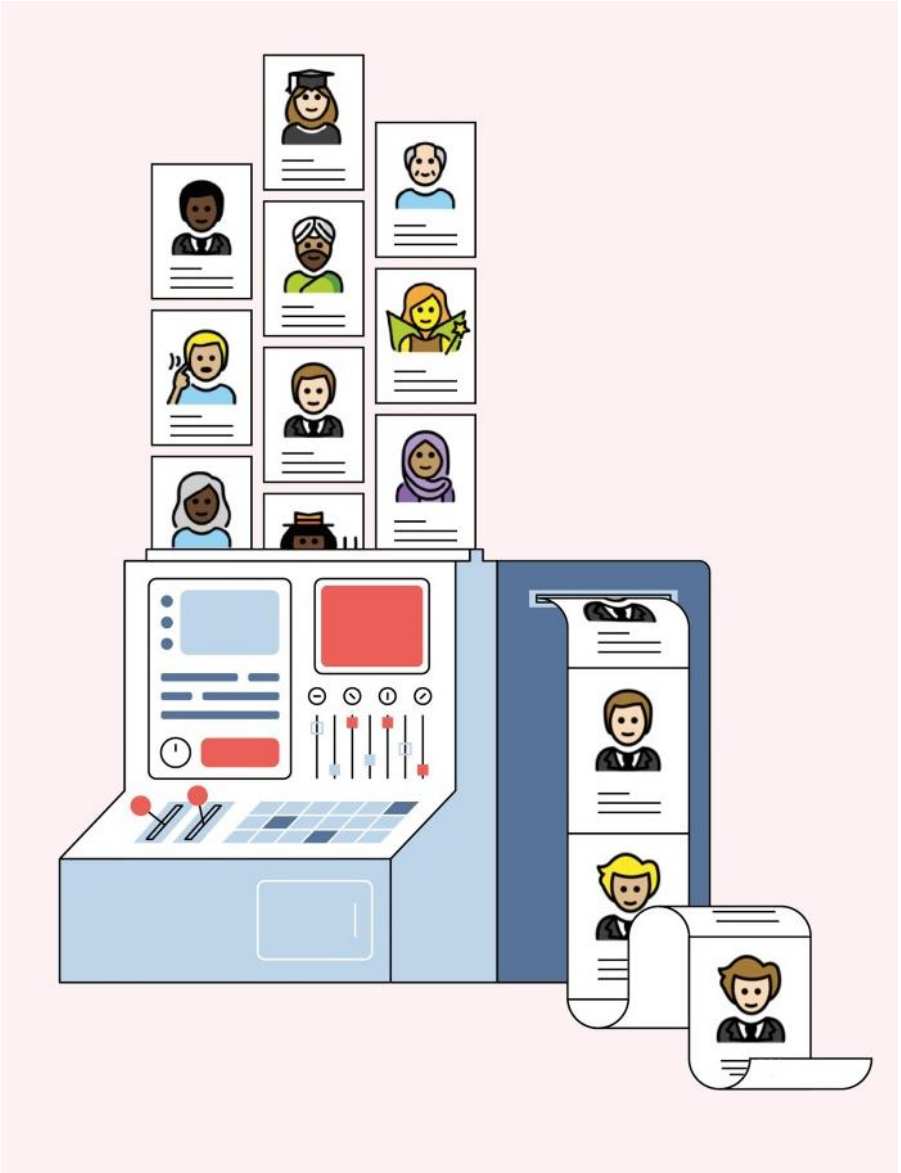
Unfair you say?
Explain yourself!

Toon Calders toon.calders@uantwerpen.be

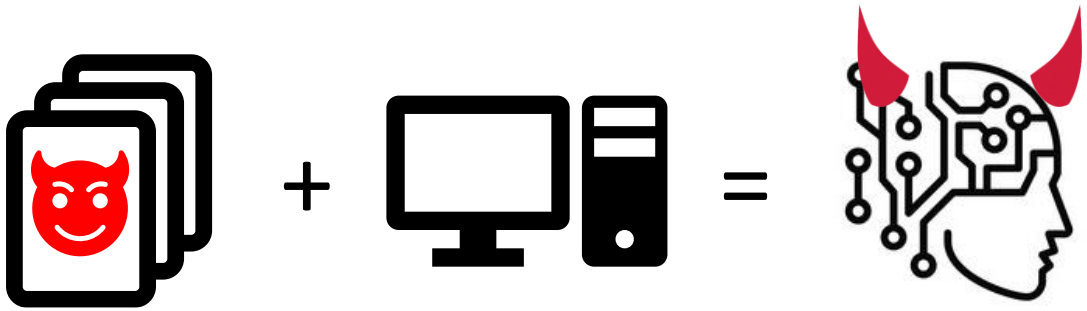
University of Antwerp



Data Bias

What happens when data do not represent reality or are biased?

Biased data lead to biased models



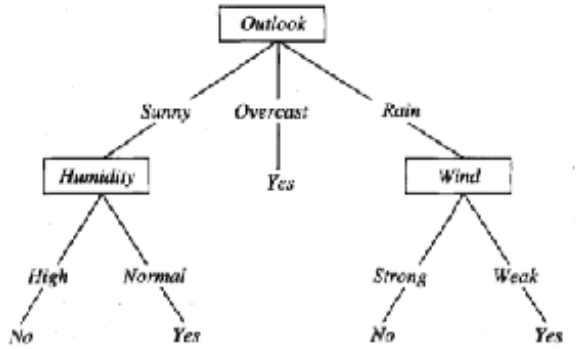
Racial bias found in health care risk algorithm

- Identify patients that will benefit from "high-risk care management" programs
- Black individuals: predicted to have 26.3 percent more chronic illnesses than white ones
- Race wasn't a variable, but **healthcare cost history**. […]
  - Black patients incurred lower health-care costs than white patients with the same conditions on average

https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

The “Old Way”: Fairness by design

Learning + fairness intervention;
e.g., reweighing

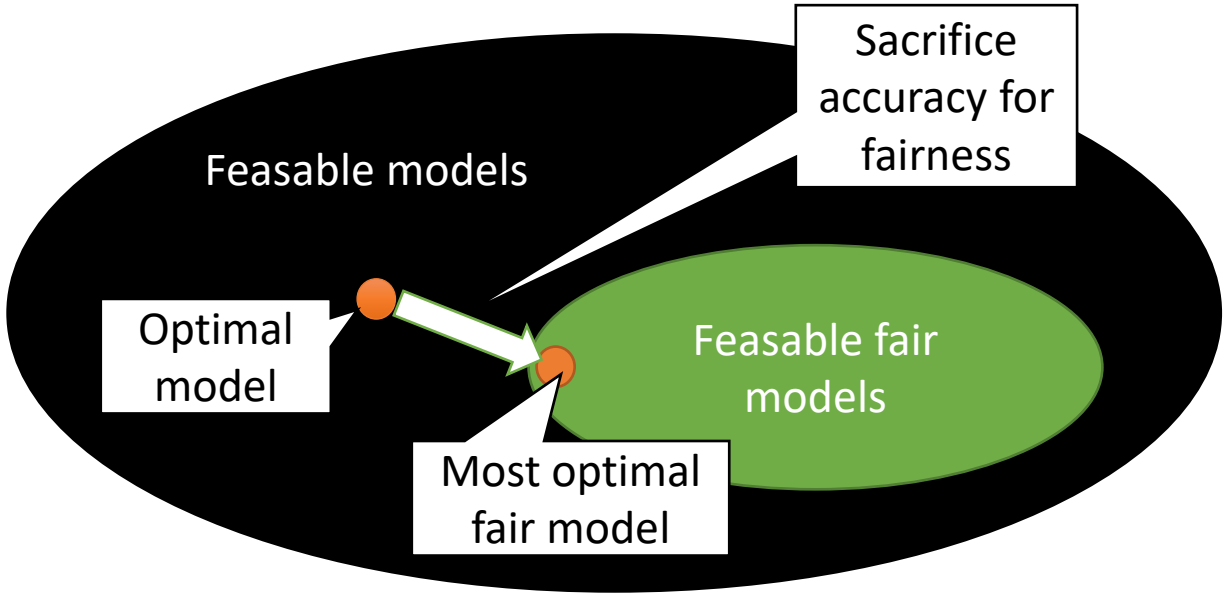


Bias detected by measure;
e.g., demographic parity (DP)

Model that is “*fair by design*”;
e.g., predictions satisfy DP

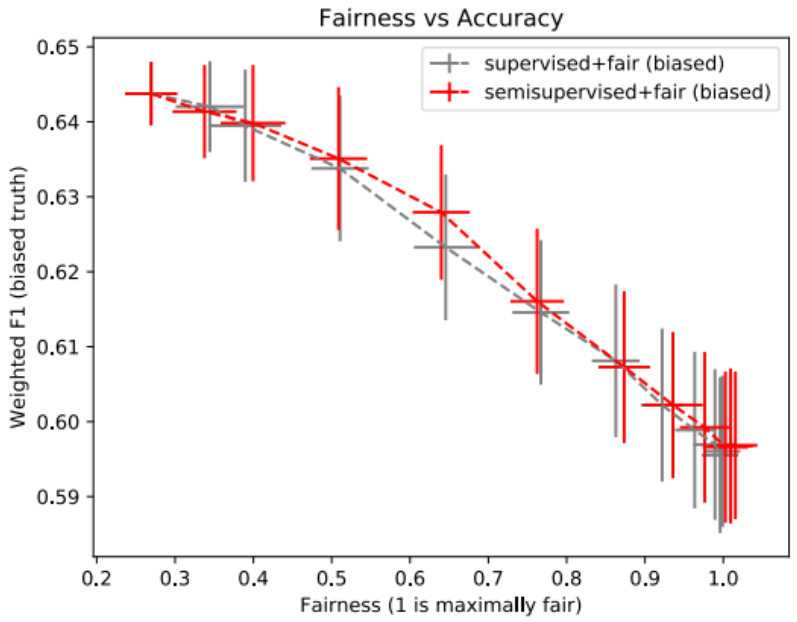
Common Approach to Fairness

- Select a fairness measure
- Optimize the fairness measure while keeping accuracy as high as possible



Accuracy – Fairness Trade-Off

- Common assumption: Most accurate *fair* model is less accurate than most accurate model overall • … however …



(a) COMPAS (biased ground truth)

Is this dataset biased?

Gender	X	Y
F	e	1
F	e	1
F	h	0
F	h	1
M	e	1
M	h	0
M	h	1
M	h	0
M	h	1
M	l	1
M	l	0

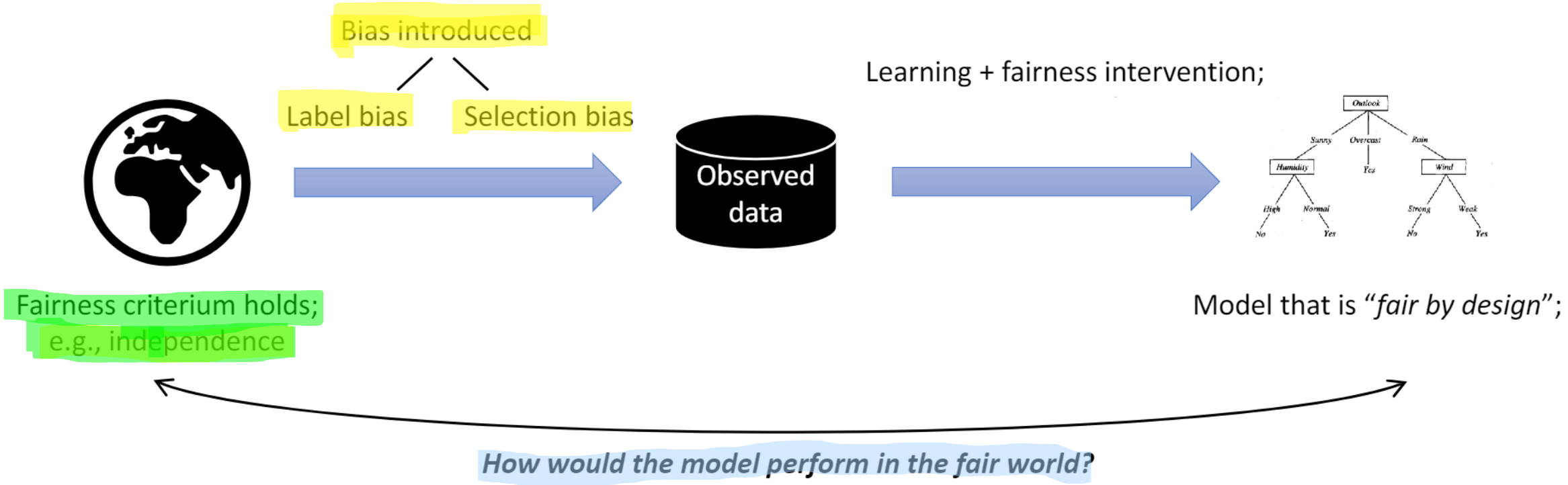
- Demographic parity difference = $\left|\frac{3}{4} - \frac{4}{7}\right| \approx 18\%$
- Conditional Demographic parity : $Y \perp\!\!\!\perp G \mid X$
- Females underrepresented
- Distribution of X differs between M and F

Preferred Debiassing Technique

Bias\Action	Ignore bias	Resample dataset before training	Affirmative action
No bias; X explains DPD	✓	F discriminated	M get preferential treatement
Selection bias; only qualified F apply	F get preferential treatment	✓	M get preferential treatement
M are discriminated	bias persists	bias persists	✓

- Preferred action when training depends on the type of bias
- Some fairness interventions will make things worse

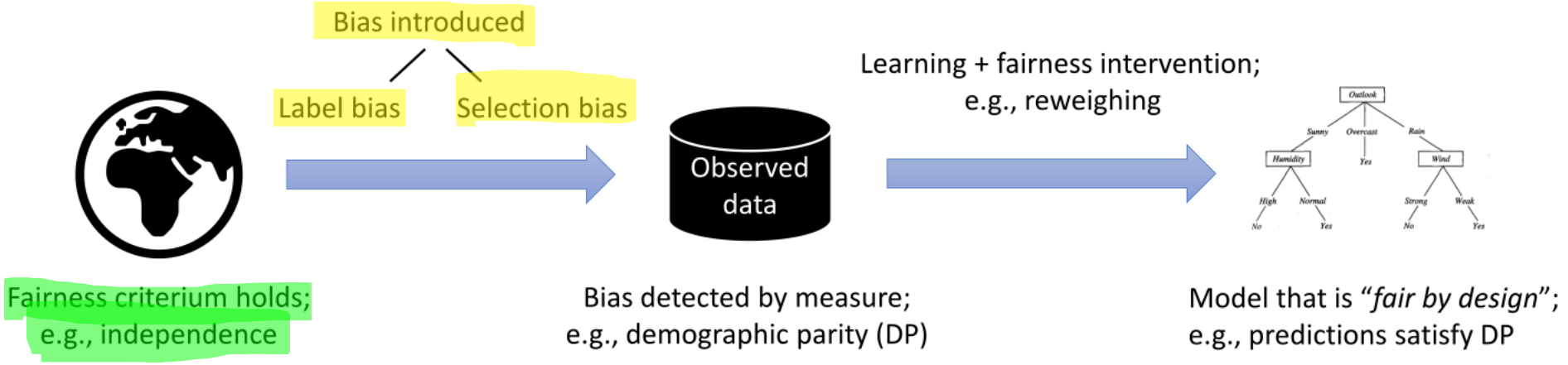
A New Way to be Fair



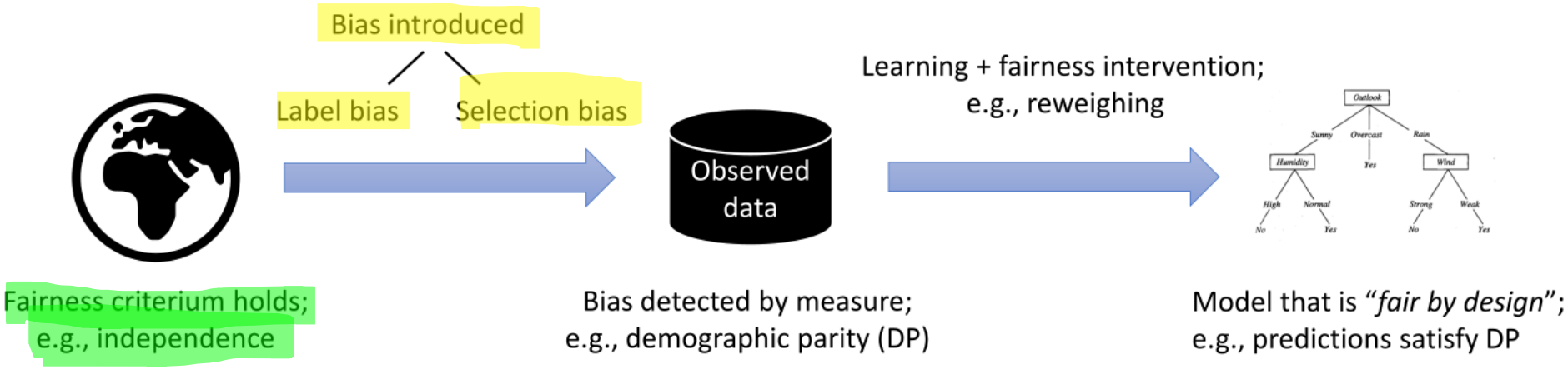
Favier, M., Calders, T., Pinxteren, S., & Meyer, J. (2023). How to be fair? A study of label and selection bias. Machine Learning, 1-24.

Fairness Framework

- Different types of bias introduction
- Different fairness assumptions

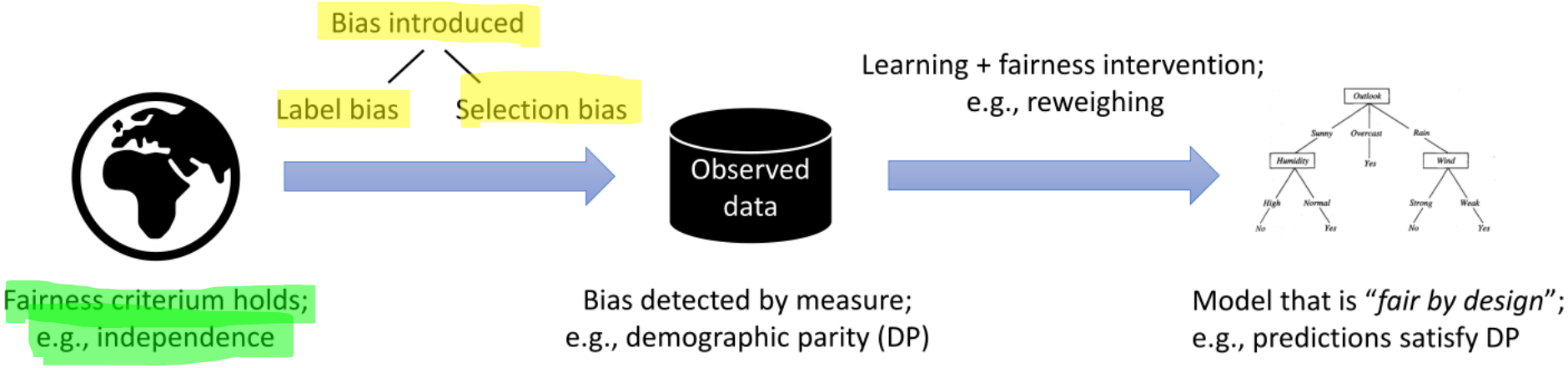


Research Questions



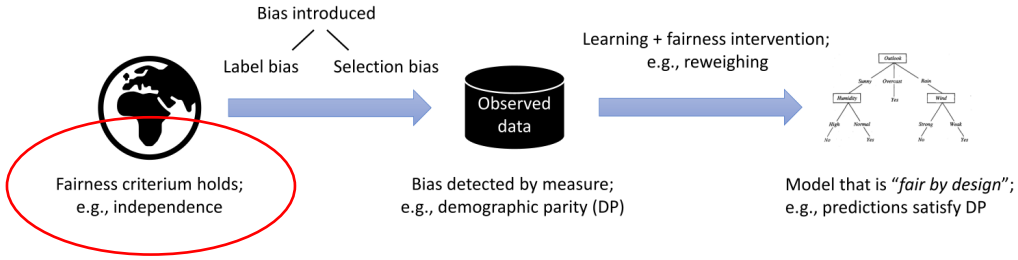
- How to formalize the sources of bias?
- Is the *observed data* consistent with the bias assumptions?
- What should we optimize w.r.t. the biased data?

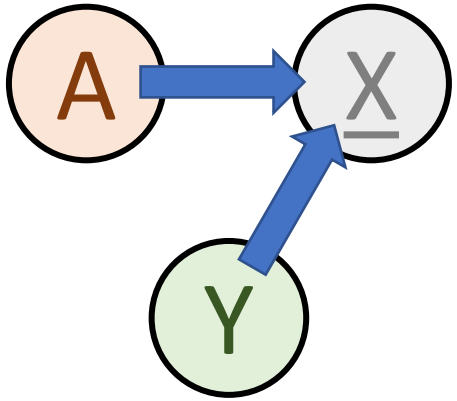
Research Questions



- **How to formalize the sources of bias?**
- Is the *observed data* consistent with the bias assumptions?
- What should we optimize w.r.t. the biased data?

Two definitions of Fair Data

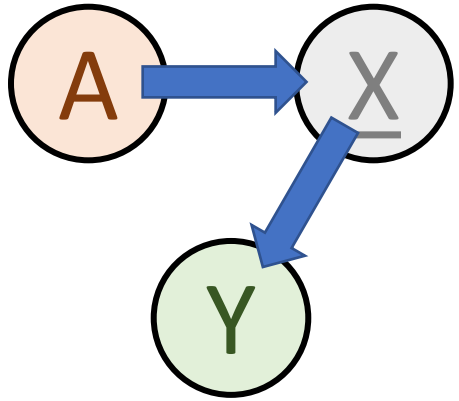




Statistical Parity

$$Y \perp\!\!\!\perp A$$

"The Label is equally distributed between sensitive groups"

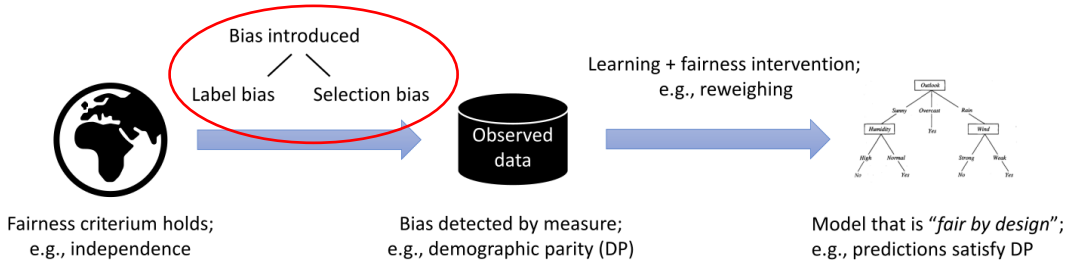


We're All Equal

$$Y \perp\!\!\!\perp A \mid X$$

"Identical individuals with different sensitive attributes are treated equally"

Label Bias

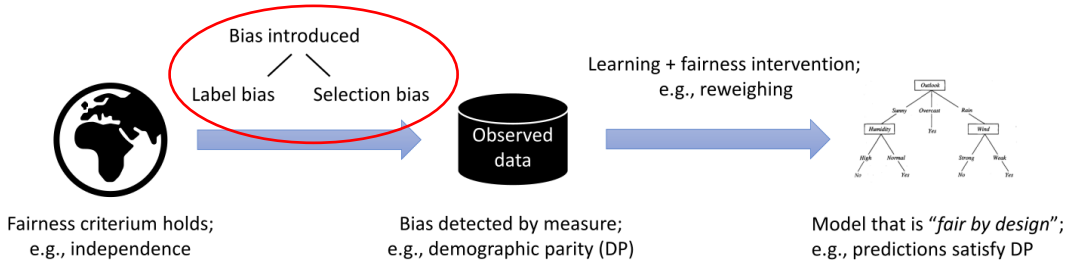


The label that some individuals received does not represent the label they deserved.

Example: Job Hiring
If a racist person is responsible for hiring people, he will deny the positive label to valuable workers from the discriminated group. Those people did not receive the label they deserved.

X_1	...	X_n	A	Y
0	...	1	♂	✓
1	...	1	♀	✗
0	...	0	♂	✓
0	...	1	♀	✓
1	...	0	♂	✗
1	...	0	♂	✓
0	...	0	♀	✗
1	...	1	♀	✗

Selection Bias

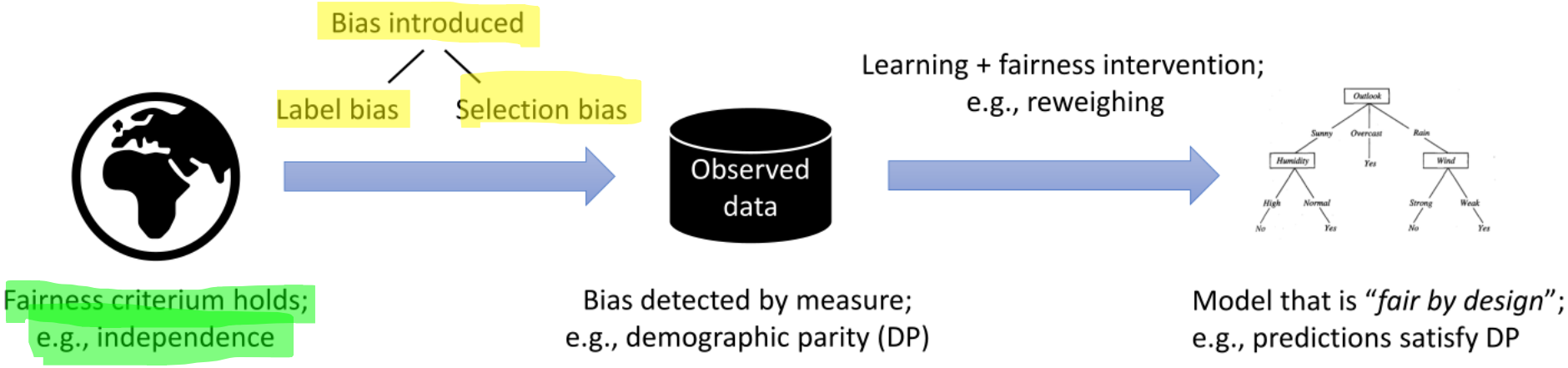


the selection of the individuals in the dataset is not independent from the individuals' features.

Example: the *toeslagenaffaire*
The data collected relied on anonymous tips, which may lead discriminated groups to be over-represented.

X_1	...	X_n	A	Y
0	...	1	♂	✓
1	...	1	♀	✗
0	...	0	♂	✓
0	...	1	♀	✓
1	...	0	♂	✗
1	...	0	♂	✓
0	...	0	♀	✗
1	...	1	♀	✗

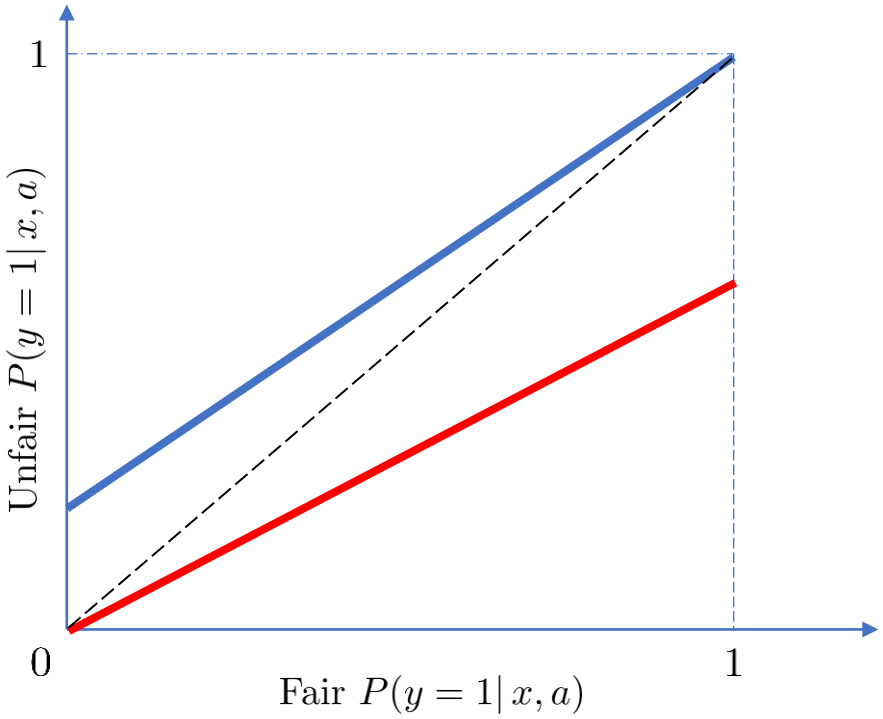
Research Questions



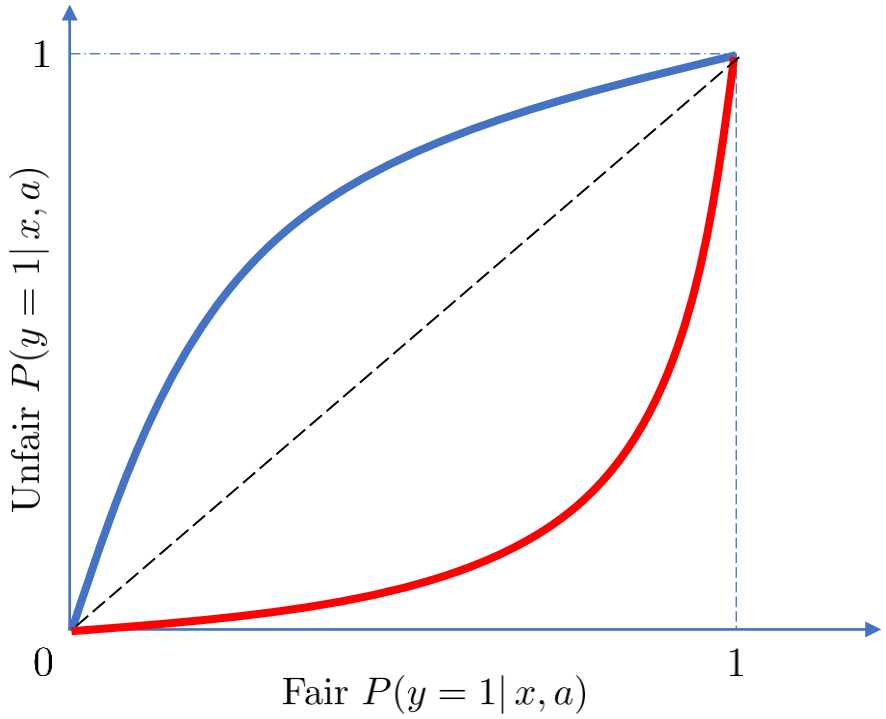
- How to formalize the sources of bias?
- **Is the *observed data* consistent with the bias assumptions?**
- What should we optimize w.r.t. the biased data?

How Bias Changes Probabilities

Unprivileged Group



Label Bias



Selection Bias

Results: Observability

- Positive & Negative results:
  - Some combinations can be refuted from data only, others cannot

	Statistical Parity	We're all equal
Label Bias	• Can be detected? ✓	• Can be detected? ✓
Sampling Bias	• Can be detected? ✗	• Can be detected? ✓

Favier, M., Calders, T., Pinxteren, S., & Meyer, J. (2023). How to be fair? A study of label and selection bias. Machine Learning, 1-24.

Negative Results

- Some combinations cannot be verified from data
  - Sample bias and Statistical Parity

Gender	X	Y
F	e	1
F	h	0
F	h	1
F	h	0
M	e	1
M	h	0
M	h	1
M	l	0

Gender	X	Y
F	e	1
F	h	0
F	h	1
M	e	1
M	h	0
M	l	0

Statistical Parity + Label Bias

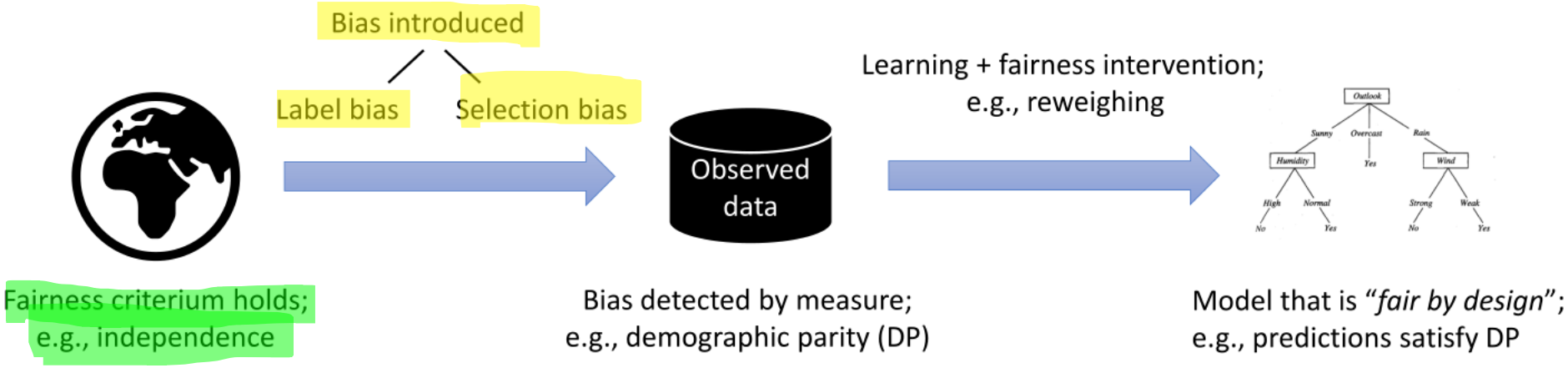
- What does it mean?
  - Necessary condition that allows us to check if label bias has occurred.
  - The amount of bias is tied with how difficult it is to make predictions.
  - In general, the process that generates the biased data might be not unique.

Theorem:

The following set must be non-empty:

$$\left[1 - \frac{P_U(y_0 \mid a_1)}{\max_{x \in X} P_U(y_0 \mid x, a_1)}, \frac{P_U(y_1 \mid a_1)}{\max_{x \in X} P_U(y_1 \mid x, a_1)}\right] \cap \left[1 - \frac{P_U(y_0 \mid a_0)}{\max_{x \in X} P_U(y_0 \mid x, a_0)}, \frac{P_U(y_1 \mid a_0)}{\max_{x \in X} P_U(y_1 \mid x, a_0)}\right]$$

Research Questions



- How to formalize the sources of bias?
- Is the *observed data* consistent with the bias assumptions?
- **What should we optimize w.r.t. the biased data?**

Logical Error

Fairness criterium holds;
e.g., independence

Assumption:
gender independent from label

Bias introduced

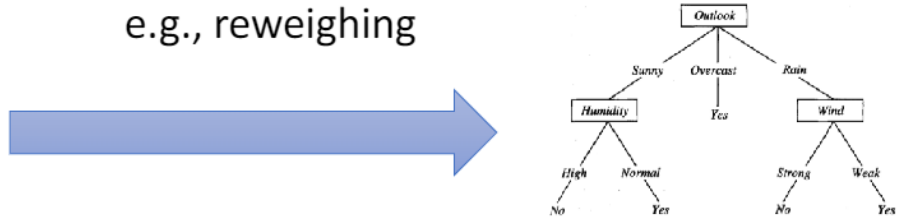
Label bias Selection bias

Bias detected by measure;
e.g., demographic parity (DP)

Observation:
gender dependent from label

Require:
gender independent from prediction

Learning + fairness intervention;
e.g., reweighing



Model that is "*fair by design*";
e.g., predictions satisfy DP

Logical Error

Gender	X	Y
F	e	1
F	h	0
F	h	1
F	h	0
M	e	1
M	h	0
M	h	1
M	l	0

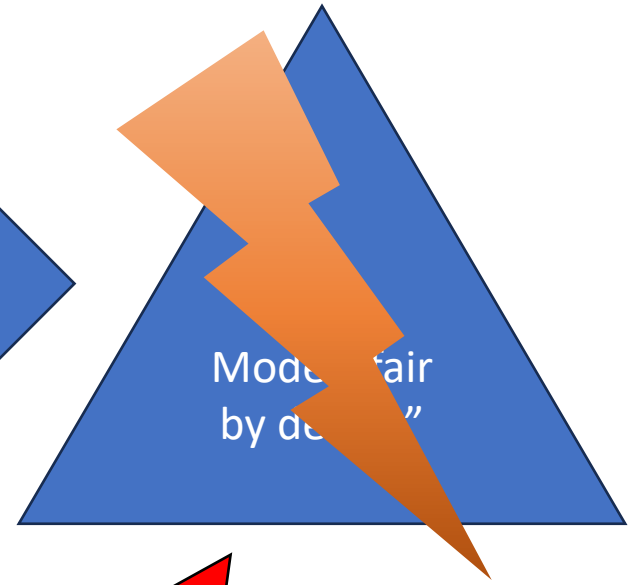
Assumption:
gender independent from label

Self-selection bias

Gender	X	Y
F	e	1
F	h	0
F	h	1
M	e	1
M	h	0
M	l	0

Observation:
gender dependent from label

Use constraint DPD = 0



Discriminates against women

Results

- Again, depends on situation

	Statistical Parity	We're all equal
Label Bias	• Still satisfied? ✓	• Still satisfied? ✓
Sampling Bias	• Still satisfied? ✗	• Still satisfied? ✓

Favier, M., Calders, T., Pinxteren, S., & Meyer, J. (2023). How to be fair? A study of label and selection bias. Machine Learning, 1-24.

Intermediate Conclusion

- Fully technological solution is impossible
  - Unable to unambiguously detect biases
- Neccesity for human involvement
  - Assumptions on “ideal world”
  - Assumptions on bias process
- Needs?
  - Language to express assumptions → causal networks ?
  - Methods to verify consistency of assumptions and data

Fairness as an Optimization Problem

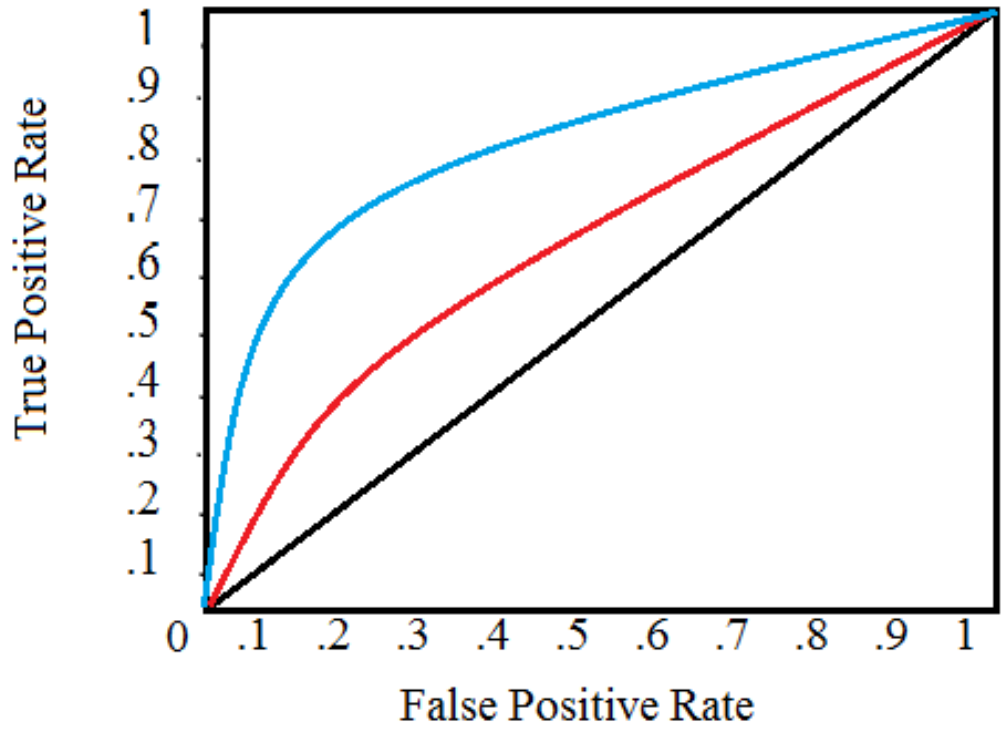
- Suppose:
  - All assumptions are right
  - Optimal model satisfies fairness criterion on data

- Still not out of the woods!
  - Optimizing fairness criterion has *undesirable side-effects*
  - E.g.: “Fair classifier” making mistakes on purpose

Favier, Marco, and Toon Calders. "Cherry on the Cake: Fairness is NOT an Optimization Problem." *arXiv preprint arXiv:2406.16606* (2024).

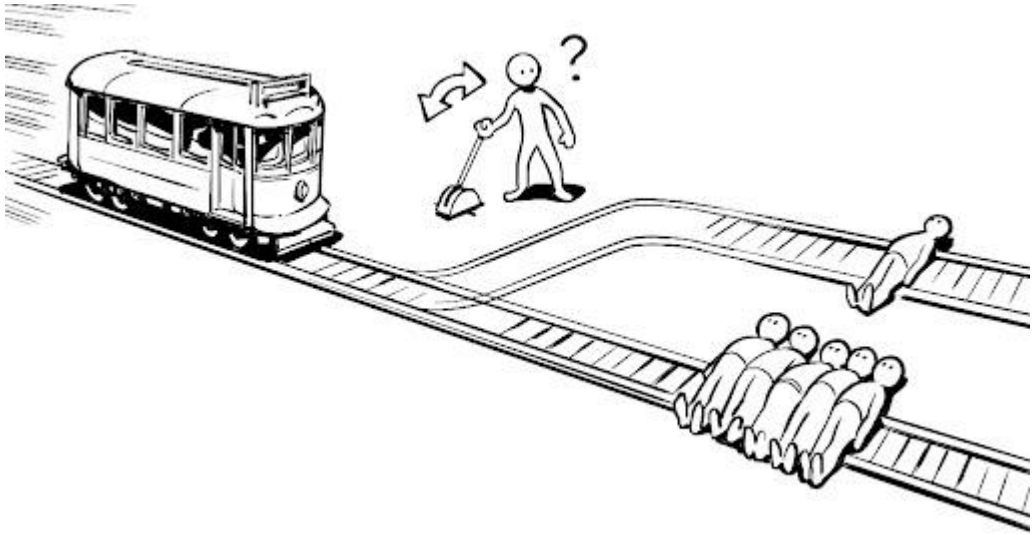
Example: Optimizing for Equal Odds

- Suppose:
  - Sensitive attribute can be used for classification
- Draw AUC curves for both sensitive groups
  - What if they do not intersect?
- Model has to make mistakes *on purpose*

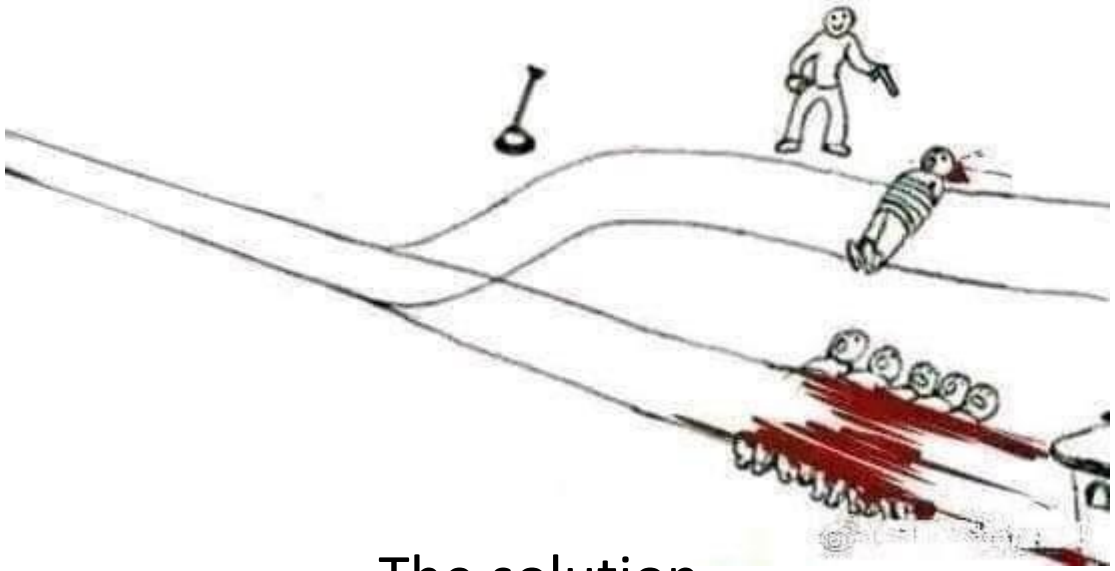


Optimal Equal Odds Classifier

- Provable optimal classifier makes mistakes on purpose



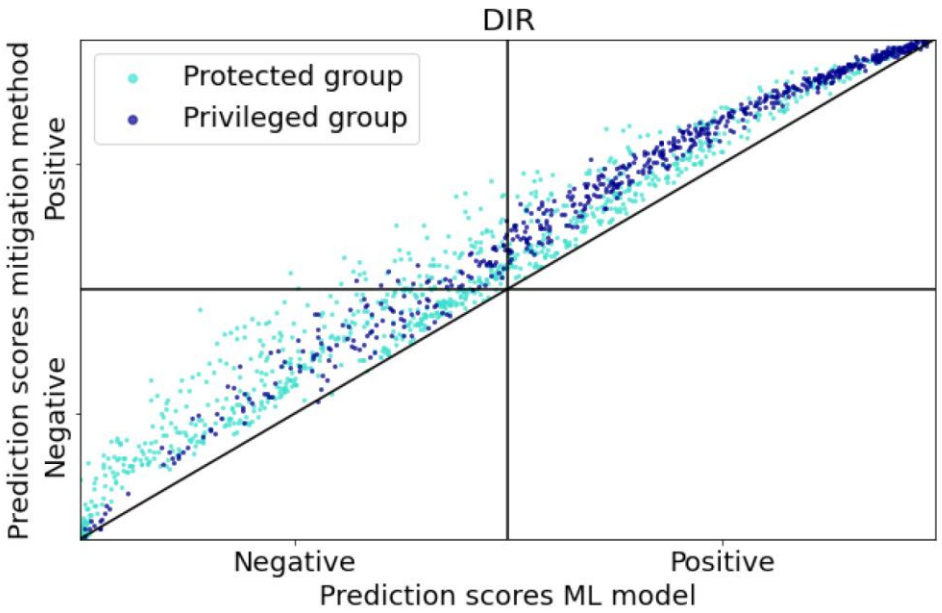
The dilemma



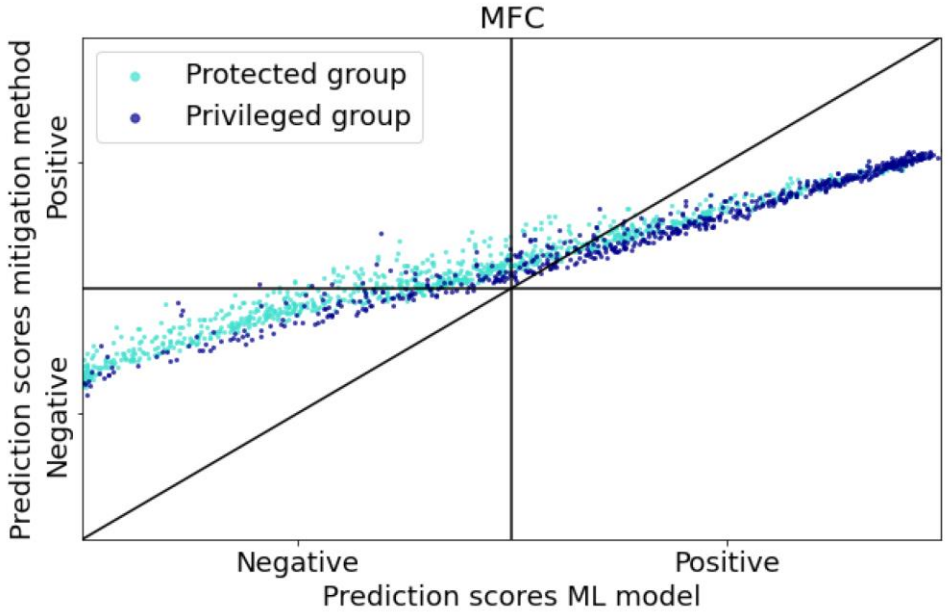
The solution

Different “Optimal” Solutions (COMPAS dataset)

Disparate Impact Remover (DIR)



Meta Fair Classifier (MFC).



Goethals, Sofie, Toon Calders, and David Martens. "Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics." *arXiv preprint arXiv:2401.13391* (2024).

Intermediate Conclusion 2

- Fully technological solution is impossible
  - Unable to unambiguously detect biases
  - Human input required
- But, even if all assumptions made and right, optimization has undesirable side-effects
  - Errors on purpose
  - Wildely different approaches scoring equally well
- Need for explanations / model interpretation

The Need for Explanations

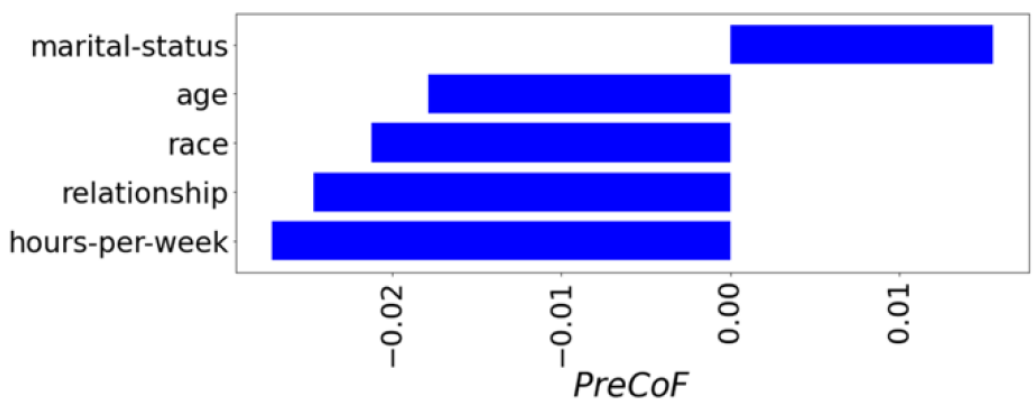
- Very simple idea: PreCoF
  - Generate counterfactuals for all instances
  - Which attributes most often changed?

$$CoF(j,s) = \frac{\left|\left\{i | \exists \mathbf{c} : x_{iz} = s, M\left(\mathbf{x}'_i\right) = -, \mathbf{c} \text{ is a counterfactual of } \mathbf{x}'_i\right\}\right|}{\left|\left\{i | x_{iz} = s, M\left(\mathbf{x}'_i\right) = -\right\}\right|}$$

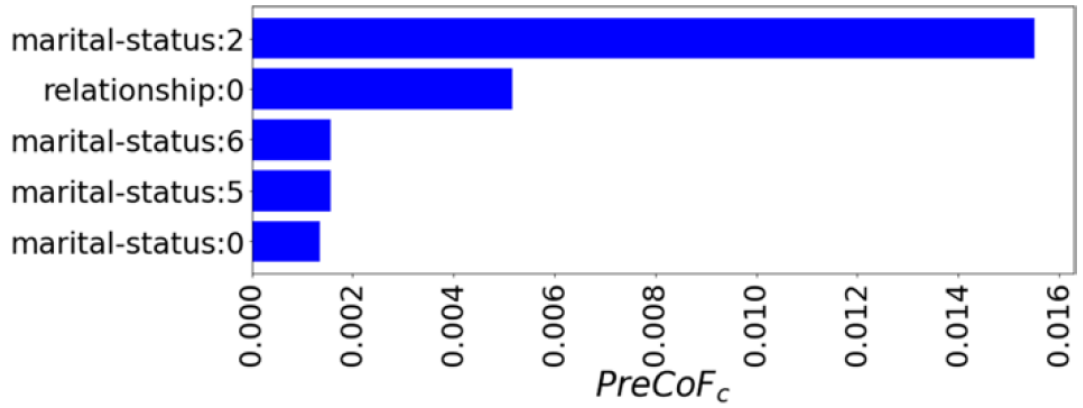
$$PreCoF(j) = CoF(j,s) - CoF(j,ns)$$

Goethals, Sofie, David Martens, and Toon Calders. "PreCoF: counterfactual explanations for fairness." *Machine Learning* 113.5 (2024): 3111-3142.

Adult Dataset - PreCoF

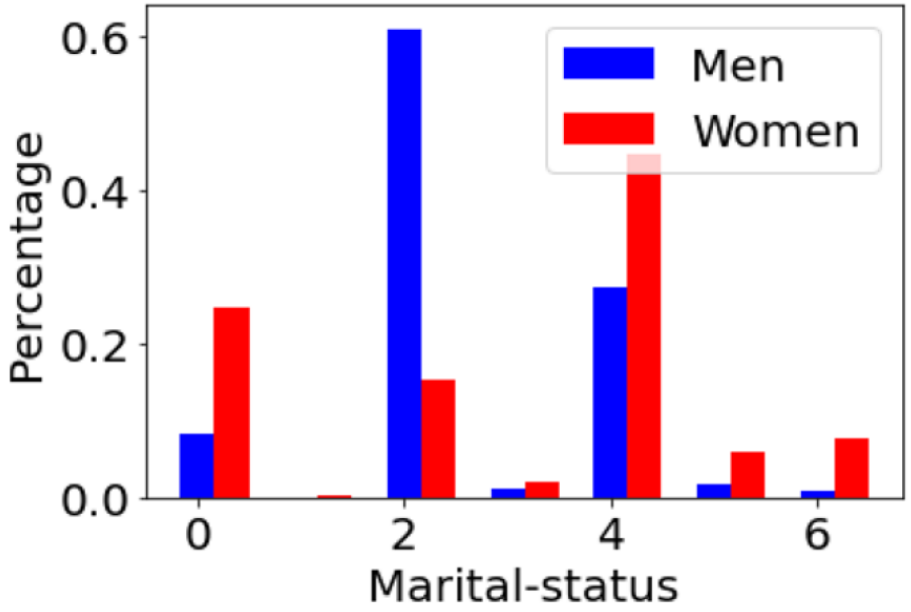


(a) $PreCoF$: attributes in the counterfactual explanations

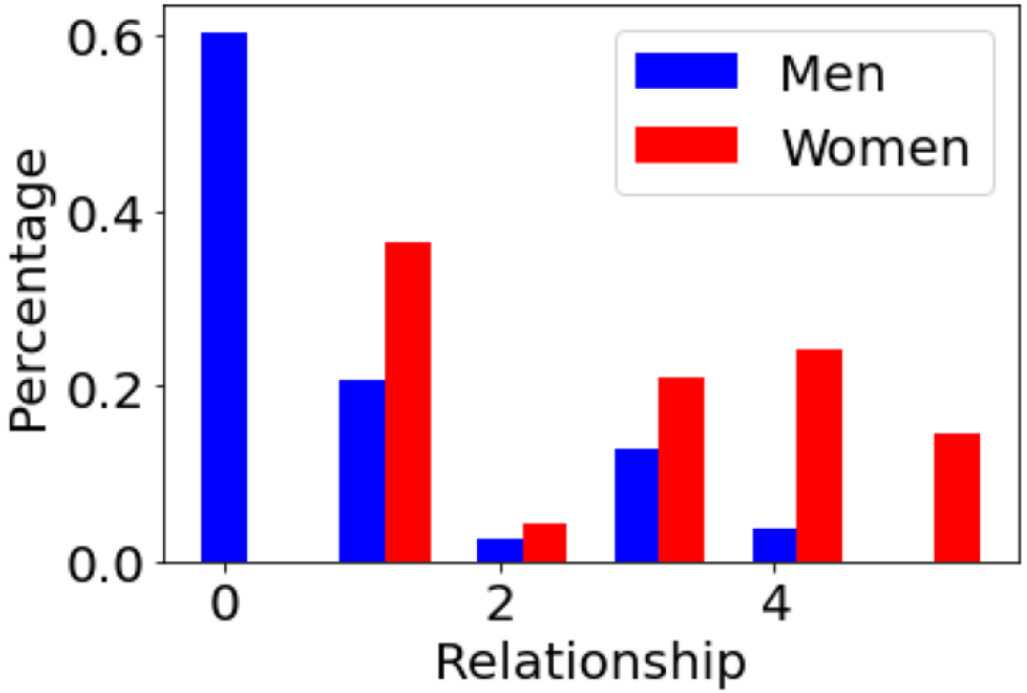


(b) $PreCoF_c$: attribute values in the counterfactual explanations

Proxies for Gender



(a) 0 = Divorced, 1 = Married to AF, 2= Married to Civ. Spouse, 3= Married to Abs. Spouse, 4 = Never married, 5 = Separated, 6 = Widowed



(b) 0 = Husband, 1 = Not in a family, 2 = Other relatives, 3 = Own children, 4= Unmarried, 5 = Wife

Student performance dataset

- Model trained with sensitive attribute

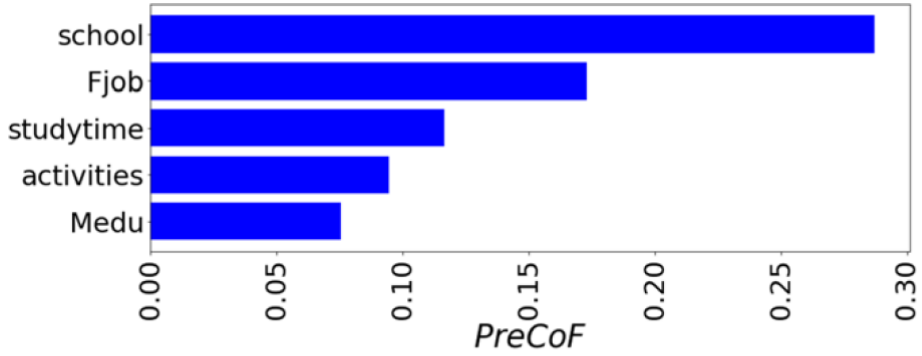
Table 5 Accuracy and fairness metrics for the model trained on the Student Performance Dataset

	Situation 2 Model without sensitive attribute
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.115 (0.646–0.531)**
Accuracy of the model	*71.28%*

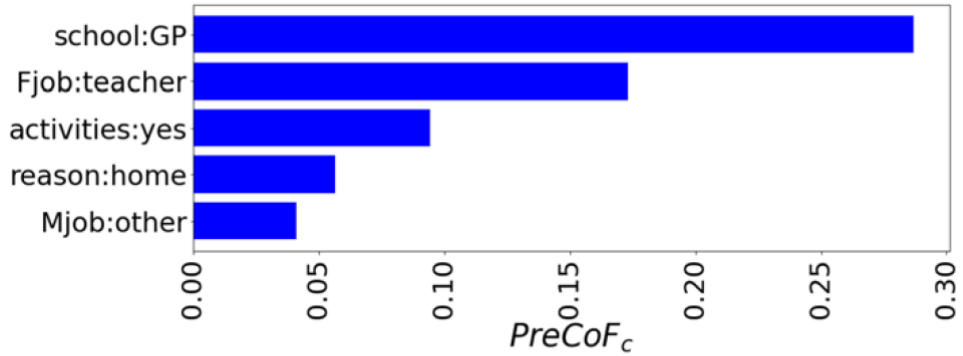
Females discriminated against

However …

- Direct discrimination:
  - 3 times *If you would have been a girl, you would have been predicted as scoring above average instead of below*
  - reverse does not happen
- How come women *indirectly* disadvantaged?



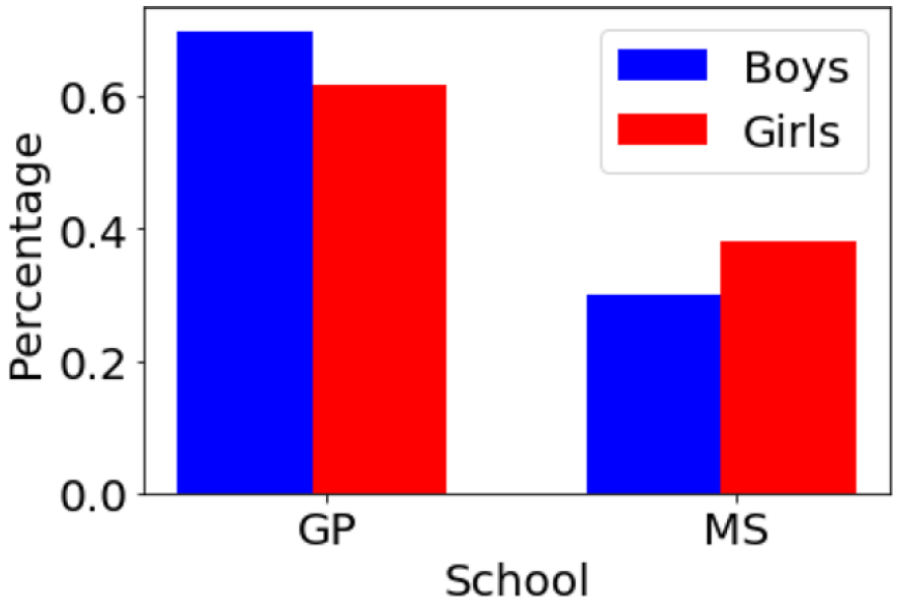
(a) $PreCoF$: attributes in the counterfactual explanations

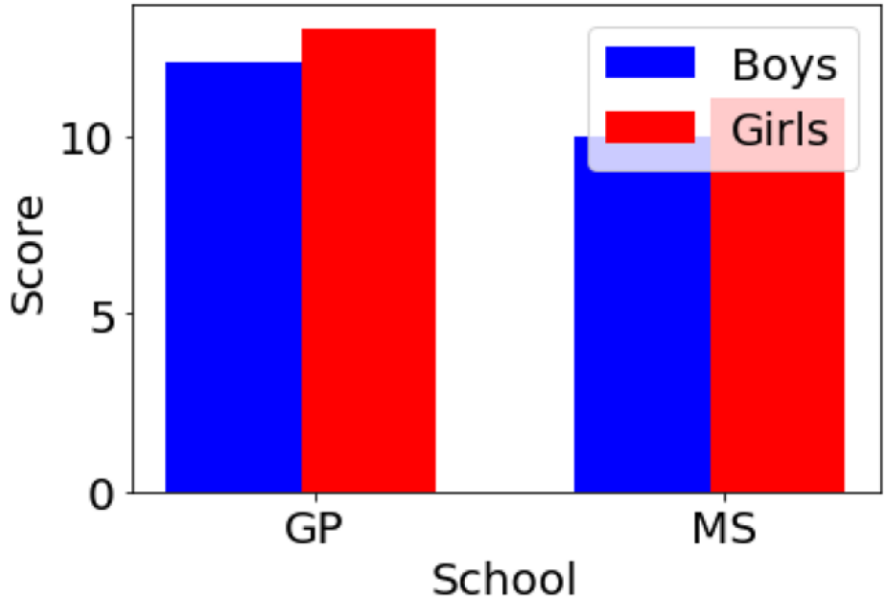


(b) $PreCoF_c$: attribute values in the counterfactual explanations

Explanation …

- Indirect effect of school:





- Girls outperform boys in both schools, but more boys go to the "better school"

Student performance dataset

Table 5 Accuracy and fairness metrics for the model trained on the Student Performance Dataset

	Situation 1 Model with sensitive attribute	Situation 2 Model without sensitive attribute	Situation 3 Model without sensitive attribute and $PreCoF_1$
Demographic disparity (Positive rate unprotected group - positive rate protected group)	**0.043 (0.610−0.566)**	**0.115 (0.646−0.531)**	**0.066 (0.659−0.593)**
Accuracy of the model	*73.85%*	*71.28%*	*70.26%*

Conclusion

- Fully technological solution to fairness: **impossible!**
- Need for human input:
  - Fairness assumptions
  - Bias assumptions
- Blindly optimizing fairness as a constraint: **undesirable!**
  - Errors on purpose
  - Different ways to reach same effect

- **Explanations for understanding fairness of models**

Many thanks to my collaborators

Marco Favier

Sofie Goethals

David Martens


Ethical machine learning is maximizing accuracy in a fair world