When Mitigating Bias is Unfair

Studying the Impact of Bias Mitigation Algorithms

Thibault Laugel – thibault.laugel@axa.com

About this work

TRAIL
Trustworthy and Responsible AI Lab

SORBONNE UNIVERSITÉ

When Mitigating Bias is Unfair: A Comprehensive Study on the Impact of Bias Mitigation Algorithms

Natasa Krco[1*] Thibault Laugel[1*] Jean-Michel Loubes[2]
Marcin Detyniecki[1,3,4]

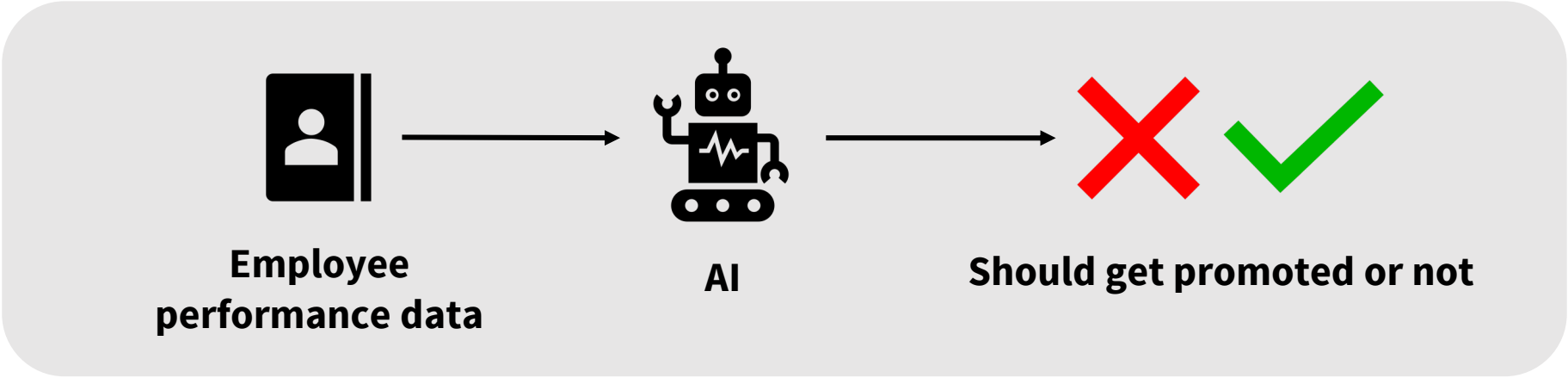



INSTITUT DE MATHÉMATIQUES DE TOULOUSE

Preprint link: https://arxiv.org/abs/2302.07185v1

Context: Algorithmic Fairness

Example: HR case in a company



Observation: certain groups are ”**privileged**” = more likely to be put in the positive class

Technical solution: design models that decrease bias, but preserve accuracy

Context: Algorithmic Fairness

Enforcing Fairness in Practice

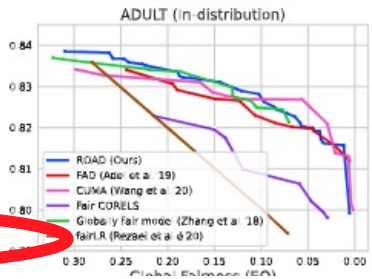
Old model

Choose a fairness objective

Apply bias mitigation algorithms

Select a Fair Model

	accuracy	fairness
method 1	...	...
method 2	...	...
method 3	...	...
method 4	...	...



Context: Algorithmic Fairness

Enforcing Fairness in Practice

Old model

Choose a

Apply bias

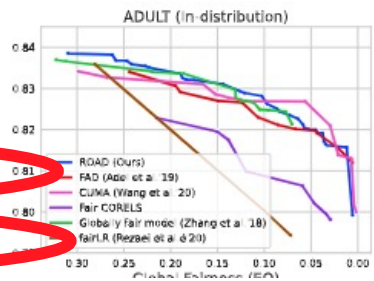
Select a Fair

el

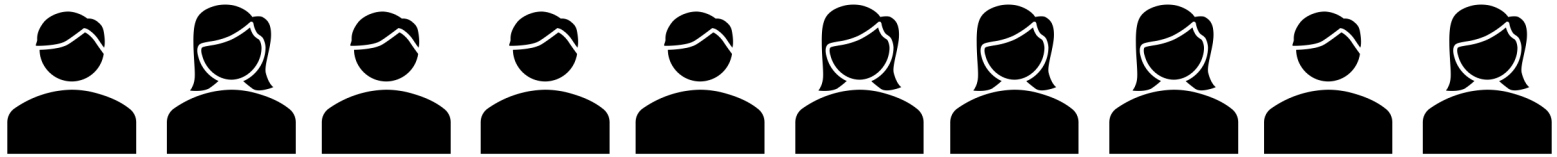
Issues/points of interest:

1. Debiasing is not a deterministic process
2. This process is not transparent

	accuracy	fairness
method 1	...	...
method 2	...	...
method 3	...	...
method 4	...	...



Traceability in the bias mitigation process



Promoted

Not promoted

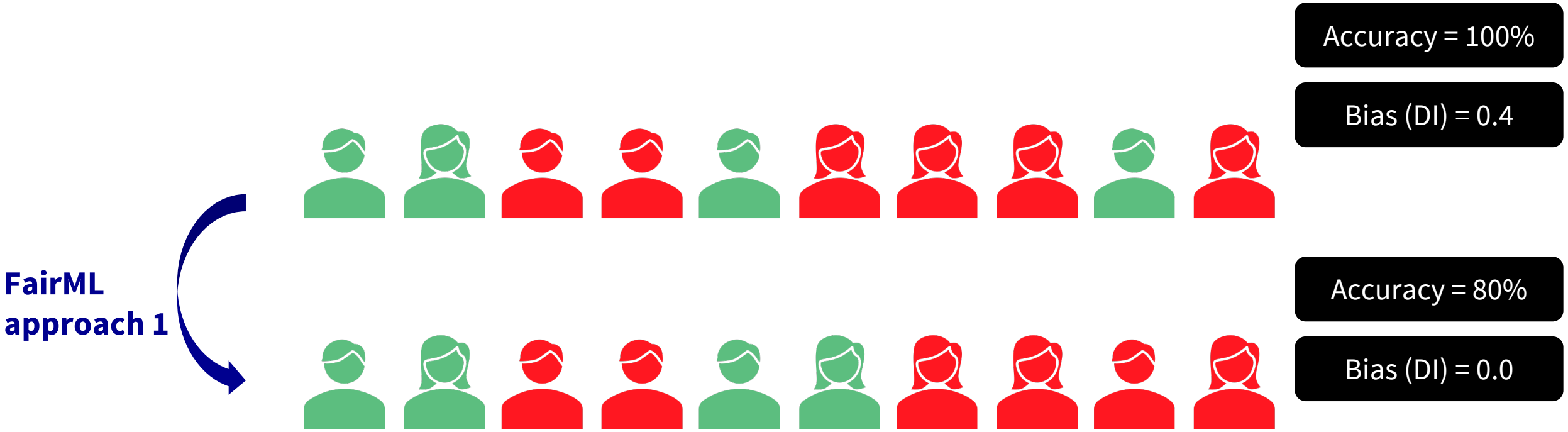
Traceability in the bias mitigation process

Traceability in the bias mitigation process

Accuracy = 100%

Bias (DI) = 0.4

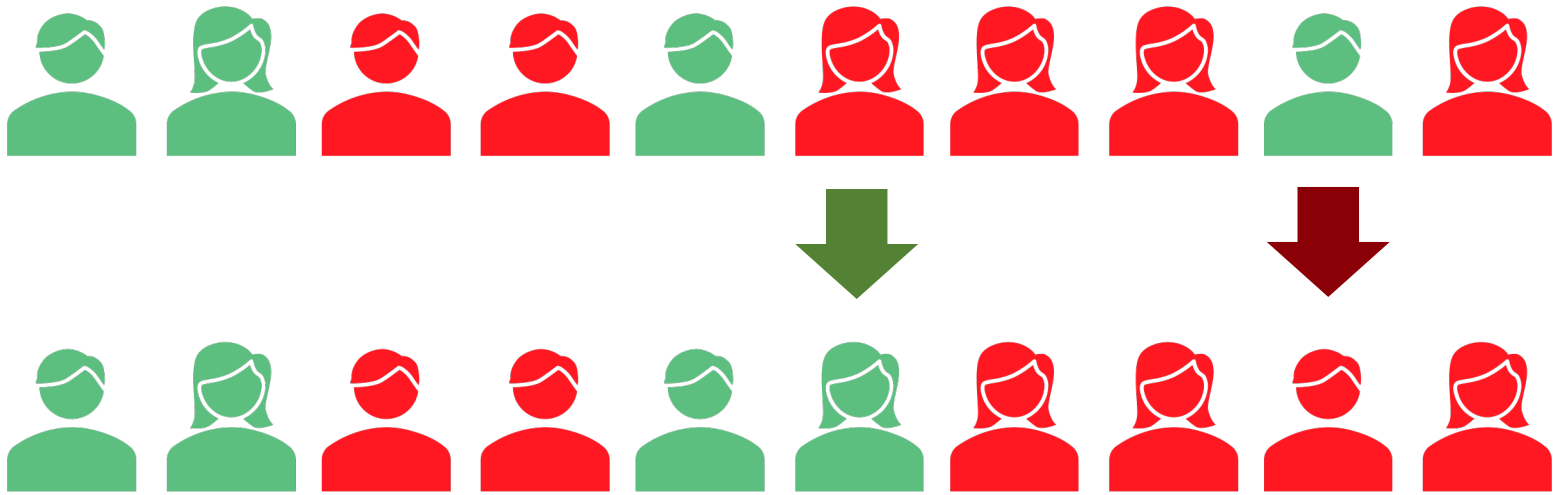
Traceability in the bias mitigation process



Traceability in the bias mitigation process

Accuracy = 100%

Bias (DI) = 0.4



FairML approach 1

Accuracy = 80%

Bias (DI) = 0.0

2 changes:

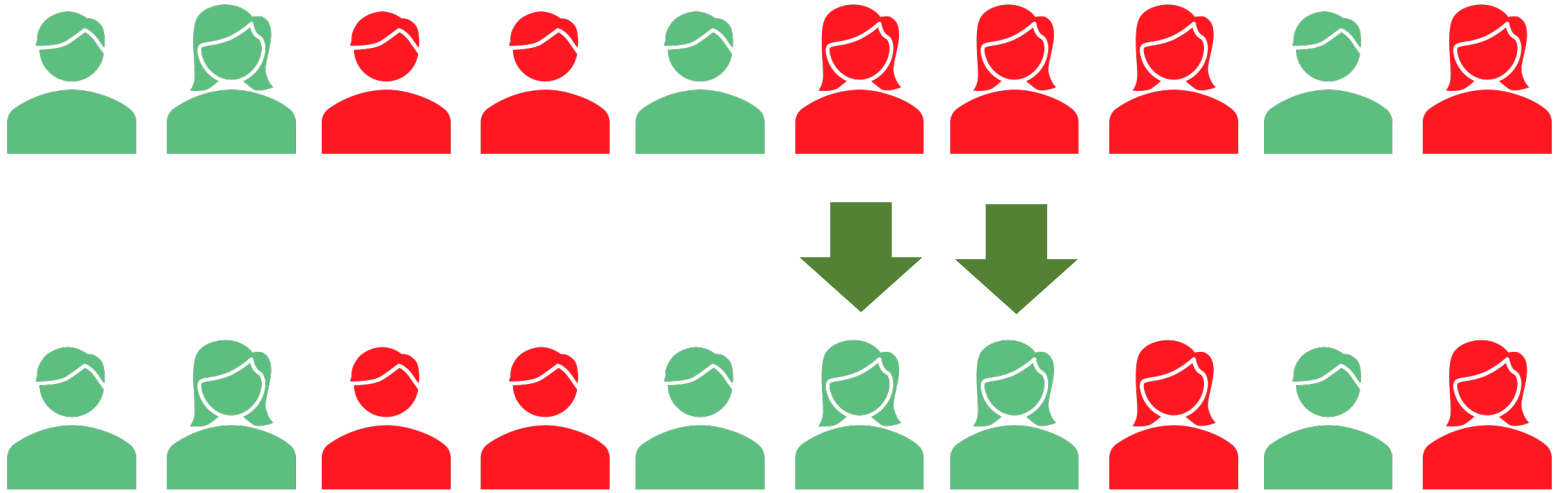
Promoting women more and men less

Traceability in the bias mitigation process

Accyracy, fairness

Accuracy = 100%

Bias (DI) = 0.4



FairML approach 2

Accuracy = 80%

Bias (DI) = 0.0

2 changes:

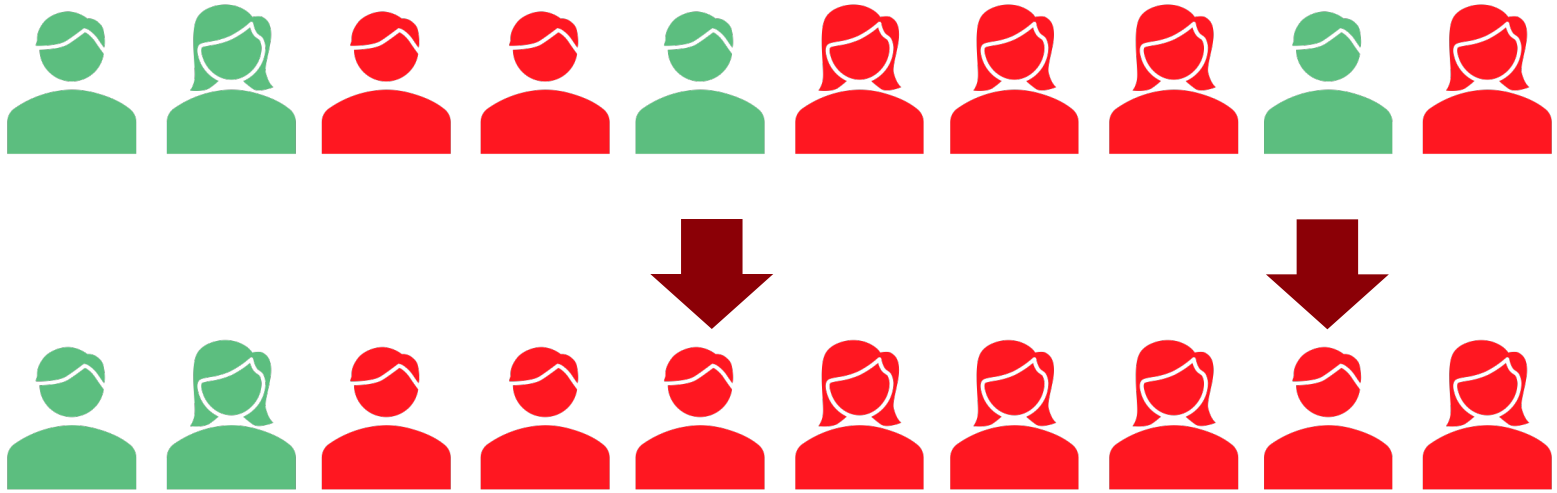
Promoting more women

Traceability in the bias mitigation process

Accyracy, fairness

Accuracy = 100%

Bias (DI) = 0.4



FairML approach 3

Accuracy = 80%

Bias (DI) = 0.0

2 changes:

Promoting less men

Multiplicity in Debiasing

2 changes: 

Promoting women more and men less

2 changes: 

Promoting more women

2 changes: 

Promoting less men

$\neq$

Multiplicity in Debiasing

Why it's bad

If we are not looking, algorithmic fairness methods then fail to achieve their goal of true fairness

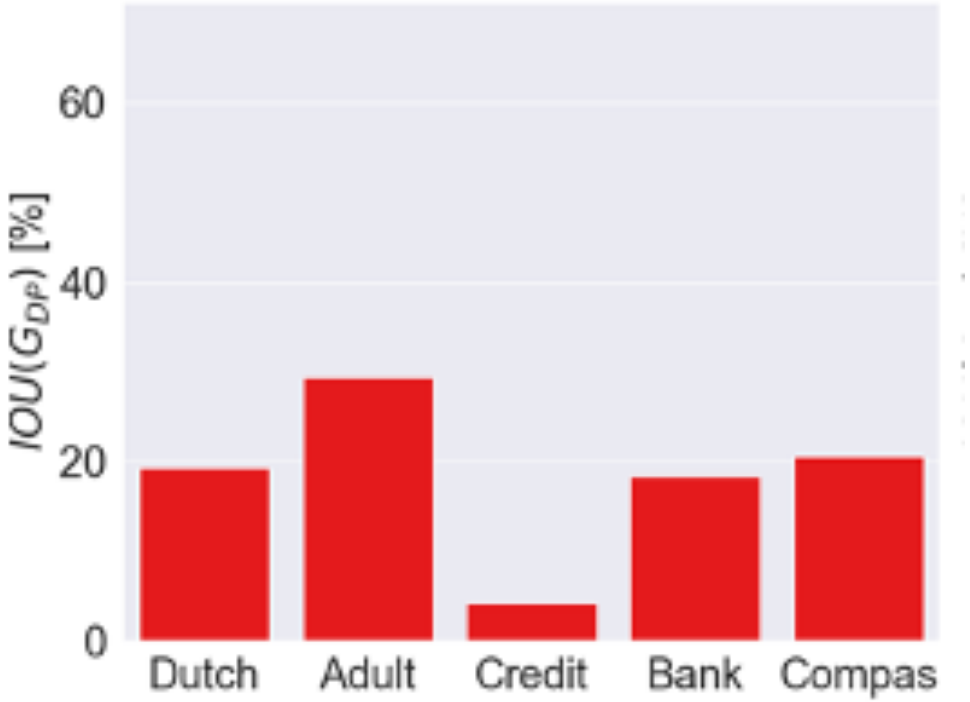
- Blind « Levelling down » effect [1]
- Blind discrimination on other factors [2]
- Arbitrariness in general

[1] The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default, Mittelstadt et al. 2021
[2] On the Fairness Road: Robust Optimization for Adversarial Debiasing, Grari et al. 2023

Multiplicity in Debiasing

Empirical Study

We measure a **very small overlap** in people “treated” between Fairness approaches



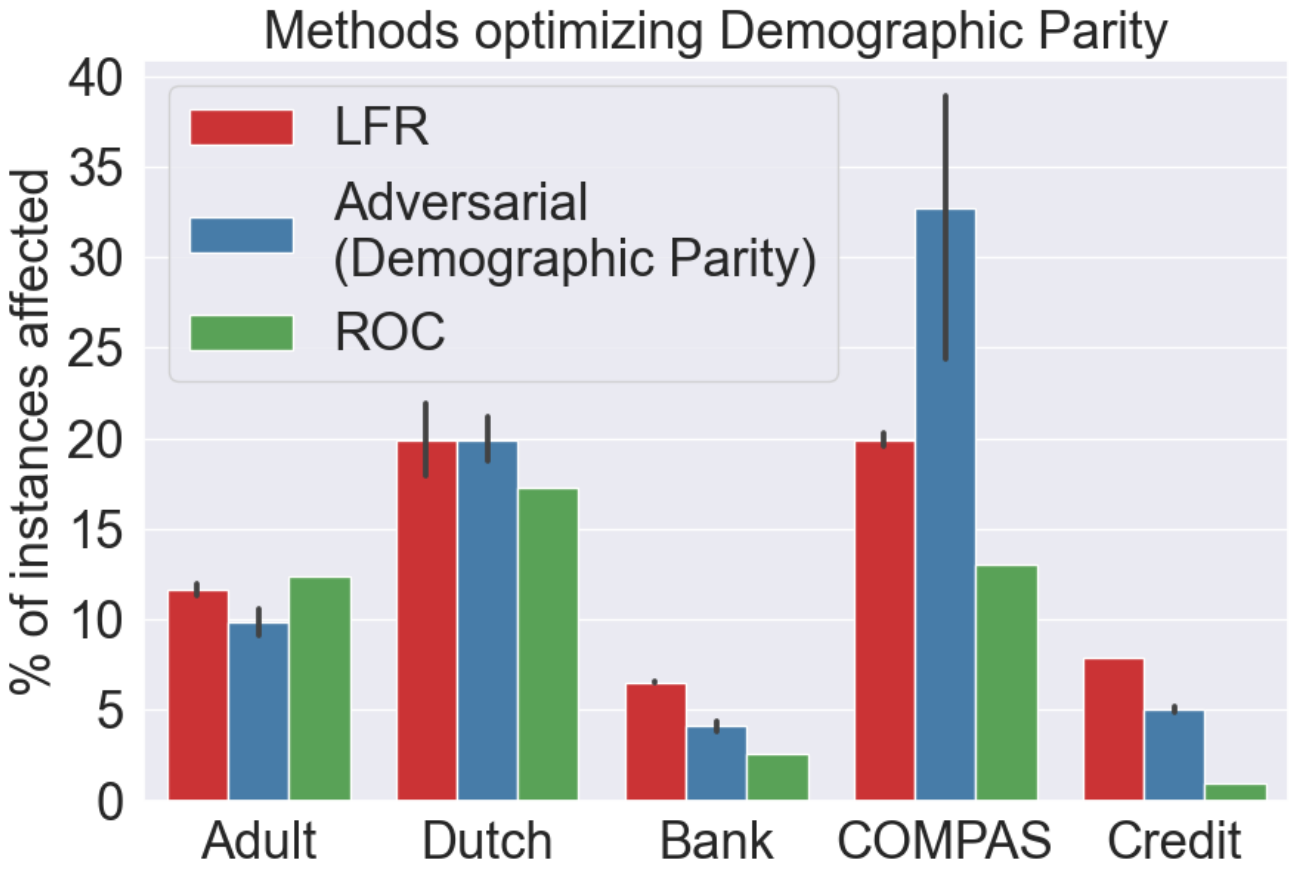
➡ **How are these strategies different?**

Making the debiasing process more transparent

How: **characterizing the debiasing processes to understand their differences**

- Proposed “audit” questions:
  - Q1) How many individuals are affected by the debiasing?
  - Q2) How are the sensitive groups affected?
  - Q3) What consequences for the decision model?
  - Q4) Who are the populations affected?

Q1: How many people are affected by debiasing?



What?

Impact size of the bias mitigation

Why is it important?

Decision consistency, robustness

Trust

Q2: Who are the targeted people ?

	LFR		Adversarial (DP)		ROC	
	Female	Male	Female	Male	Female	Male
Positive Difference	39.25%	0.16%	24.83%	4.18%	0.0%	0.0%
Negative Difference	5.47%	55.1%	10.22%	60.75%	0.0%	100%

What?

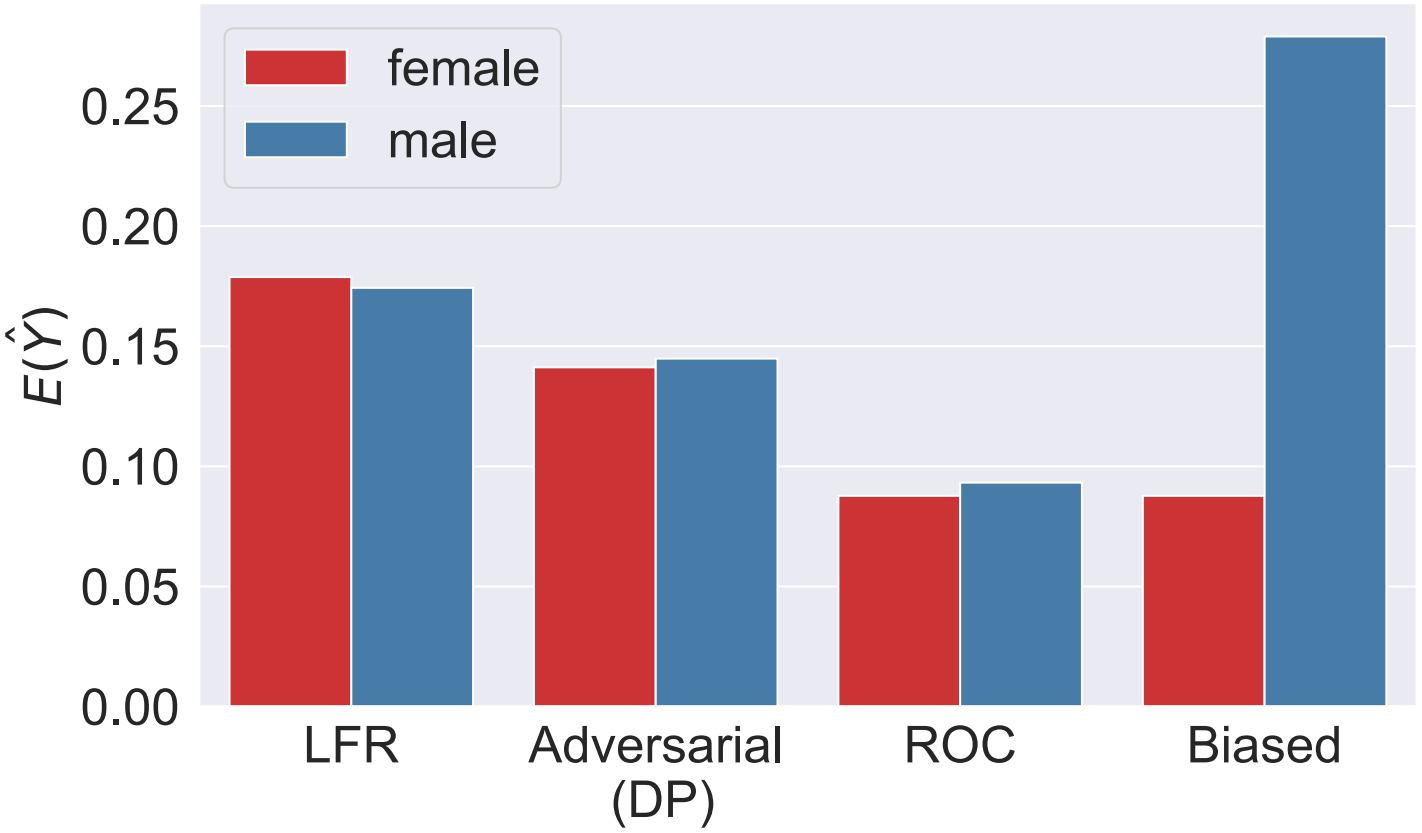
Levelling up vs Levelling down

Why is it important?

Degradation of the service

Q3: What consequences for the decision model?

Subgroup and difference direction



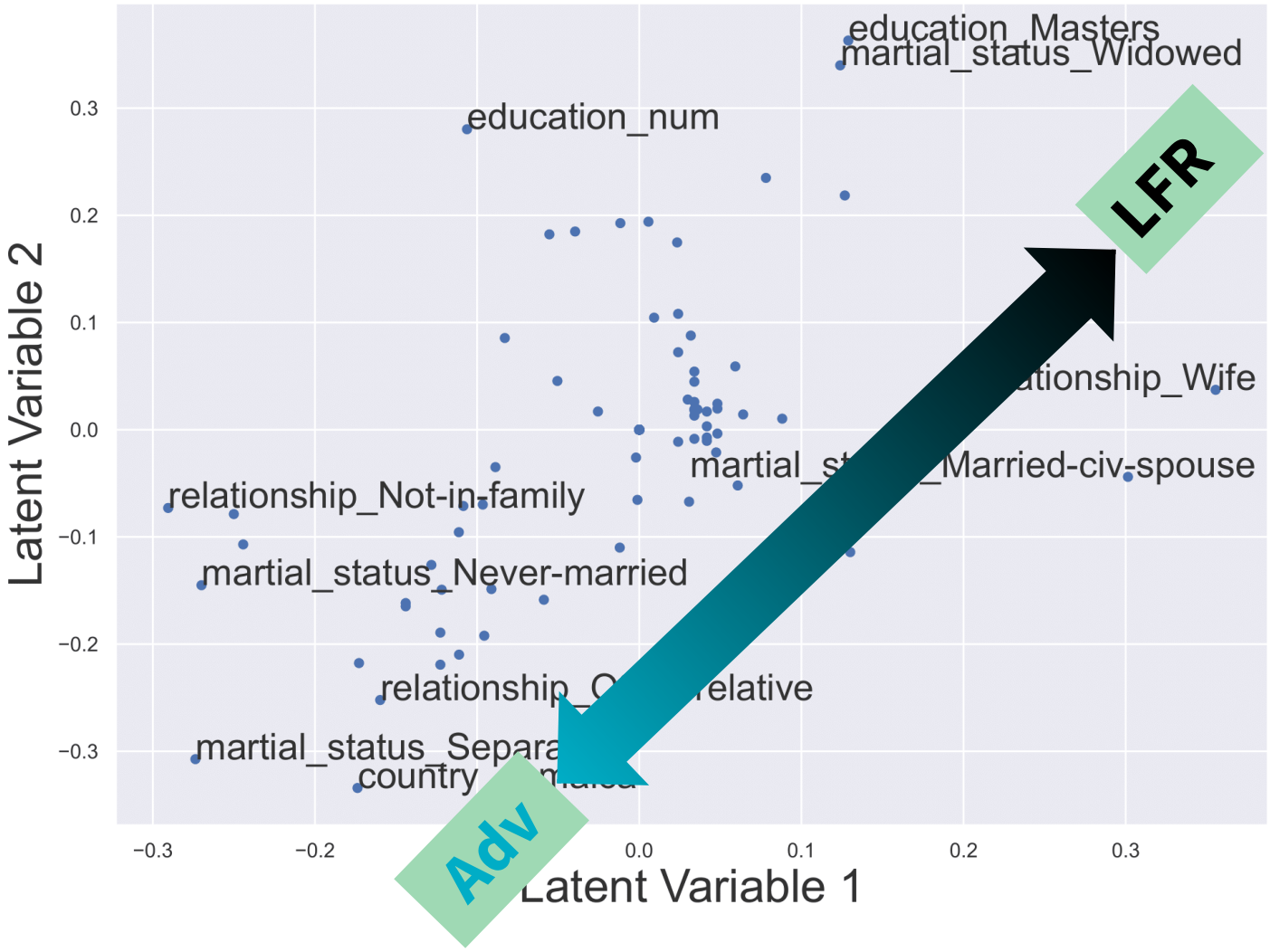
What?

Final acceptance rate of the model

Why is it important?

Broader impact on the general task: budget, resources, rights, etc.

Q4) Who are the populations affected?



What?

XAI to identify affected populations

Why is it important?

Better understanding of the bias

Highlighting possible new biases

Recap

	Impact Size (Q1)	Up vs down (Q2)	Final model state (Q3)	Populations targeted
LFR	++	Balanced	0.17	Married & educated
Adv	+	Balanced	0.14	single
ROC	-	Male down	0.8	all

Conclusion