigcup_{i=1}^{M} (q_i \cup c_i)} num\_of\_predicates(e)$<br>$\textbf{numrsets}(R) = \lvert rset(R)\rvert$ where $rset(R) = \bigcup_{i=1}^{M} q_i$ |

$(q, c, c')$ $q$:subgroup descriptor, $(c, c')$ recourse rule
$cost(c)$: cost associated with an action; domain-dependent

# Counterfactual Explanation Trees (CET)



| | HowToChange | Effectiveness | |
|---|---|---|---|
| | | Cost | Flip rate |
| Action 1 | MonthlyIncome : + 1282$ | 0.17 | 83 % |
| Action 2 | BusinessTravel : "Frequently" → "Rarely" | 0.19 | 80 % |
| Action 3 | OverTime : True → False | 0.27 | 86 % |



A CET is a tree assigning an action to each instance in the group

*Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, Yuichi Ike: Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees. AISTATS 2022: 1846-1870*

# CET: Algorithm

Set of possible actions are given as input

Learns a CET by optimizing objectives on cost, coverage, correctness (called validity), size (number of leaves)

Uses stochastic local search

# Fairness Aware Counterfactuals for Subgroups (FACTS)

Emphasis is on providing refined definitions of fairness

- In the micro viewpoint, the instances in the group are considered independently, and *each may choose the action* that benefits itself the most (1-1 mapping between instances and actions)
- In the macro viewpoint, the group is considered as a whole, and *an action is applied collectively to all* instance in the group (a single action for the whole group)

*Loukas Kavouras, Konstantinos Tsopelas, Giorgos Giannopoulos, Dimitris Sacharidis, Eleni Psaroudaki, Nikolaos Theologitis, Dimitrios Rontogiannis, Dimitris Fotakis, Ioannis Z. Emiris: Fairness Aware Counterfactuals for Subgroups. NeurIPS 2023*

# FACTS: Fairness definitions

Several new fairness definitions for both the micro and macro level based on how difficult it is for a group to achieve recourse

Equal Effectiveness (aka coverage)
A classifier is fair if the same proportion of individuals in the protected subgroups can achieve recourse

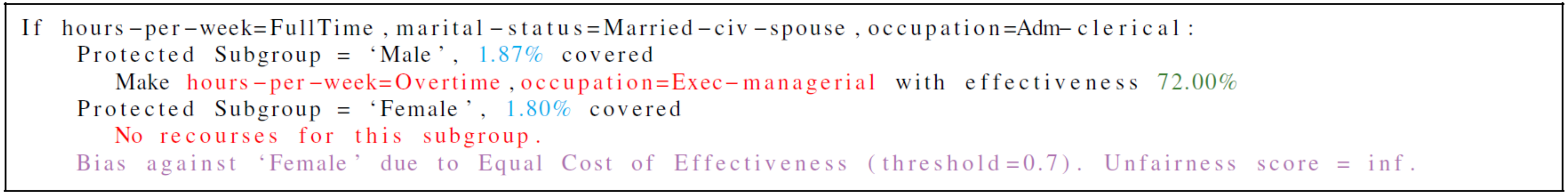
Equal Choice for Recourse (only for the macro)
A classifier is fair if the groups can choose among the same number of sufficiently effective actions to achieve recourse, where sufficiently effective means the actions should work for at least ϕ% of the instances

Equal Effectiveness within Budget
The classifier is fair if the same proportion of individuals in the protected subgroups can achieve recourse with a cost within a specified budget

Fair Effectiveness-Cost Trade-Off
The classifier is fair if the protected subgroups have the same effectiveness-cost distribution

# FACTS: algorithm

Frequent itemset algorithm (FP Growth) for both the predicates and the actions



```
If hours-per-week=FullTime, marital-status=Married-civ-spouse, occupation=Adm-clerical:
    Protected Subgroup = ‘Male’, 1.87% covered
       Make hours-per-week=Overtime, occupation=Exec-managerial with effectiveness 72.00%
    Protected Subgroup = ‘Female’, 1.80% covered
       No recourses for this subgroup.
    Bias against ‘Female’ due to Equal Cost of Effectiveness (threshold=0.7). Unfairness score = inf.
```

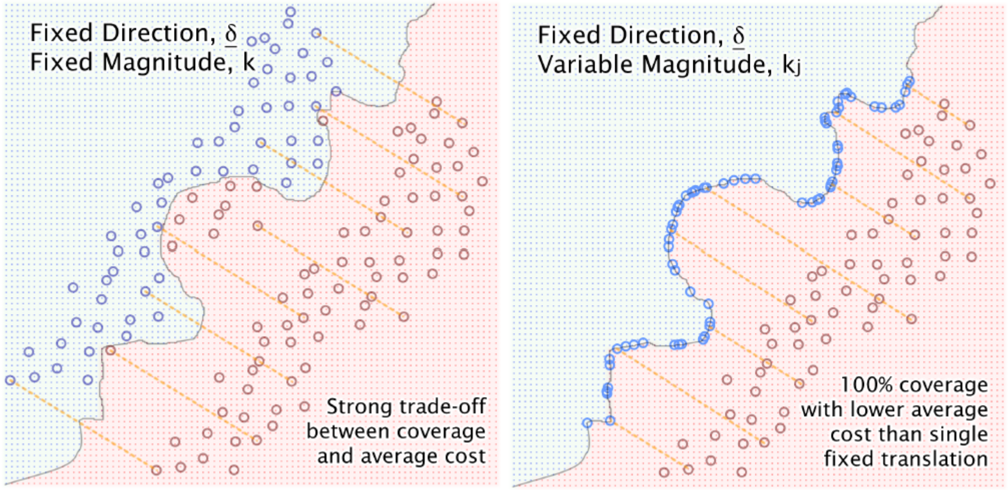
Covered refers to the percentage of population that satisfies the predicate

# GLOBE-CE: A Translation Based Approach

$$\arg min_{x' \in F}\ distance(x, x')\ s.t.\ f(x') \neq f(x), x' = x + k\ \delta$$

where $\delta$ is a translation vector and $k$ a scalar

Previous work

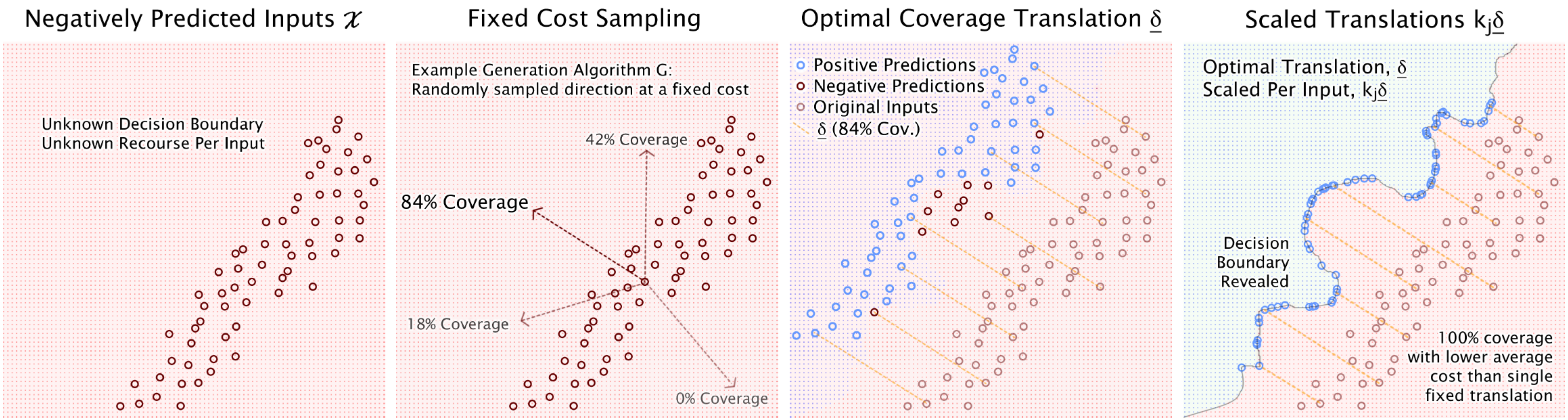

Fixed Direction, δ
Fixed Magnitude, k
Strong trade-off between coverage and average cost
Fixed Direction, δ
Variable Magnitude, kj
100% coverage with lower average cost than single fixed translation

GLOBE-CE

*Dan Ley, Saumitra Mishra, Daniele Magazzeni: GLOBE-CE: A Translation Based Approach for Global Counterfactual Explanations. ICML 2023: 19315-19342*

# GLOBE-CE: Algorithm

(a) Assign a single vector direction ($\delta$) to an entire group of inputs,
(b) Travel along this vector, and
(c) analyze the minimum costs required for successful recourses per instance in the subgroup


Negatively Predicted Inputs $\mathcal{X}$
Unknown Decision Boundary
Unknown Recourse Per Input
Fixed Cost Sampling
Example Generation Algorithm G:
Randomly sampled direction at a fixed cost
42% Coverage
84% Coverage
18% Coverage
0% Coverage
Optimal Coverage Translation $\underline{\delta}$
Positive Predictions
Negative Predictions
Original Inputs
$\underline{\delta}$ (84% Cov.)
Scaled Translations $k_j\underline{\delta}$
Optimal Translation, $\underline{\delta}$
Scaled Per Input, $k_j\underline{\delta}$
Decision Boundary Revealed
100% coverage with lower average cost than single fixed translation

# Counterfactuals and Actionable Recourse

## Main points so far

- Counterfactuals (recap)
- “Rename” them as actionable recourse
- Relate them with fairness:
  - 1-1 mapping and N-1 mapping (group counterfactuals)
  - burden, or cost for recourse
- Algorithms to generate group counterfactuals

# From actionable recourse to interventions

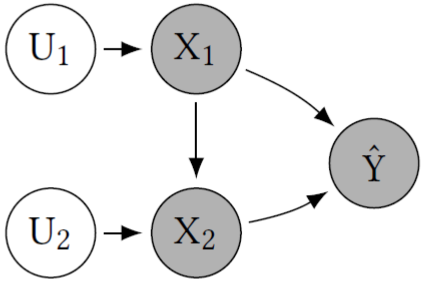
- Actionable recourse ignores dependencies among the features
- Furthermore, it can produce actions with suboptimal cost

*Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, Isabel Valera: A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. ACM Comput. Surv. 55(5): 95:1-95:29 (2023)*
*Amir-Hossein Karimi, Bernhard Schölkopf, Isabel Valera: Algorithmic Recourse: from Counterfactual Explanations to Interventions. FAccT 2021: 353-362*

# From actionable recourse to interventions

Example

Assume a Structural Causal Model (SCM) that captures inter-variable causal relationships



$$\left.\begin{aligned} X_1 &:= U_1 \\ X_2 &:= f_2(X_1) + U_2 \end{aligned}\right\} \mathcal{M}$$

$$\hat{Y} = h(X_1, X_2)$$

$X_1$ annual salary
$X_2$ bank balance
$\hat{Y}$ is the output of a fixed deterministic predictor $h$ that predicts eligibility for a loan

- Alice with $X_1$ = \$75K and $X_2$= \$25K applies for a loan

Simple linear binary classifier ($h = \mathrm{sgn}(X_1 + 5 \cdot X_2 - \$225{,}000)$ is used and Alice is denied the loan

- Counterfactual explanations. $X'_1$ = \$100K (+%33) or $X_2$= \$30K (+%20)

Alice is encouraged to reapply when either of these conditions are met

- Actions take place in a world where home-seekers save %30 of their salary (i.e., $X_2$ =3/10 · $X_1$ + $U_2$ )

A salary \$85 (14%K) would automatically result in \$3K additional savings, with a net positive effect on the loan-granting decision

- So Alice should apply earlier (suboptimal solution)

# From actionable recourse to interventions

- Given a SCM propose that actions should be carried out through structural interventions

- A structural intervention can be though of as a transformation between SCMs

# Shapley values

# Shapley values (recap)

Set $N$ **of payers** $i$: 🟩 🔴 ▲

Cooperative **game** $g$ : players forge coalitions to achieve a common goal

Example: Kaggle competition, factories working to produce a common good, etc

**Utility function** $u_g$**:** The output of the game, some measure of the performance

After the game is over, the coalition gets a certain payout/benefit/gain for the results

How should the value function be distributed among the players?

*Lloyd Shapley, A value for n-person games, in Contributions to the Theory of Games, 1953*

# Shapley values

Set $N$ **of payers** $i$: 🟩 🔴 ▲

Proposal 1: Equal distribution among the players

$u_g$( 🟩 ) = $u_g$( 🔴 ) = $u_g$( ▲ )

Is this a good idea?

Some players may *contribute more* to the coalition than others (for example, an ML expert in the Kaggle team)

Rephrased question:

How can we estimate the contribution of each player?

*Lloyd Shapley, A value for n-person games, in Contributions to the Theory of Games, 1953*

# Shapley values

Set $N$ **of payers** $i$: 🟩 🔴 ▲

Value function $\varphi_g(i)$

$\varphi_g($🟩$)$ : the contribution of player 🟩 to the game

Proposal 2: Leave-one-out?

$$\varphi_g(🟩) = u_g(\{🟩, ▲, 🔴\}) - u_g(\{▲, 🔴\})$$

In general:

$$\varphi_g(i) = u_g(N) - u_g(N \backslash i)$$

Is this a good idea?

*Lloyd Shapley, A value for n-person games, in Contributions to the Theory of Games, 1953*

# Shapley values

Set $N$ **of payers** $i$: 🟩 🔴 ▲

## Desired properties:

Zero element: if adding 🟩 to any subset has no impact on utility

$$\varphi_g(\text{🟩}) = 0$$

Symmetry: if adding 🟩 or 🔴 to any subset of data always leads to the same change on utility

$$\varphi_g(\text{🟩}) = \varphi_g(\text{🔴})$$

Efficiency: the utility shall be fully allocated to all players

$$\varphi_g(\text{🟩}) + \varphi_g(\text{🔴}) + \varphi_g(\text{▲}) = u_g(N)$$

Additivity: the utility in two tasks (games) shall be the sum of the utilities in each task

*Lloyd Shapley, A value for n-person games, in Contributions to the Theory of Games, 1953*

# Shapley Values

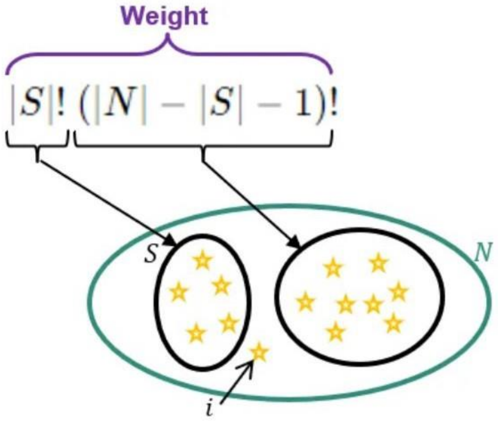
Given a set of players N, for all possible coalitions S of players, we get two values:

- Including the player $i$: $u_g(S \cup i)$
- Excluding the player $i$: $u_g(S)$

The Shapley value of player *i* is a weighted average of the marginal contributions of *i* over all subsets *S* of *N*.

for a subset *S*, the weight is the *product* of the number of permutations of *S* and the number of permutations of the complement of *S* and *i* (i.e.; N\{S∪{i}})

the marginal contribution of player *i* to the subset *S*



$$) = \sum_{S \subseteq N\setminus\{i\}} \frac{|S|!\;(|N|-|S|-1)!}{|N|!}\left[u_g(S \cup i) - u_g(S)\right]$$

|*N*|! is the number of permutations of the set *N*

# Shapley Values in Classification

Game: ?
Players: ?
Utility: ?

# Shapley Values in Classification

Game: prediction model $f_y(x)$

Players: The features

Utility: The performance (prediction) of the model

- Shapley values tell us how to distribute the prediction of the model to the features
- Feature attribution method: contribution of each feature in the prediction

# Shapley Values in Classification

Assume a binary label $y$, a prediction model $f_y(x)$

Value function: $u_{f_y(x)}(S)$

To *explain an input instance $x$*, we get for all possible subsets S of features two values:

- Including the feature $i$: $u_{f_y}(x_{s\cup\{i\}})$
- Excluding the feature $i$: $u_{f_y}(x_s)$

Local Shapley value

$$\varphi_{f_y(x)}(i) = \sum_{\mathrm{S} \subseteq \mathrm{N}\setminus\{i\}} \frac{|\mathrm{S}|!\ (|\mathrm{N}| - |\mathrm{S}|\ - 1)!}{|\mathrm{N}|!} [u_{f_y}(x_{s\cup\{i\}}) - u_{f_y}(x_s)]$$

# Shapley Values in Classification

Global Shapley Values

$$\Phi_f(i) = \mathrm{E}_{p(x,y)}[\varphi_{f_y(x)}(i)]$$

Global explanation: Contribution of feature $i$ to the model predictions

# Shapley Values in Classification

## How to exclude a feature from a ML model?

- We cannot just remove a feature, will affect the representation
- Key idea: instead of removing a feature, set its value to a random value

## How to calculate the Shapley values?

Two important challenges

- The number of utility evaluations is exponential to the number of features
- The computation cost for a single utility evaluation may be high

Solution: Sampling and approximate computation

# Shapley Values in Fairness

Game: ?

Players: ?

Utility: ?

# Shapley Values in Fairness

Game: Prediction model $f_y(x)$

Players: The features

Utility: The fairness of the model

# Shapley Values in Fairness

Demographic parity

$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$

Quantification of (dis)parity:

$$\Delta_{DP} = P(\hat{Y} = 1 | A = 1) - P(\hat{Y} = 1 | A = 0)$$

Use signed difference to show which group is privileged

Attribute based explanations for demographic disparity: Use the difference

$$\Phi_f(i) = \mathrm{E}_{X|A=1}\left[\Phi_{f(x)}(i)\right] - \mathrm{E}_{X|A=0}\left[\Phi_{f(x)}(i)\right]$$

*Tom Begley, Tobias Schwedes, Christopher Frye, Ilya Feige: Explainability for fair machine learning. CoRR abs/2010.07389 (2020)*

# Shapley Values: Mitigation

Main idea

Use linearity of Shapley values:

Fairness of Shapley values of a linear ensemble of models are the corresponding linear combinations of Shapley values of the underlying models

How:

Learn an additive perturbation of an existing model to impose fairness (in-processing approaches)

# Shapley Values: Mitigation

Given $f$ learn perturbation $\delta_\theta$ to make $f$ fair

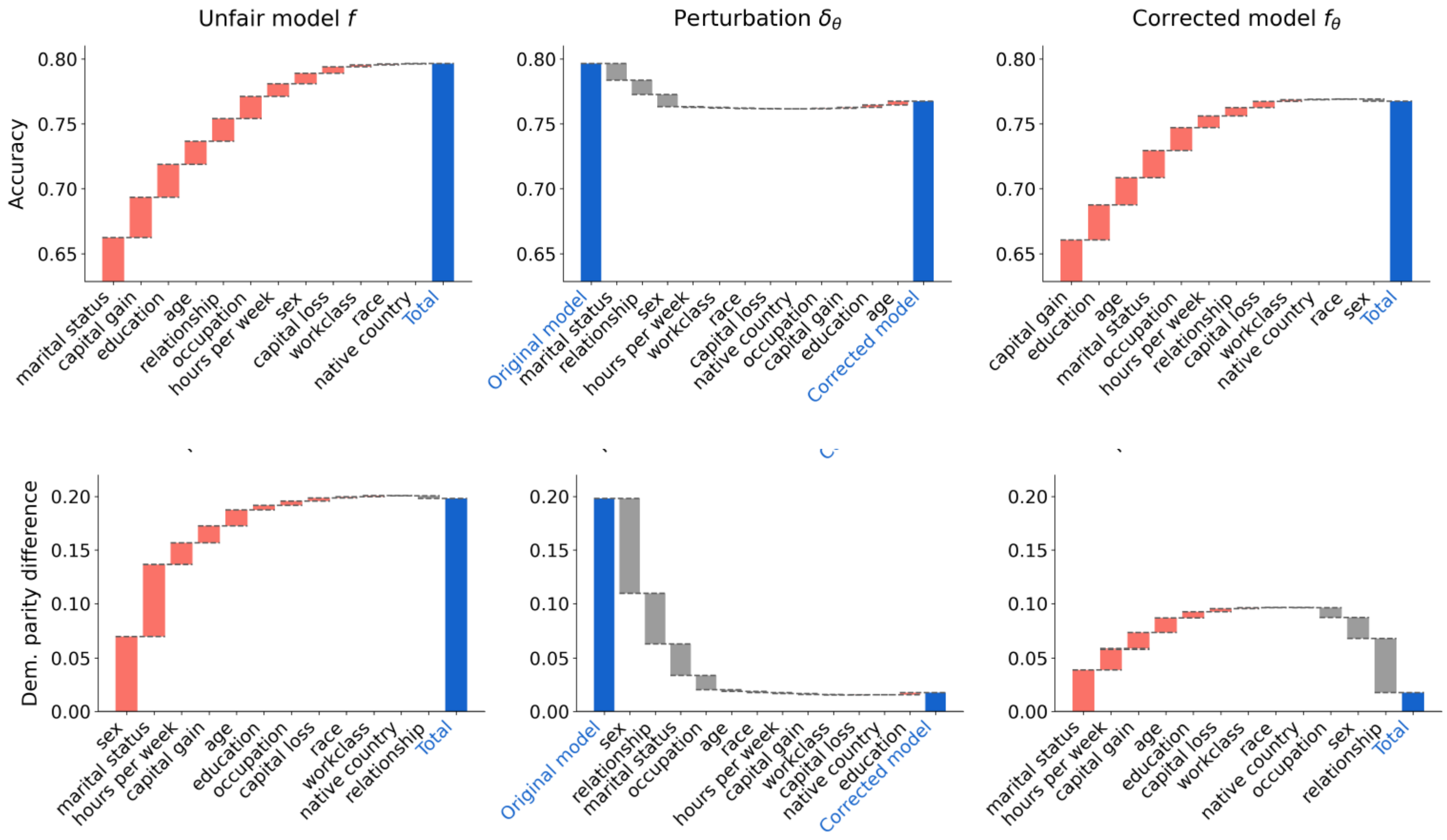
$$\underset{\text{fair}}{f_\theta} = \underset{\text{original}}{f} + \underset{\text{perturbation}}{\delta_\theta}$$

Explanations for $f$ : why is the original model fair?
Explanations for $f_\theta$: why is the “corrected” model fair?

$\delta_\theta$: corrections/trade-off

# Shapley Values: Mitigation

Adult dataset: Predict whether income exceeds $50K/yr based on census data



Unfair model $f$

Perturbation $\delta_\theta$

Corrected model $f_\theta$

# Shapley Values in Fairness

Game: A  prediction model $f_y(x)$

Players: The features

Utility: The fairness of the model

But what about causal relationships among features?

Game: A  prediction model $f_y(x)$

Players: Paths modeling dependencies among features

Utility: The fairness of the model

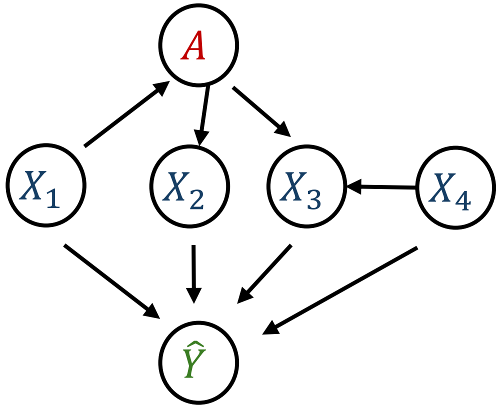
# Shapley Values: Path-specific explanations

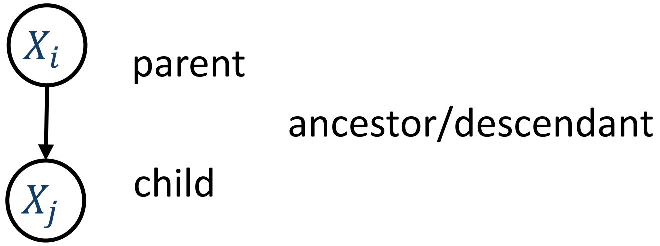
$A$: protected (binary) attribute
$X = \{X_{1,} \dots, X_M\}$: input features
$Y$: output, $\hat{Y}$: predicted output

Input causal graph $G$:
obtained based on domain knowledge or learned from the training data with existing causal discovery algorithms

Causal graph

Nodes: Variables
Edges: Variable relations





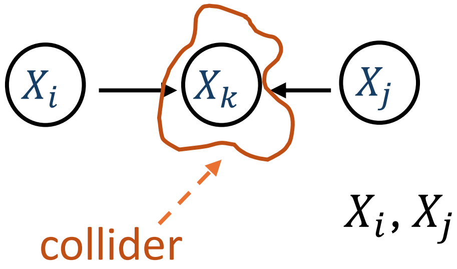
parent

ancestor/descendant

child

$\hat{Y}$ is the child of all $X_i$'s

Causal path: directed path



collider

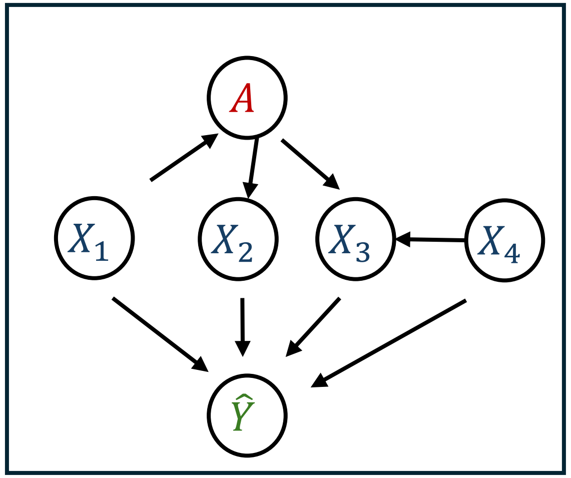
$X_i, X_j$ independent

Active paths relative to a conditioning set of nodes C may contain colliders if the collider, or any of its descendants belong to the conditioning set C

Assumption: Faithfulness
For all $X_i$, $X_j$, C, $X_i$ is conditional independent with $X_j$ on C, if there exists **no active path** from $X_i$ to $X_j$ in the graph.

*Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, Fei Wang: Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition. KDD 2021: 1287-1297*

# Shapley Values: Path-specific explanations



Demographic parity

$$P(\hat{Y} = 1 | A = 1) = P(\hat{Y} = 1 | A = 0)$$

Quantification of (dis)parity:

$$\Delta_{DP} = P(\hat{Y} = 1 | A = 1) - P(\hat{Y} = 1 | A = 0)$$

Use signed difference to show which group is privileged

**Assumption:**
**Any non-zero $\Delta_{DP}$ comes from the set P of feasible active paths (FACTs) linking the protected attribute $A$ and $\widehat{Y}$**

*Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, Fei Wang: Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition. KDD 2021: 1287-1297*

# Shapley Values: Path-specific explanations

Decompose the contribution to unfairness among the FACTs in P

Local explanation for path $p_i$

$\varphi_{f(x)}(p_i)$: the contribution of individual active path $p_i$ to $f(x)$

$$\Phi_f(p_i) = \mathrm{E}_{x|A=1}\left[\varphi_{f(x)}(p_i)\right] - \mathrm{E}_{x|A=0}\left[\varphi_{f(x)}(p_i)\right]$$

The computation $\varphi_{f(x)}(p_i)$ involves considering permutations of paths in P

# Shapley Values: Path-specific mitigation

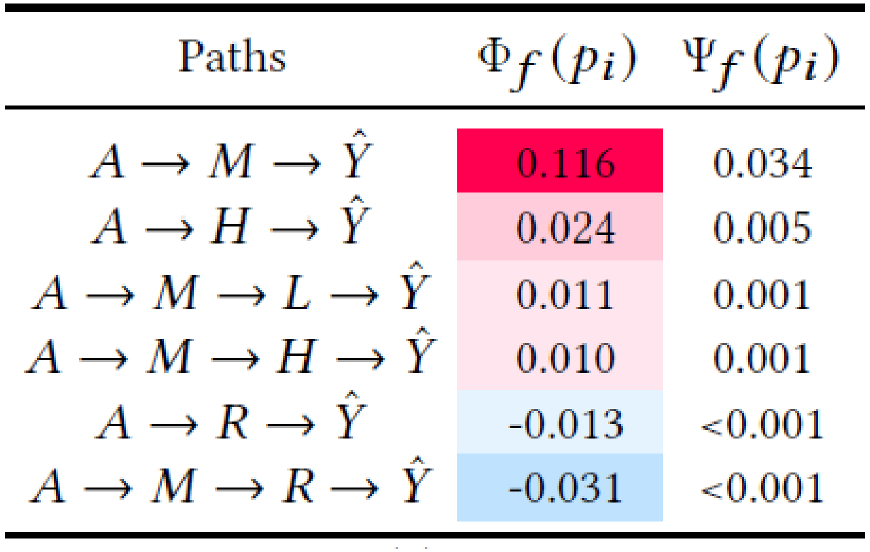
- Decompose the utility of the model by: $U(f) = \Psi_f(\emptyset) + \sum_{p_i \epsilon P} \Psi_f\ (p_i)$

- Train a fair model with a subset of paths $(T)$ by minimizing:

$$L(T) = -\sum_{p_i \epsilon T} \Psi_f\ (p_i) + \lambda | \sum_{p_i \epsilon \mathrm{T}} \Phi_f(p_i) |$$

*Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, Fei Wang: Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition. KDD 2021: 1287-1297*

# Shapley Values: Path-specific mitigation



| Paths | $\Phi_f(p_i)$ | $\Psi_f(p_i)$ |
|---|---|---|
| $A \rightarrow M \rightarrow \hat{Y}$ | 0.116 | 0.034 |
| $A \rightarrow H \rightarrow \hat{Y}$ | 0.024 | 0.005 |
| $A \rightarrow M \rightarrow L \rightarrow \hat{Y}$ | 0.011 | 0.001 |
| $A \rightarrow M \rightarrow H \rightarrow \hat{Y}$ | 0.010 | 0.001 |
| $A \rightarrow R \rightarrow \hat{Y}$ | -0.013 | <0.001 |
| $A \rightarrow M \rightarrow R \rightarrow \hat{Y}$ | -0.031 | <0.001 |

Adult dataset (with a subset of features)
Top-paths

A: sex, M: marital status, L: level of education, H: working hours per week, R: relationship

*Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, Fei Wang: Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition. KDD 2021: 1287-1297*

# Fairness of explanation methods

# Fairness of Explanations

What does it mean?

# Fairness of Explanations

Individual fairness

If two instances are similar, they should receive similar explanations

Error-based (accuracy-based) group fairness

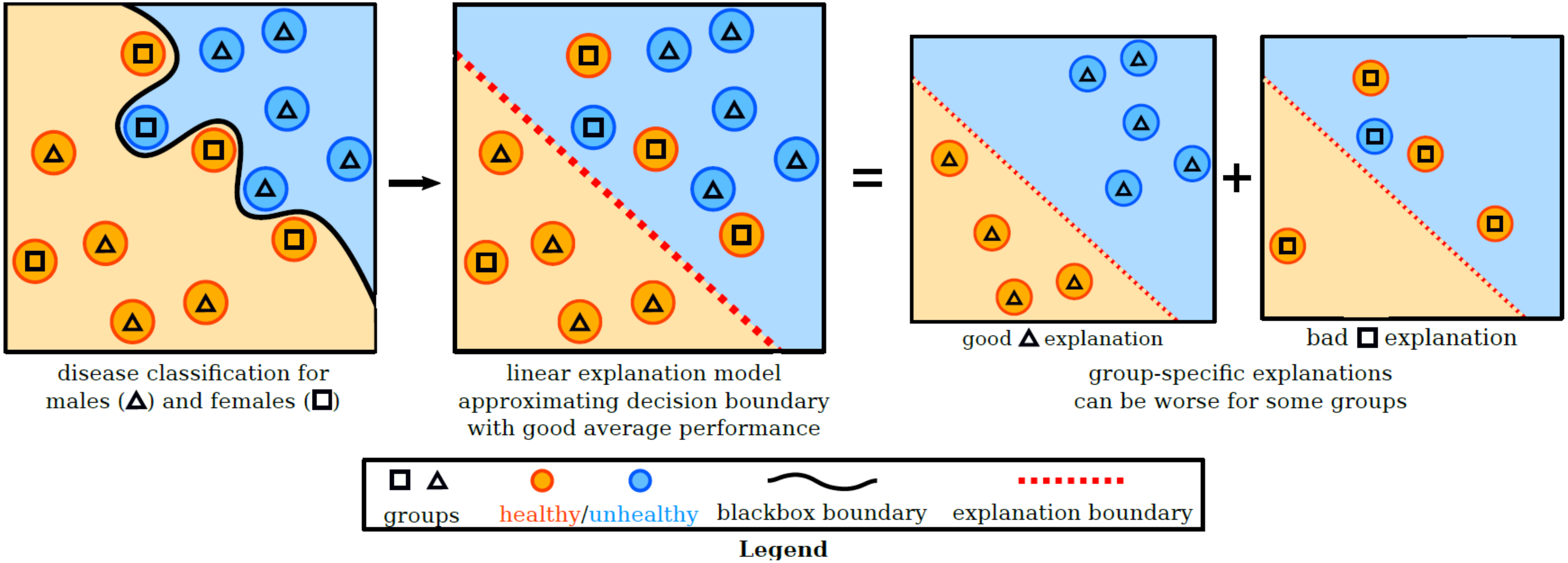
Given group $R$ and $B$, an explanation method is *group fair* when it produces *equally good explanations* for both groups.

# Quality of Explanations

Measures of explanation quality

- Sparsity: how well an explanation can be understood and interpreted by users and is often measured by the *complexity, or size* of the explanations, e.g., number of important features
- Fidelity: that measures how well the explanations *capture the model* they explain. e.g., for methods that create an interpretable surrogate model to explain a black-box model, fidelity compares the prediction of the surrogate and the original model on the instances used to train the original model
- Stability: asks that similar instances receive similar explanations
- Consistency: if an explanation for an instance is calculated multiple times, each of the calculated explanations should be similar.

# Quality of Explanations: Fidelity


→
=
+
good △ explanation
bad □ explanation
disease classification for males (△) and females (□)
linear explanation model approximating decision boundary with good average performance
group-specific explanations can be worse for some groups
□ △ groups
healthy/unhealthy
blackbox boundary
explanation boundary
Legend

*Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. FAccT 2022. 1194–1206.*

# Quality of Explanations: Fidelity

Two metrics:

- maximum fidelity gap from average that measures the extent to which the fidelity of the disadvantaged group is lower than the average, and
- mean fidelity gap among subgroups that measures how much the fidelity differs across subgroups.

Several observations

- The fidelity gaps are largest for the least-fair black boxes.
- However, even when training fair models, fidelity gaps are still observed.
- To reduce fidelity gaps across groups they propose robust training but also data-distribution aware training methods that leverages causal knowledge.

*Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. FAccT 2022. 1194–1206.*

# Procedural-oriented fairness

*Procedural-oriented fairness* measures the gap between the explanation quality for different subgroups where quality is measured by fidelity (attribution-based methods)

$$L(G) = L_u(G) + \alpha\, L_f(G_0, G_1) + \beta\, L_e(G_0, G_1)$$

utility

fairness

explanation fairness

Main idea is to look at the hidden representations of the instances for each group

- For fairness, the goal is to minimize the difference (gap) between the hidden representations of instances from the two groups with the same label
- For explanation fairness, the gap is considered small if the hidden representations for instances with/without masking the most important features in subgroups are close to each other.

*Yuying Zhao, Yu Wang, Tyler Derr: Fairness and Explainability: Bridging the Gap towards Fair Model Explanations. AAAI 2023: 11363-11371*

# Fairness of Explanations: empirical results

Regarding the quality of post-hoc explanations.

- Disparities in the quality of explanations between groups
- Such disparities are more likely to occur when the models being explained are complex and non-linear

*Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, Marzyeh Ghassemi: The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. FAccT 2022: 1194-1206*
*Jessica Dai, Sohini Upadhyay, Ulrich Aïvodji, Stephen H. Bach, Himabindu Lakkaraju: Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post hoc Explanations. AIES 2022: 203-214*

# Individual Fairness: Robustness

Are counterfactual explanation robust?

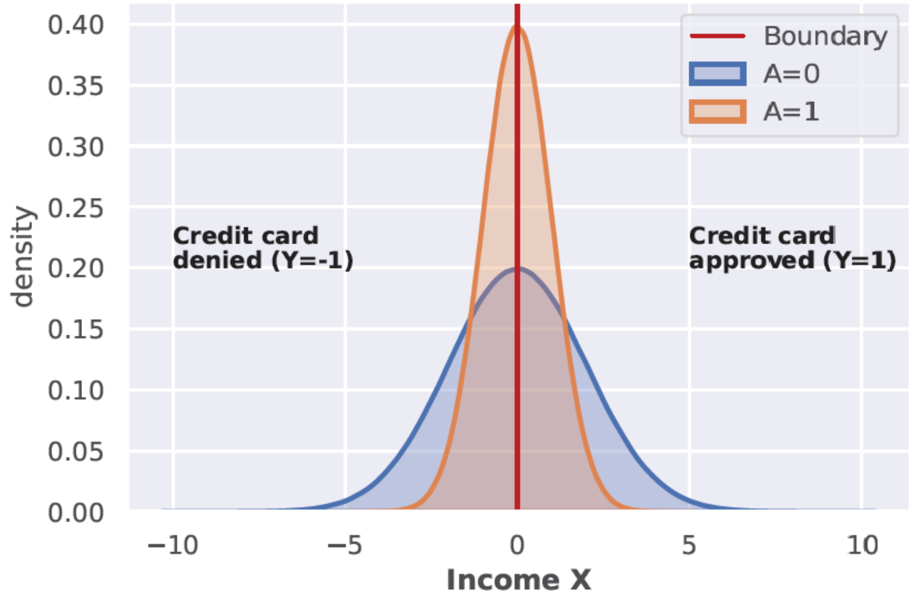
Individual fairness: Do similar factuals get counterfactuals with similar costs?

What is the result  of small perturbation of a factual at the counterfactual?

*André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, Barbara Hammer: Evaluating Robustness of Counterfactual Explanations. SSCI 2021: 1-9*

# Unfair Recourse

The recourse actions necessary for transitioning to the positive class may exhibit greater variation in one group



Employ a regularized objective while training the classifier to ensure an equal average distance to the decision boundary for different groups

*Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, Bernhard Schölkopf: On the Fairness of Causal Algorithmic Recourse. AAAI 2022: 9584-9594*
*Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, Suresh Venkatasubramanian: Equalizing Recourse across Groups. CoRR abs/1909.03166 (2019)*

# Truthfulness of Fairness

Does *an explanation method preserve the fairness of the original model that it explains?*

Jessica Dai, Sohini Upadhyay, Stephen H. Bach, Himabindu Lakkaraju: What will it take to generate fairness-preserving explanations? CoRR abs/2106.13346 (2021)

# How explanations affect fairness judgments

An empirical study using COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), two racial groups (Caucasian and African- Americans

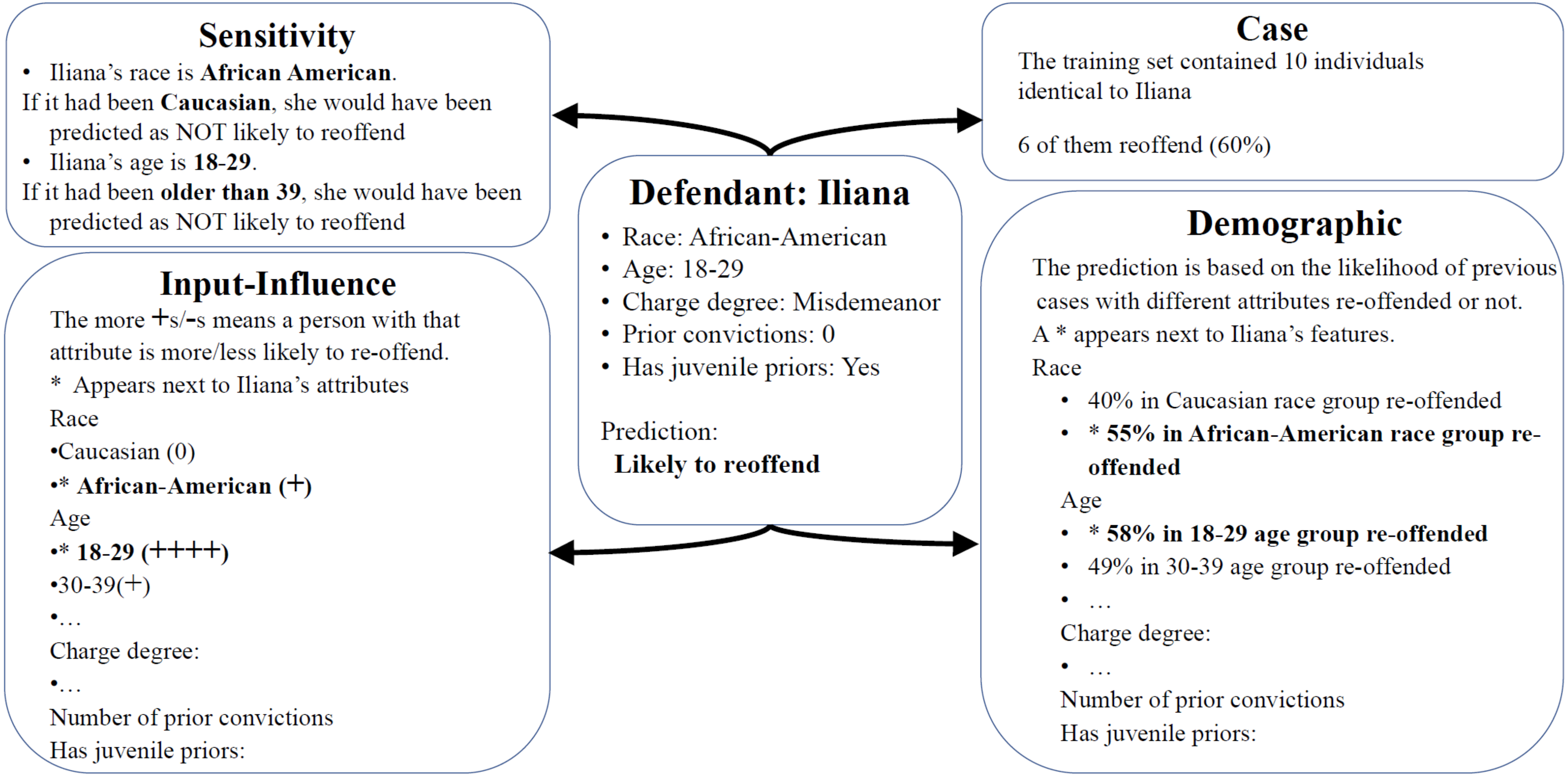
Compare fairness judgment for a model trained on raw data and on pre-processed data

Disparate impact: if two individuals with identical profile features but different racial categories receive different predictions, it should be considered unfair

*Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, Casey Dugan: Explaining models: an empirical study of how explanations impact fairness judgment. IUI 2019: 275-285*

Counterfactual (local)

Nearest-neighbor search (local)



**Sensitivity**

- Iliana's race is **African American**.

If it had been **Caucasian**, she would have been predicted as NOT likely to reoffend

- Iliana's age is **18-29**.

If it had been **older than 39**, she would have been predicted as NOT likely to reoffend

**Input-Influence**

The more +s/−s means a person with that attribute is more/less likely to re-offend.

* Appears next to Iliana's attributes

Race

•Caucasian (0)

•* **African-American (+)**

Age

•* **18-29 (++++)**

•30-39(+)

•…

Charge degree:

•…

Number of prior convictions

Has juvenile priors:

**Defendant: Iliana**

- Race: African-American
- Age: 18-29
- Charge degree: Misdemeanor
- Prior convictions: 0
- Has juvenile priors: Yes

Prediction:

**Likely to reoffend**

**Case**

The training set contained 10 individuals identical to Iliana

6 of them reoffend (60%)

**Demographic**

The prediction is based on the likelihood of previous cases with different attributes re-offended or not.

A * appears next to Iliana's features.

Race

- 40% in Caucasian race group re-offended
- * **55% in African-American race group re-offended**

Age

- * **58% in 18-29 age group re-offended**
- 49% in 30-39 age group re-offended
- …

Charge degree:

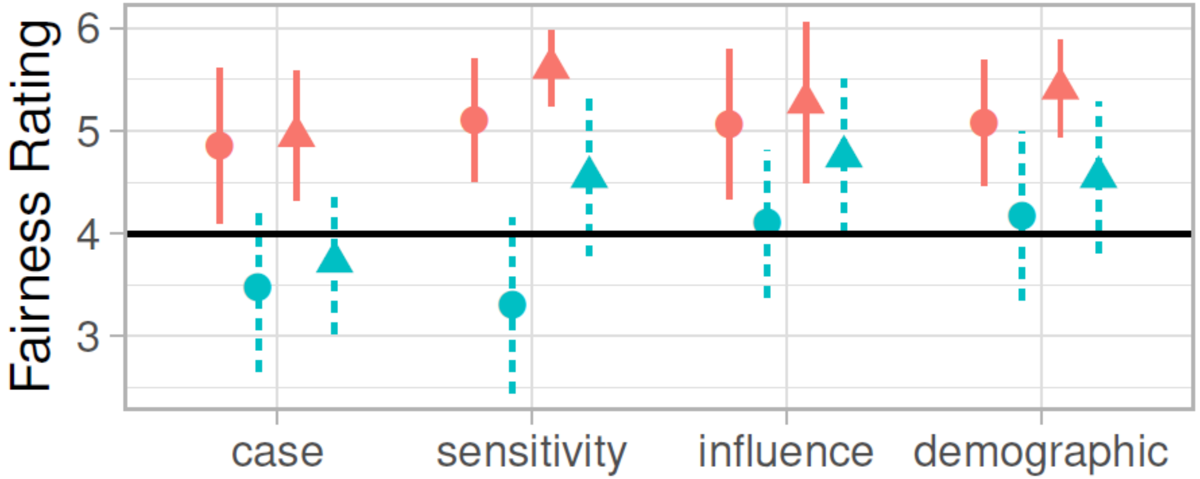
- …

Number of prior convictions

Has juvenile priors:

Use the relative importance of each feature (the logistic regression weights) global

how it is distributed with respect to the decision boundary (global)



● raw

▲ processed

Blue dashed: impacted (unfair)
Red solid: non impacted (fair)

- Predictions made on the processed data (triangles) were rated fairer than those on the raw data (circles).
- Predictions made on cases with disparate impact (blue dashed lines) were rated less fair than those without it (red solid lines).
- Explanation styles made some difference. Local (sensitivity-counterfactuals) and case (nearest-neighbor) more effective especially on raw data

# Fairness of Explanations (recap)

- Definitions
  - Quality of explanations
  - Difference in recourse
- Truthfulness
- How they affect fairness judgement

# More info

Check out our recent survey

- Christos Fragkathoulas, Vasiliki Papanikou, Danae Pla Karidi, Evaggelia Pitoura: *On Explaining Unfairness: An Overview.* CoRR abs/2402.10762 (2024) (also ICDE FAIR Workshop)

Work in progress, comments welcome

# Concluding remarks (in one slide)

Basic Background on Fairness and Explainability

Complex issues with many open questions (both in terms of formalism and in algorithmic terms)

Increasingly important with the increasing complexity of the models and  the prevailing use of foundational models

Any questions?