

110° Nexa Lunch Seminar

Towards Responsible AI in Banking: Addressing Bias for Fair Decision-Making

Alessandro Castelnovo



Politecnico
di Torino

Nexa Center
for Internet & Society

Business Worldwide Declares AI as a Strategic Goal

SCALE ADOPTION OF AI-BASED SOLUTIONS LEAD TO MANY BENEFITS...

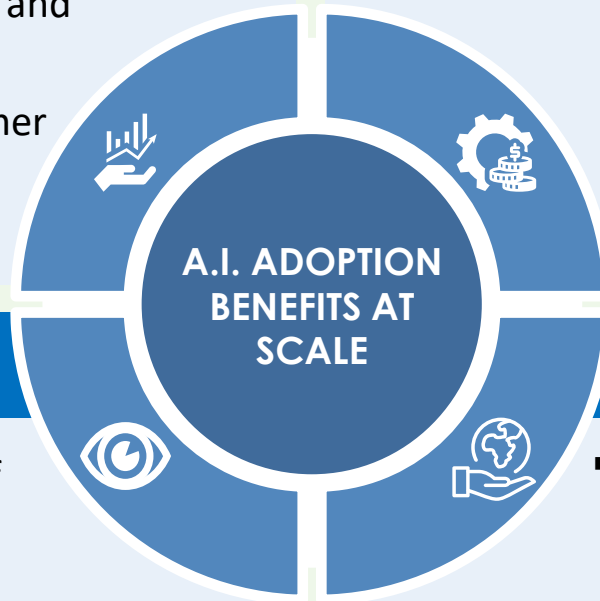
...WHOSE VALUE IS NOW RECOGNIZED WITHIN THE INDUSTRY

SUPPORT FOR REVENUE GROWTH

SIGNIFICANT COSTS REDUCTION

- Improvement of the Customer Experience through personalized offers and experiences
- Improvement of the Customer Satisfaction through dedicated marketing

- Optimization of operational processes by reducing manual activities



OPTIMIZATION OF RISK PROFILES

SUSTAINABILITY AND COMPLIANCE

- Increase in the accuracy of the Credit Score through enrichment of the database for risk analysis

- Automation of Conformity Checks for compliance purposes
- Optimization of profiling for ESG purposes, also through external data

Executives see Artificial Intelligence at the center of business evolution...

84%

of Executives believes in the centrality of A.I. to achieve growth goals¹

... with a potential incremental value of enormous magnitude for the Banking Industry

\$1 Trillion

Potential annual incremental value from AI and analytics for the global banking industry²

AI Implementation Doesn't Always Guarantee Expected Benefits

IBM Watson Flops For Cancer Treatment:
Why Did AI Fail?



IBM's Watson for Oncology cancelled

The dark side of Google Ads

AdFisher: tool to automate the creation of behavioral and demographic profiles.

Used to demonstrate that selling gender = female results in less ads for high-paying jobs.



© 2014 Google. All rights reserved. Data (2014) <https://www.google.com/adsense/adsense/adsense>
Photoshops on PhotoShelter.com. Technology, 2014/11/23-11.2

Google's AdFishertool served significantly fewer ads for high paid jobs to women than men

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

98.7% 68.6% 100% 92.9%

amazon



**DARKER
MALES**



**DARKER
FEMALES**



**LIGHTER
MALES**



**LIGHTER
FEMALES**

Amazon Rekognition Performance on Gender Classification

Amazon's Facial recognition works better for white males

Microsoft's bot Tay taken offline after racist tweets

This Could Represent a Risk for People

data as a social mirror

ML could amplify and perpetuate biases already present in data, at large scale

sample size imbalances

ML could disregard minority groups, effectively producing bias even if absent in the data

this can have a huge impact on people's lives
e.g. Recruiting / Loans approval

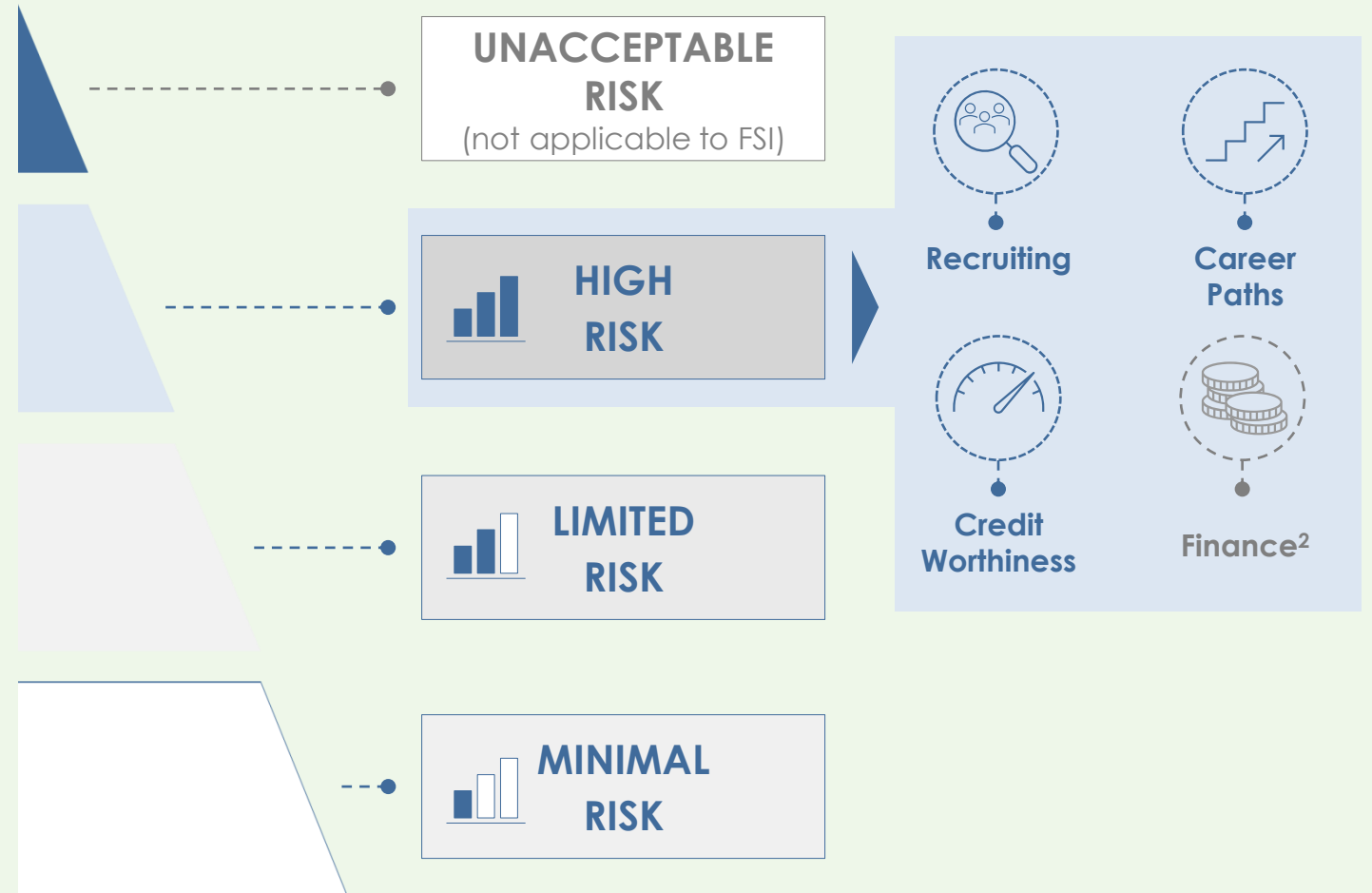
Regulation Risk for Companies - The New European Regulation on AI



The importance of **limiting AI risks** is unequivocally demonstrated by the European Union's proactive efforts to regulate AI, aiming to create a **more favorable environment for the development and deployment of AI**



- Provides for **different levels** of **risk** based on possible **discrimination** and **impacts** on fundamental **human rights** such as
 - Dignity
 - Freedom
 - Equality
 - Solidarity (including health protection)
- Identifies **cross-sectoral "high risk" A.I. systems** and contains no specific **provisions** for **FSI**



Reputation Risk for Companies

HOME > TECH

Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women

Isobel Asher Hamilton Oct 10, 2018, 11:47 AM



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

INSIDER

GOOGLE IS POISONING ITS REPUTATION WITH AI RESEARCHERS

The firing of top Google AI ethics researchers has created a significant backlash

By James Vincent | Apr 13, 2021, 9:30am EDT

THE VERGE

WILL KNIGHT BUSINESS 11.19.2019 09:15 AM

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute.

Control Measures for promoting trust in A.I. Solutions



The adoption of AI by a company should be contingent upon widespread understanding, not only among data scientists and developers but also within governance and compliance structures

CONTROL MEASURES PROVIDED DEPENDING ON THE RISK LEVEL

FAIRNESS

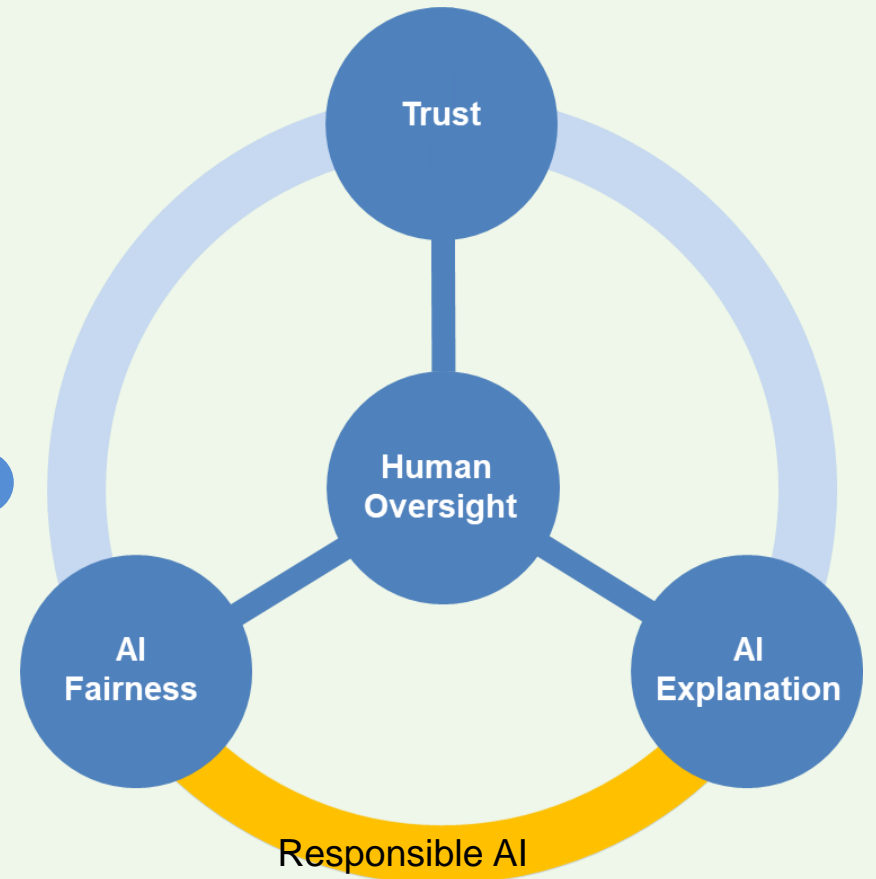
EXPLAINABILITY

HUMAN OVERSIGHT

fairness enables the **measurement** and **mitigation** of **undesired biases**, ensuring that AI systems exhibit **desirable ethical characteristics**

explanations for an AI system provide **human-understandable interpretations** of its inner workings and outcomes

Definition of **supervision** and **human monitoring mechanisms** on **A.I. systems**, on the outputs **produced** and on the **resulting decisions**



Promoting trust in AI-based decision-making requires the integration of three key elements: fairness, explainable Artificial Intelligence (XAI), and human oversight, which must be combined harmoniously

Contributions Landscape of my thesis

Understanding Bias

Bias and Moral Framework in AI-based Decision Making

1

A framework for generating synthetic data that emulates Fundamental biases

2

Fairness Metrics Landscape

3

Mitigating Bias

Fairness Mitigation

4

FFTree: A Flexible Tree to Mitigate Multiple Fairness Criteria

5

Accounting for Bias

Criteria for choosing a fairness metric

6

Towards Fairness Through Time

7

Addressing Fairness in the Banking Sector

8

Part I

Understanding Bias

This part aims to deepen our understanding of how bias is generated, how it manifests in the data, and how it impacts the outcomes of AI systems.

Bias and Moral
Framework in AI-based
Decision Making

1

A framework for generating
synthetic data that emulates
Fundamental biases

2

Fairness Metrics Landscape

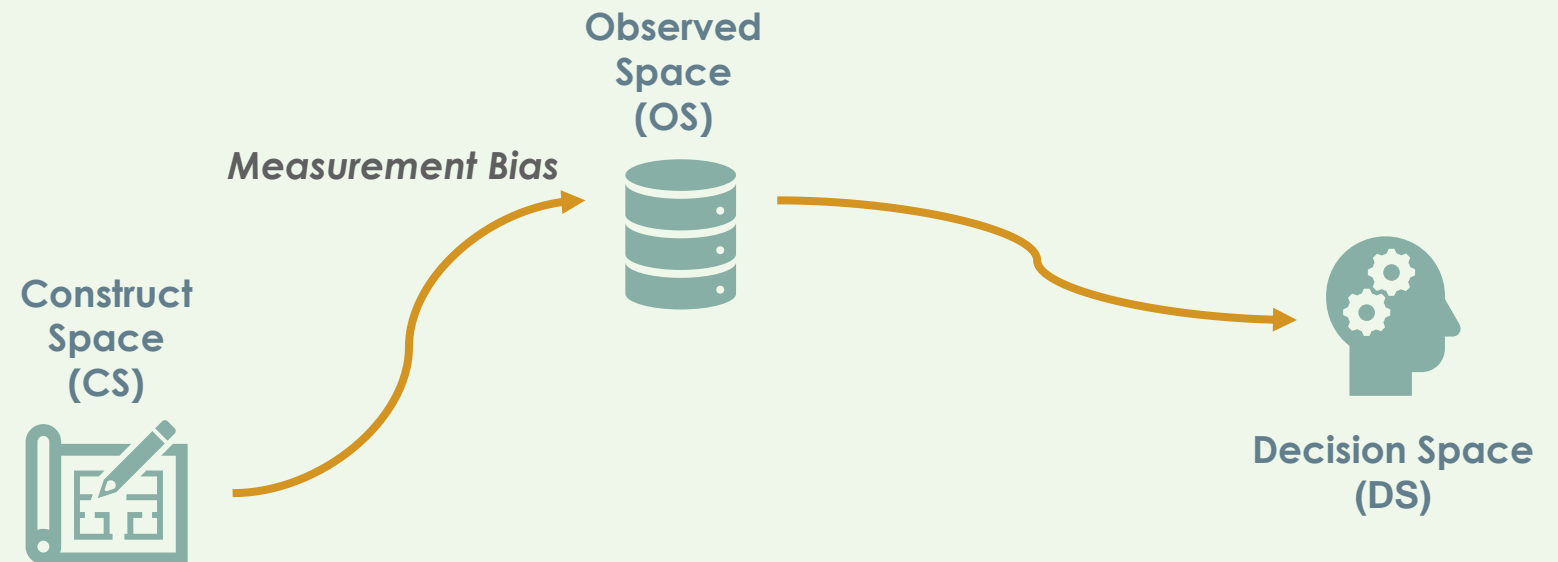
3

Publications related to this part are:

Baumann, J., Castelnovo, A., Crupi, R., Inverardi, N., and Regoli, D. (2023). **Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias**. In *2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, New York

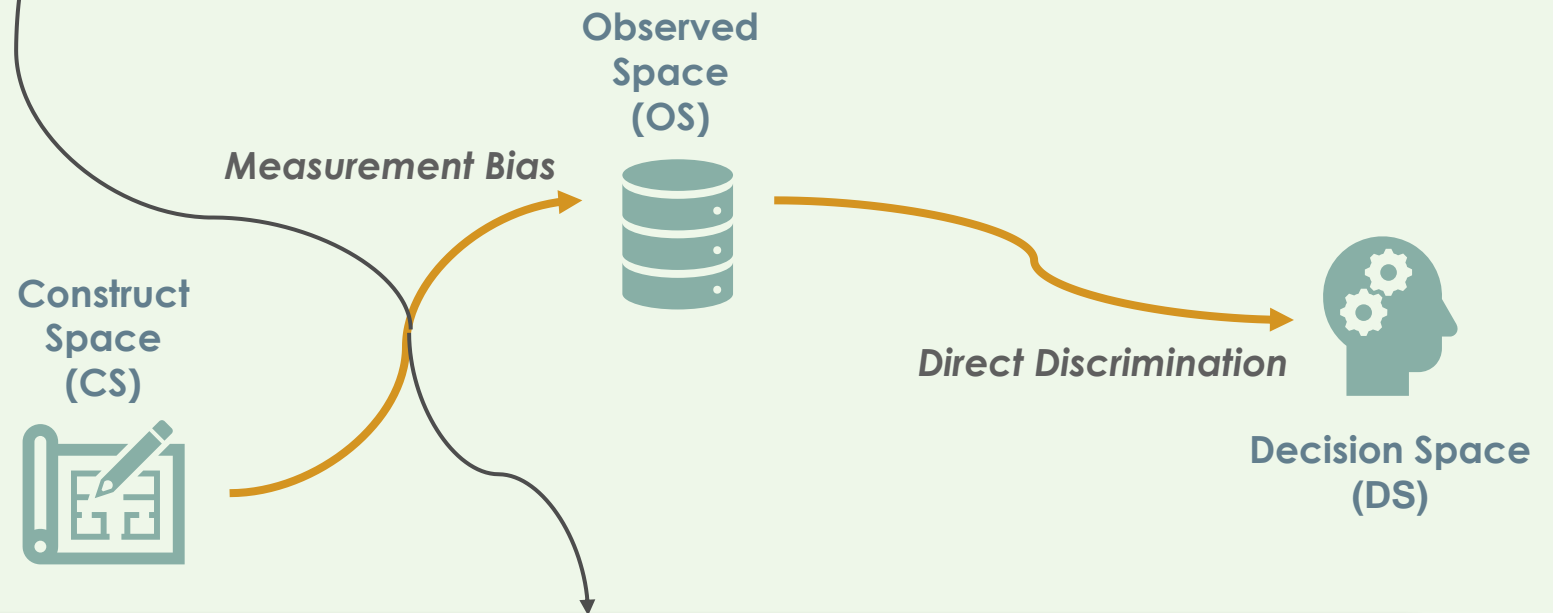
Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022b). **A clarification of the nuances in the fairness metrics landscape**. *Scientific Reports*, 12(1):1–21

Ethical Moral Frameworks for Choosing Fairness in Machine Learning



Ethical Moral Frameworks for Choosing Fairness in Machine Learning

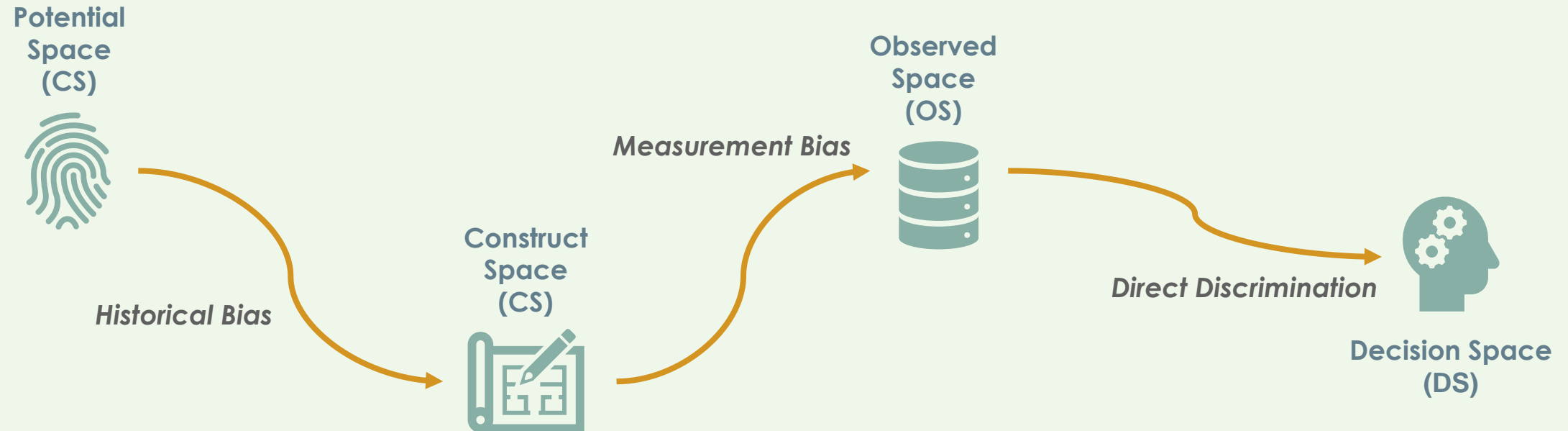
In The *What You See Is What You Get (WYSIWYG)* worldview, CS and OS must be considered equal, and any eventual difference between them is irrelevant to the fairness of the corresponding choice in DS



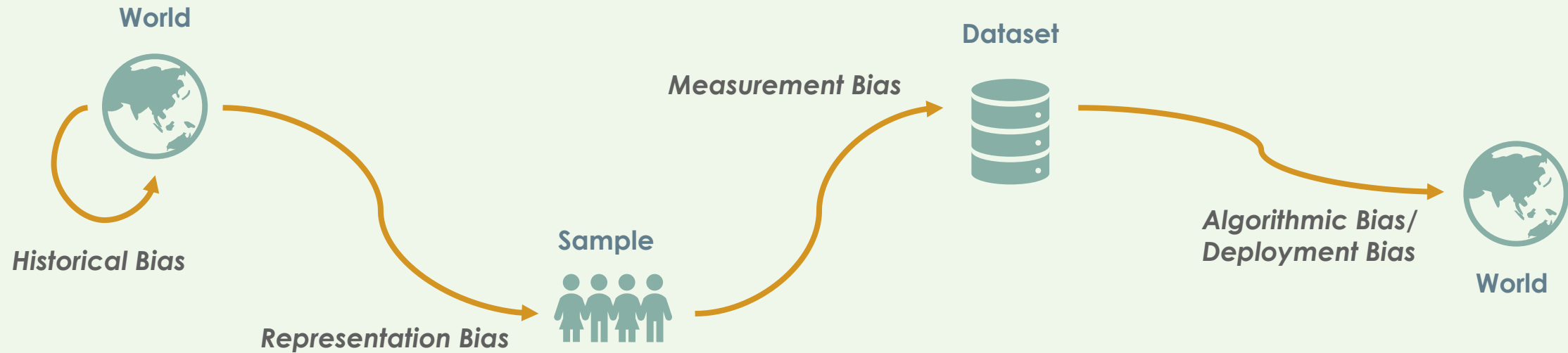
In The *We Are All Equal (WAE)* worldview, states that individuals are all equal at a certain point in time in CS. Therefore, in this perspective, any distortion detectable between CS and OS must be interpreted as caused by a biased observation method corresponding to an unfair mapping.

Ethical Moral Frameworks for Choosing Fairness in Machine Learning

In The *What You See Is What You Get (WYSIWYG)* worldview, CS and OS must be considered equal, and any eventual difference between them is irrelevant to the fairness of the corresponding choice in DS



Bias Throughout the ML Life Cycle



Family of Biases

Bias From Users to Data

Bias is present in the underline phenomenon that generates the data

$$Y = f(X) + \epsilon$$

Variables needed by the model

Historical/Life Bias

Bias From Data to Algorithm

Bias is due to the data collection mechanism

$$\tilde{X} = g(X);$$

$$\tilde{Y} = h(Y)$$

Variables used by the model

Measurement Bias

Representation/Sampling Bias

Omission Bias

Bias From Algorithm to User

Bias is due to the predictor/classification mechanism

$$\hat{Y} = \hat{f}(\tilde{X})$$

Function learned by the model

Algorithmic Bias (Aggregation bias, Learning Bias, Evaluation Bias)

Bias On Demand

***Bias On Demand** is a toolkit that permits to generate synthetic dataset with different combination of bias.*

Advantages of this approach are:

Education



Easy Toolkit to show how different biases arise in data

Investigation



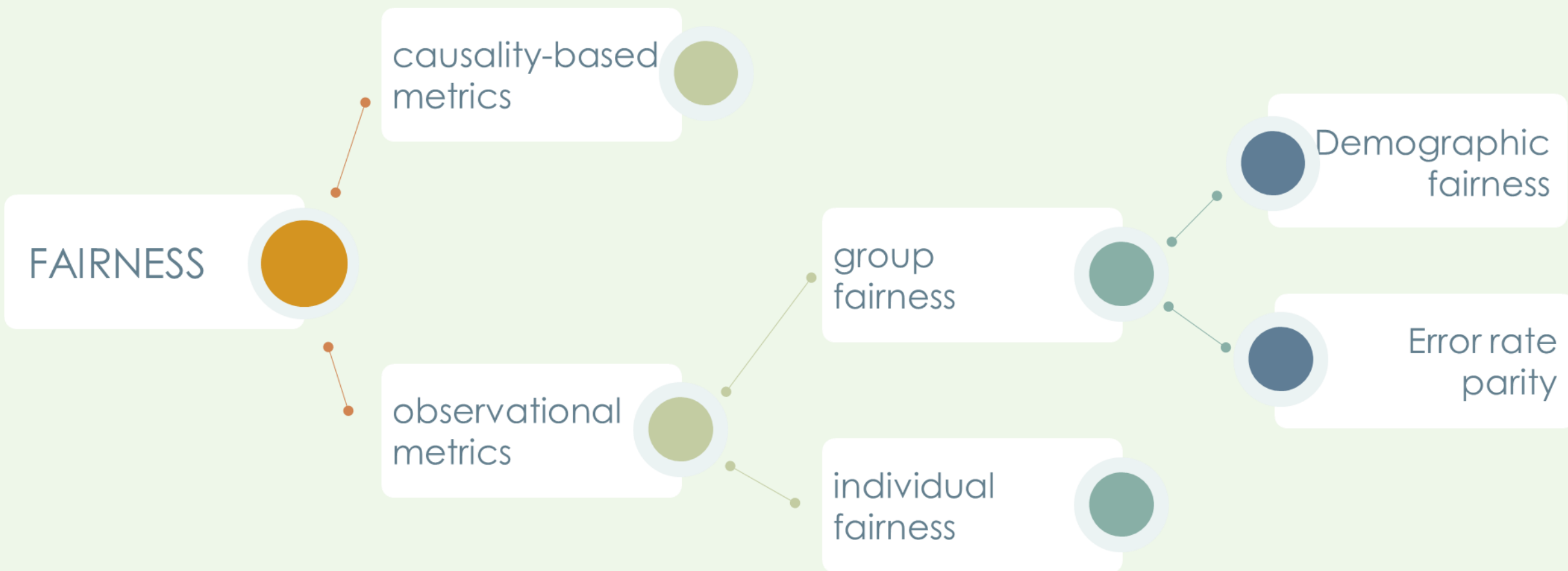
Effects of different bias combinations on performance and fairness evaluation

Research



Ad-Hoc scenarios for developing "bias-aware" systems

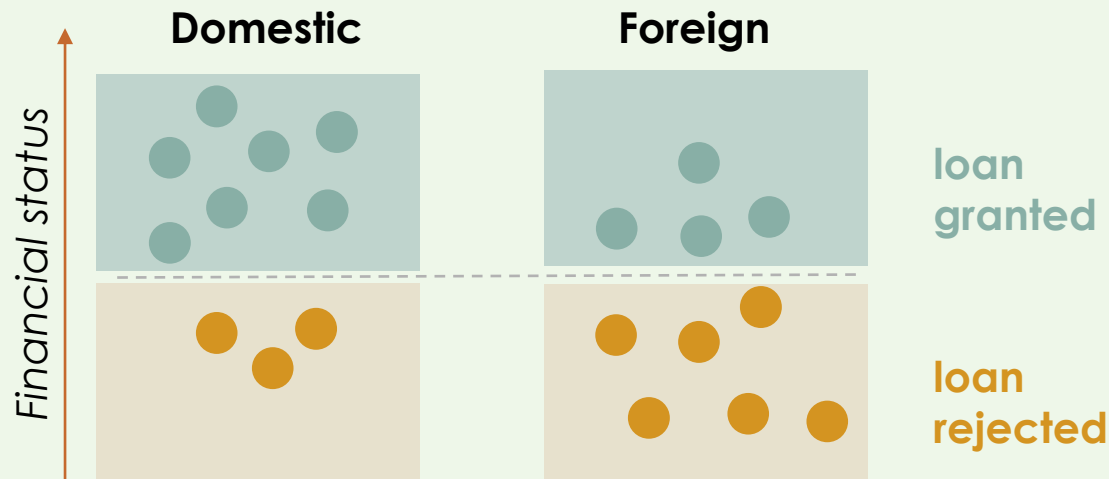
The Zoo of the Fairness Metrics



In General, Fairness Metrics are Non-Compatible With One Another

Individual Fairness

“individuals who are similar (with respect to some task) should be treated similarly (with respect to that task)” (Dwork,2012)

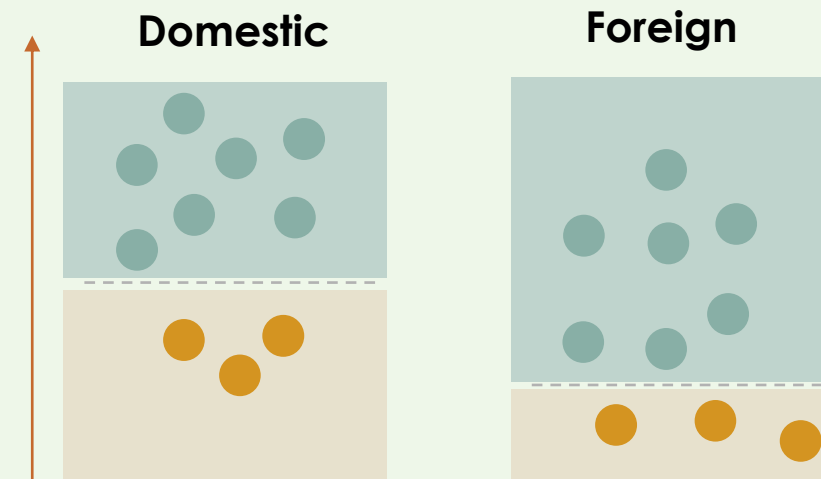


Individual Fairness maintains the **status quo**

In line with WYSIWYG worldview

Statistical Parity

“demographic groups should, *on the whole*, receive equal decisions” (inspired by civil rights law in the US and UK)



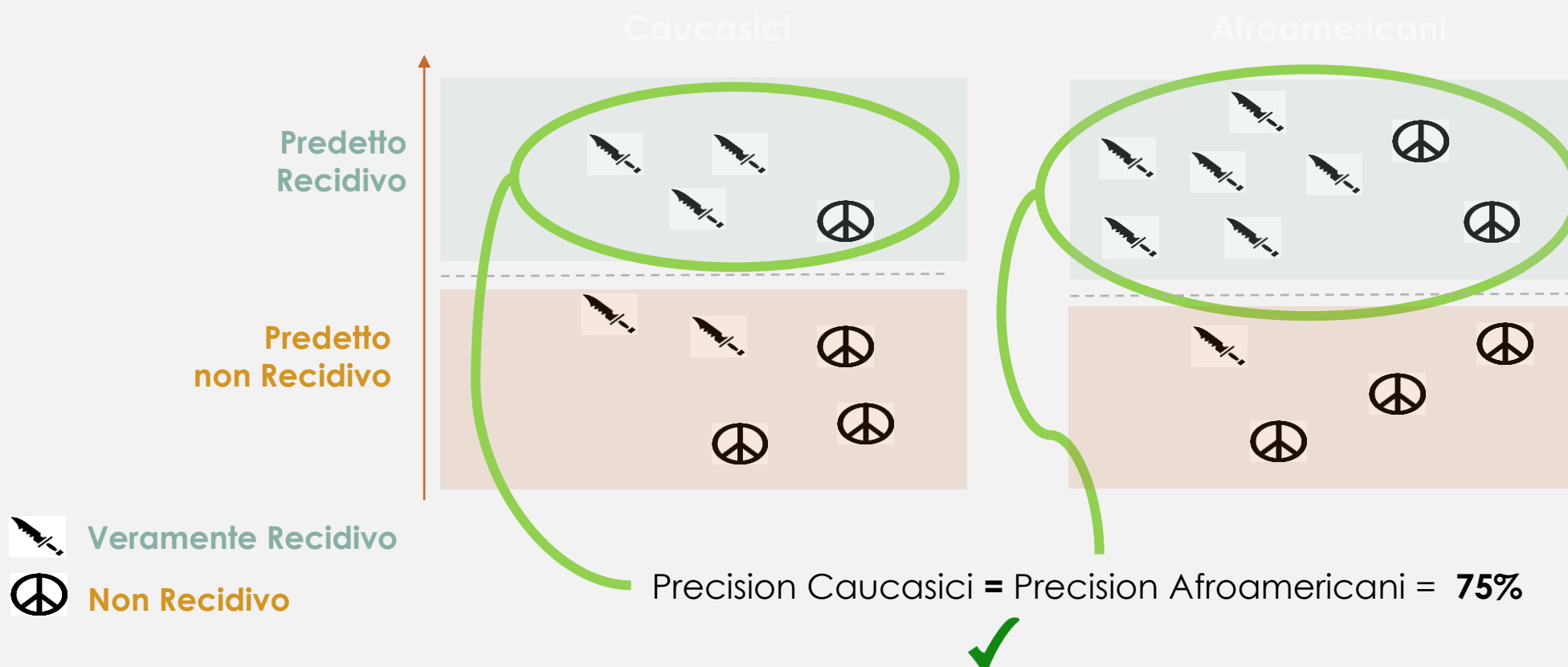
Statistical Parity breaks the **status quo**

In line with WAE worldview

Incompatibility between Error Rate Parity Metric

The Compas debate

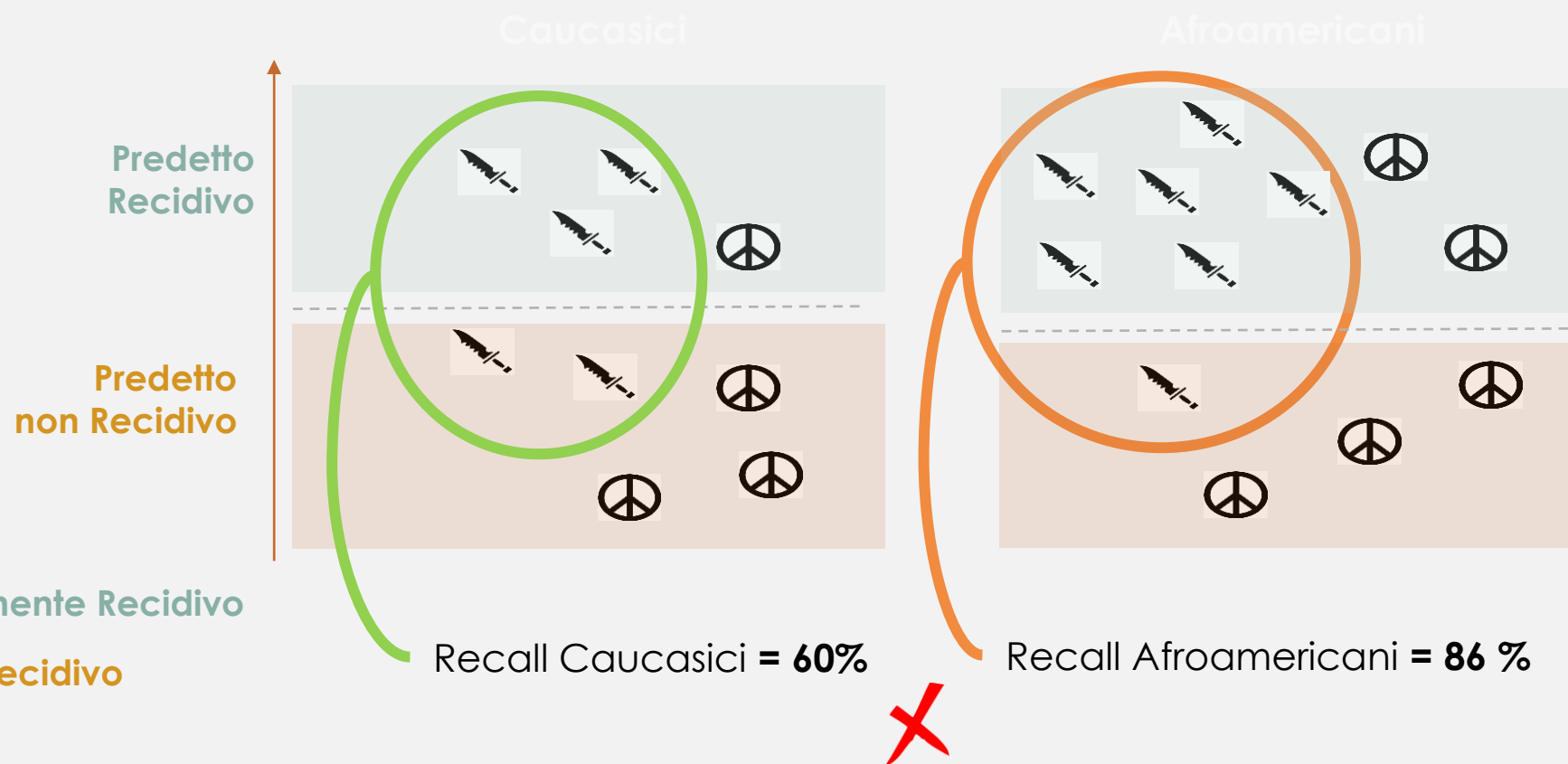
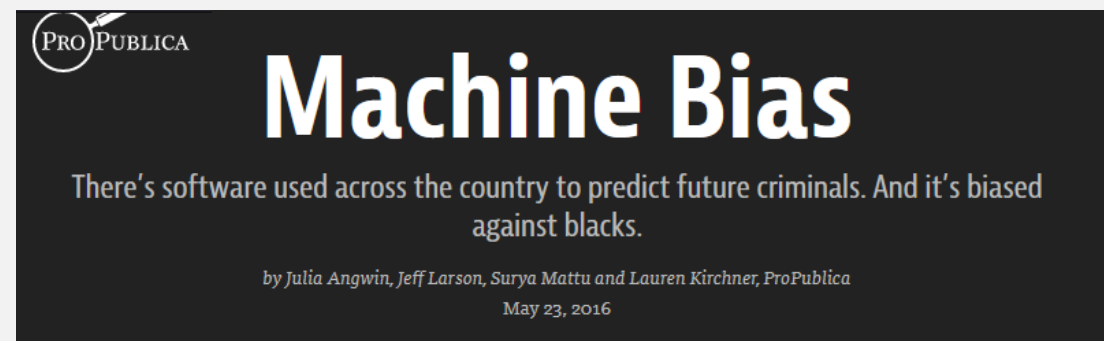
In the USA, a software was developed to predict criminal recidivism. This software, with fairness in mind, was developed to meet **Precision Parity**



Incompatibility between Error Rate Parity Metric

The Compas debate

COMPAS was criticized in the media for being discriminatory as it did not meet **Equal Opportunity**



Part II

Mitigating Bias

This part focus on addressing bias at different stages of AI decision-making, such as pre-processing, in-processing, and post-processing. They aim to mitigate bias by carefully handling data inputs, optimizing learning algorithms, and refining model outputs.

Fairness Mitigation

4

FFTree: A Flexible Tree
to Mitigate Multiple
Fairness Criteria

5

Publications related to this part are:

Castelnovo, A., Cosentini, A., Malandri, L., Mercorio, F., and Mezzanzanica, M. (2022a). Fftree: A flexible tree to handle multiple fairness criteria. Information Processing & Management

Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., and Gonzalez, B. S. M. (2020). Befair: Addressing fairness in the banking sector. In 2020 IEEE International Conference on Big Data (Big Data)

Fairness Mitigation Strategies

pre-processing:



F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," Knowledge and Information Systems, vol. 33, no. 1, pp. 1–33, 2012.

in-processing:



B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

post-processing:



M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in neural information processing systems, 2016, pp. 3315–3323.

BeFair: a Fairness Mitigation Toolkit

To assist data scientists at Intesa Sanpaolo in their efforts to achieve fairness mitigation, we have developed a comprehensive toolkit called **BeFair**

	Mitigation Technique	Demographic Parity	Error Rate Parity	Individual Fairness	Counterfactual Fairness
Pre Processing	FTU			✓	
	Suppression	✓			
	Massaging	✓			
	Sampling	✓			
In Processing	CFF				✓
	AdvDP	✓			
	AdvEO		✓		
	AdvCDP	✓		✓	
	ReductionsGS	✓			
	ReductionsEG	✓			
Post processing	ThreshDP				
	ThreshEO		✓		
	ThreshEopp		✓		
	ThreshCDP	✓		✓	

BeFair Experimentsv(1/2)

Experiment on Fairness Mitigation using Real-World Data on Credit Lending

~200,000 loan applications
 ~50 predictors, including financial variables and personal information.
 The target is the final decision of a human officer.

Throughout the analysis, we focus on

CITIZENSHIP = {0, 1}

as **sensitive attribute** with respect to which assess fairness.

Bias, measured in terms of Demographic Parity, is negligible in the original target, but amplified by the application of a ML model.

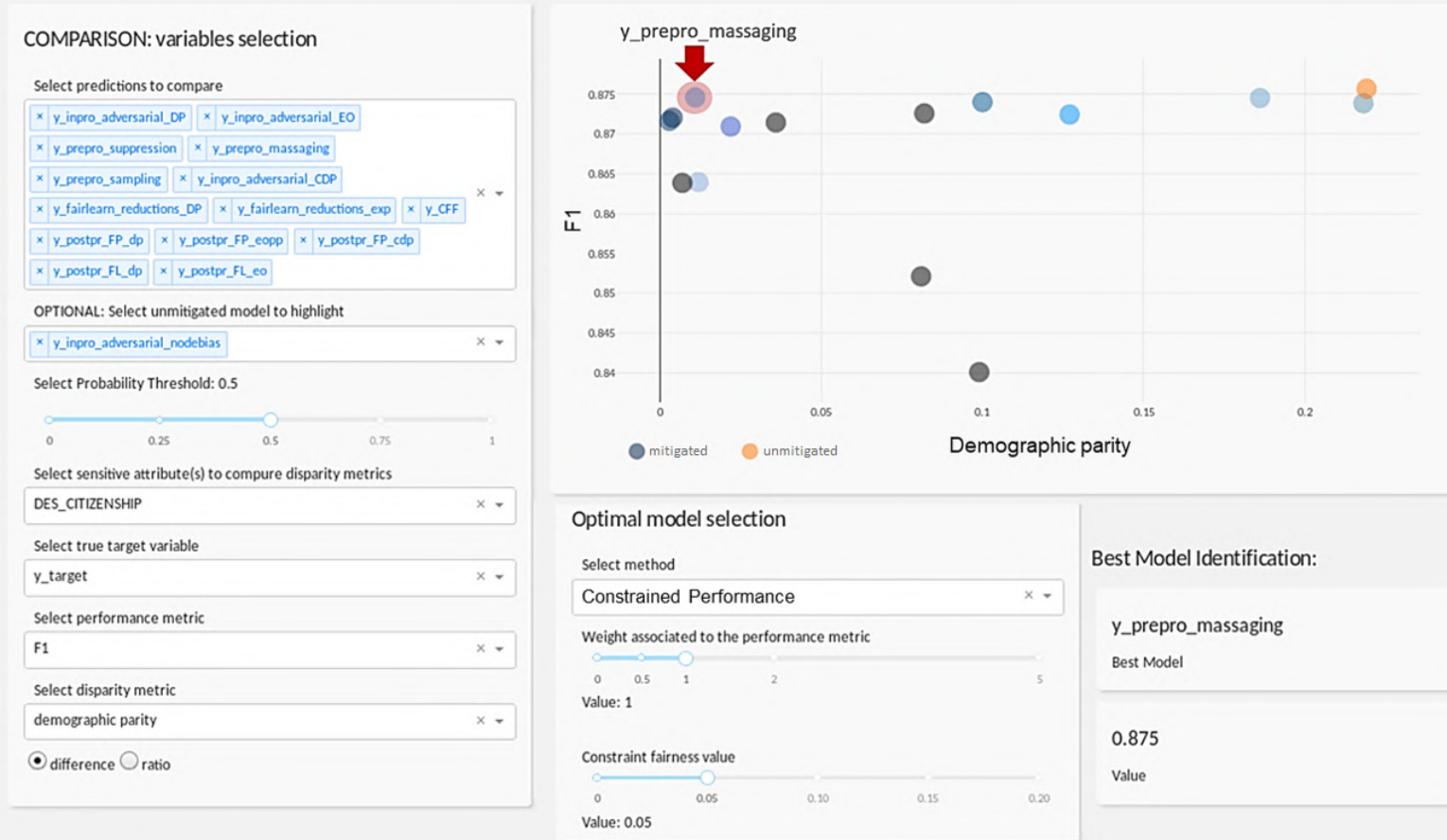
family	type	fairness				performance		
		DP	EO	EOpp	PP	AUROC	Accuracy	F1
no mitigation	Logistic	<u>0.324</u>	<u>0.272</u>	<u>0.272</u>	0.032	0.817	0.761	0.823
	Random forest	0.221	0.202	-0.104	0.068	0.838	0.804	0.875
	Neural network	0.219	0.198	0.104	0.072	0.830	0.811	0.876
pre-process	FTU	0.164	0.124	0.058	0.095	0.838	0.812	0.876
	Suppression	0.099	-0.053	0.065	0.152	<u>0.753</u>	<u>0.748</u>	0.840
	Massaging	-0.004	0.062	0.062	0.163	0.818	0.868	<u>0.803</u>
	Sampling	0.080	0.012	0.012	0.115	0.835	0.791	0.851
	CFF	0.218	0.192	0.104	0.070	0.832	0.810	0.874
in-process	AdvDP	-0.034	0.073	0.063	<u>0.176</u>	0.823	0.802	0.869
	AdvEO	0.102	0.029	-0.010	0.148	0.819	0.805	0.871
	AdvCDP	0.147	0.101	-0.050	0.112	0.830	0.807	0.872
	ReductionsGS	0.012	0.077	0.049	0.159	0.812	0.794	0.864
	ReductionsEG	0.007	0.084	0.051	0.161	–	0.794	0.864
post-process	ThreshDP	0.003	0.099	0.056	0.164	–	0.805	0.872
	ThreshEO	0.082	0.006	0.006	0.138	–	0.812	0.873
	ThreshEOpp	0.100	0.048	0.005	0.119	–	0.809	0.874
	ThreshCDP	0.186	0.159	0.072	0.083	–	0.810	0.875

BeFair Experiments (2/2)

Utilizing the BeFair Interface to Facilitate Optimal Mitigation Approach Selection

Models comparison

compare mitigations disparity and performance



Proposed methods to identify the best performance-fairness tradeoff:

Trade-off fairness-performance

$$(1 + \beta^2) \frac{(1 - |\phi|) * \pi}{\beta^2 * (1 - |\phi|) + \pi}$$

Constrained performance

$$\max_{\phi \leq \Phi} \pi$$

π and ϕ are the preferred performance and fairness metrics, respectively and β is the weight associated with the performance metric.

Common Challenges in Fairness Mitigation

As detailed in [2], the mitigation strategies proposed in prior studies typically lack flexibility with respect to the following aspects:

- ❖ They are specifically designed for only one fairness criterion, and cannot accommodate more than one simultaneously;
- ❖ They cannot ensure fairness with respect to multiple sensitive features simultaneously (e.g., gender and race);
- ❖ They are typically designed as a black box, i.e. they are not directly interpretable.

FFTree: A flexible tree to handle multiple fairness criteria

We present **FFTree**, a new transparent, flexible and fairness-aware classifier. As a novelty, **FFTree** enhances the classical approach introduced in [3] with a new approach to find a "fair" split to:

- ✓ Satisfy a fairness constraint selected from a wide range of possible definitions of fairness;
- ✓ Implement more than one fairness criterion;
- ✓ Handle more than one sensitive attribute at the same time;
- ✓ Set the required level of fairness as an input parameter to meet different business needs or regulatory requirements.

<i>State-of-the-art Fair Tree</i>	DI	DT	DM	MD	MS	BN
Kamiran et al.	✓	✗	✗	✗	✗	✗
Zhang and Ntoutsi	✓	✗	✗	✗	✗	✗
Aghaei et al.	✓	✓	✗	✗	✗	✓
FFTree (our method)	✓	✓	✓	✓	✓	✓

[3] Brieman, Friedman, Olshen, and Stone. Classification and regression trees. Wadsworth Inc.

Part III

Accounting For Bias

This part focus on proposing approaches proactively account for bias by incorporating bias-aware decision-making mechanisms. They also prioritize human involvement, allowing for human intervention and oversight, while ensuring that understandable explanations of AI outcomes are provided.

Criteria for choosing a
fairness metric

6

Towards Fairness
Through Time

7

Addressing Fairness in the
Banking Sector

8

Publications related to this part are:

Castelnovo, A., Malandri, L., Mercurio, F., Mezzanatica, M., and Cosentini, A. Towards fairness through time. In Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021

Castelnovo, A., Inverardi, N., Malandri, L., Mercurio, F., Mezzanatica, M., and Seveso, A. (2023b). Leveraging group contrastive explanations for handling fairness. In World Conference on Explainable Artificial Intelligence, pages 332–345. Springer.

Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., and Gonzalez, B. S. M. (2020). Befair: Addressing fairness in the banking sector. In 2020 IEEE International Conference on Big Data (Big Data)

Effects of Fairness Metrics (1/2)

The choice of fairness metric primarily depends on the willingness to change the status quo

Some metrics that maintain the *status quo*:

Individual
Fairness



Error Rate
Parity



"A perfect model is also perfectly fair"

In line with WYSIWYG worldview

Some metrics that change the *status quo*:

Statistical Parity



Causality-based
Metrics

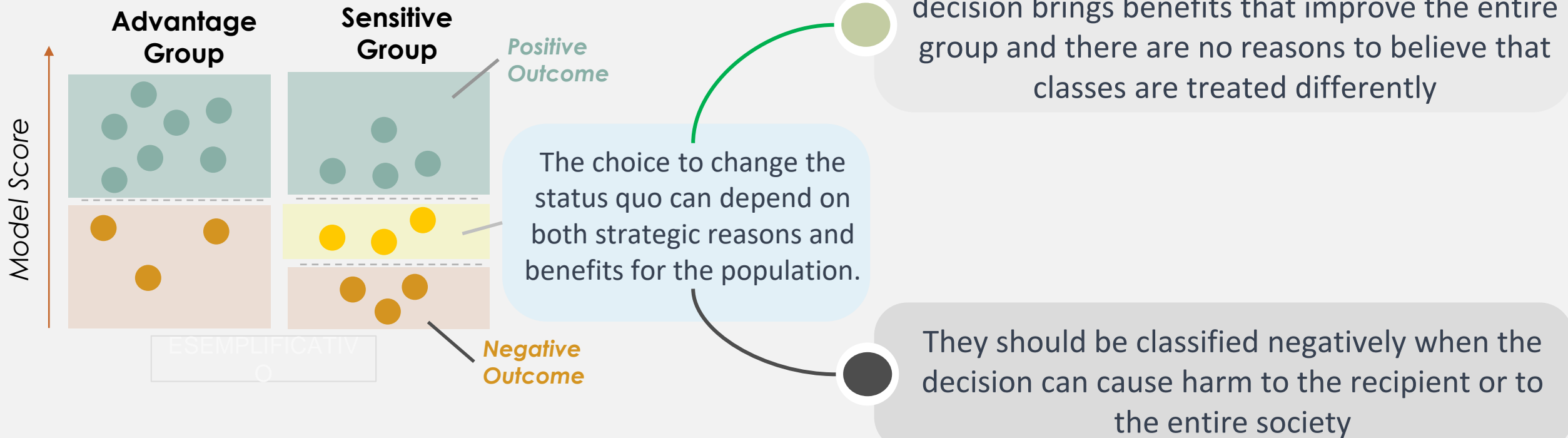


"A perfect model cannot be fair"

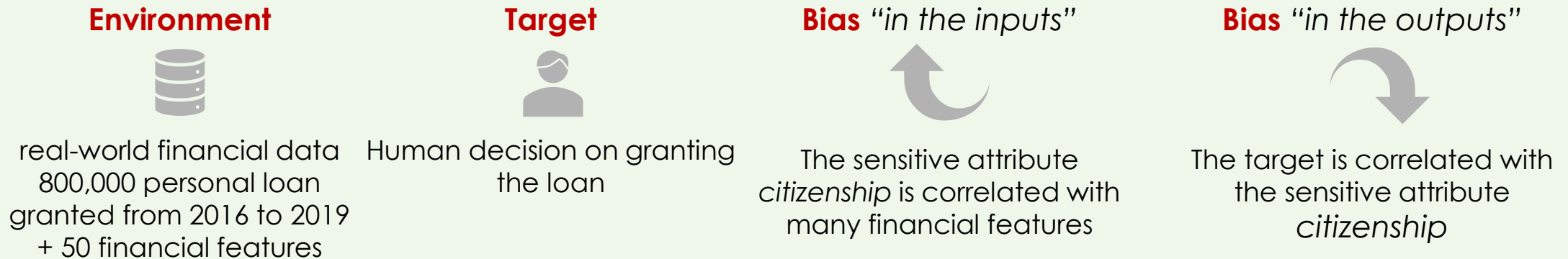
In line with WAE worldview

Effects of Fairness Metrics(2/2)

La scelta di cambiare lo status quo può dipendere sia da ragioni strategiche, sia di beneficio per la popolazione. In alcuni casi, decisioni positive possono creare danni.



Monitoring Fairness Through Time



Chosen Mitigation Policy:

Deploy a ML model that ensure Demographic Parity

To lead an improvement to the vulnerable class and reach in **long-term** DP and Individual Fairness
 simultaneously - **Optimal Situation**

Challenging Questions

C1

Will the outputs of a mitigation
 model continue to ensure
 Demographic Parity over
 time?

C2

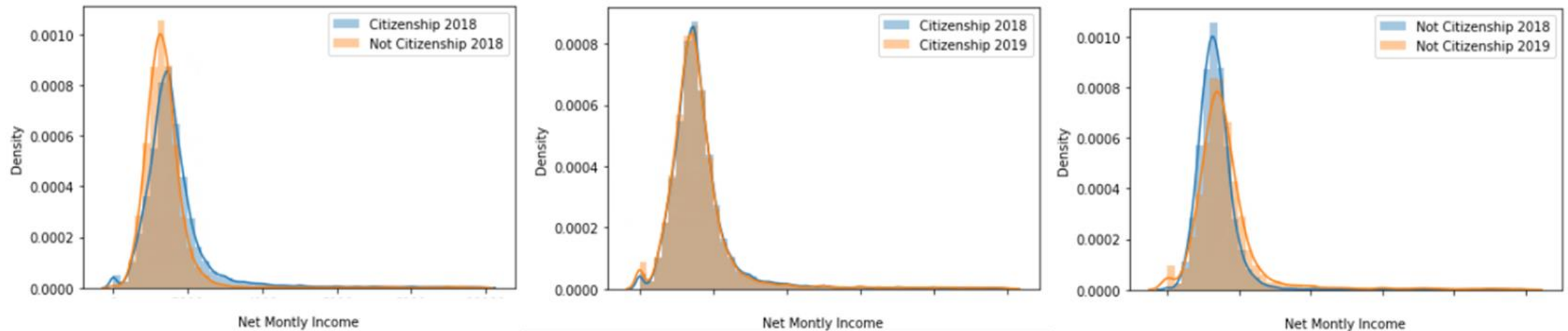
How can XAI techniques be used to verify
 that the chosen fairness policy (*ensure
 Demographic Parity*) is helping to reduce
 individual discrimination over time?

1st Challenging Questions

7

Will the outputs of a mitigation model continue to ensure Demographic Parity over time?

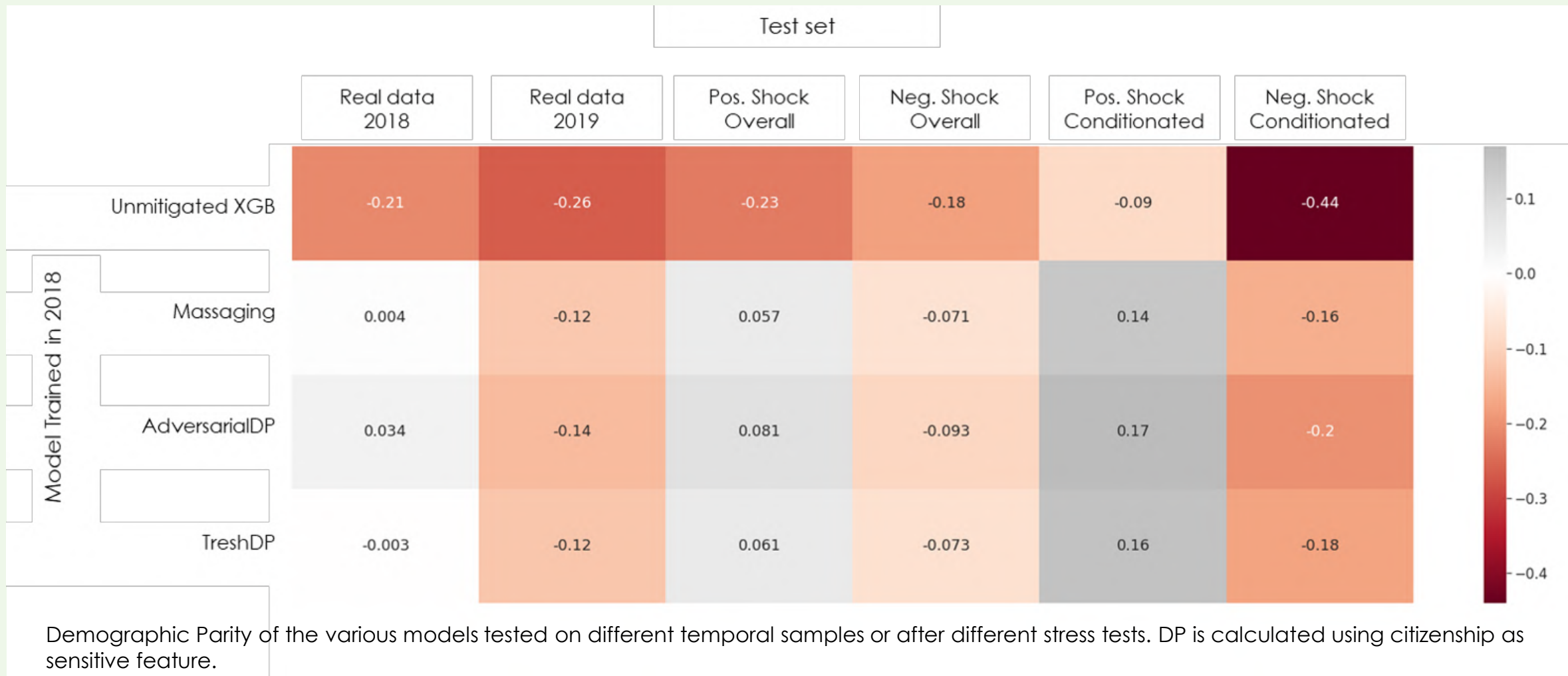
«net montly income» distribution compared between citizenship and years



Density plot of the variable net montly income conditioned to vary combination of citizenship and year. Distribution values are blinded for data privacy.

1st Challenging Questions

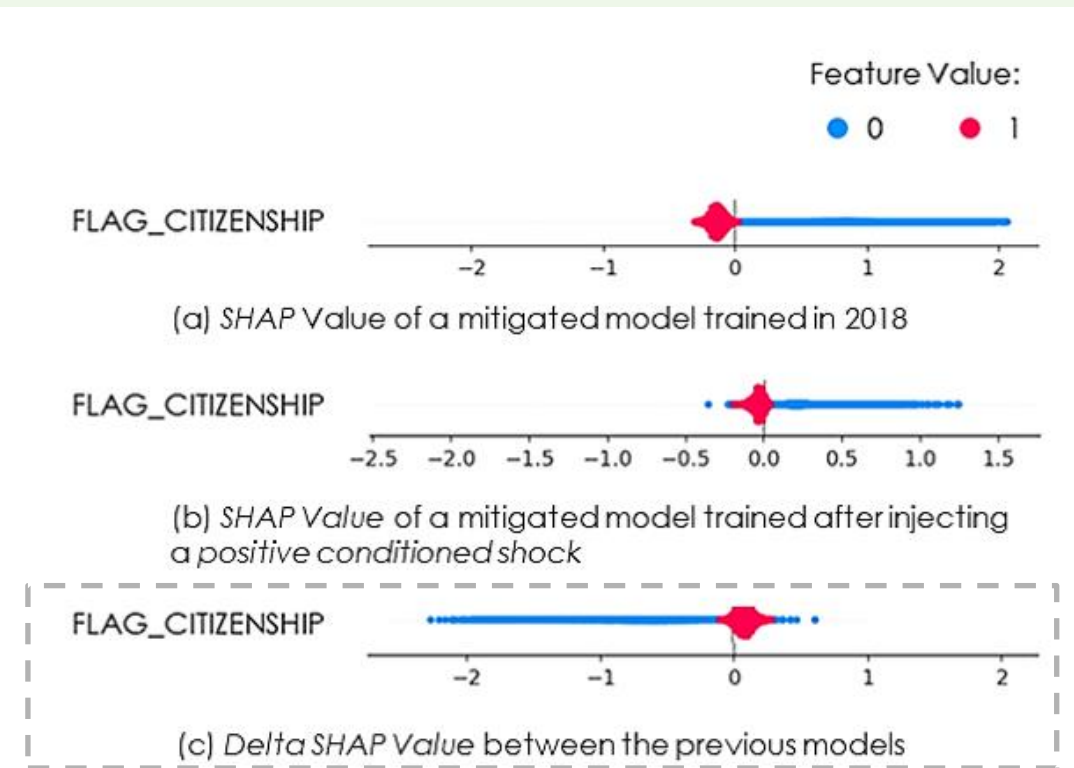
Will the outputs of a mitigation model continue to ensure Demographic Parity over time?



2nd Challenging Questions

How can XAI techniques be used to verify that the chosen fairness policy (ensure Demographic Parity) is helping to reduce individual discrimination over time?

FairX



Representation of the Shapley values of two mitigated models trained in 2018, after injecting a positive conditioned shock and the relative differences.

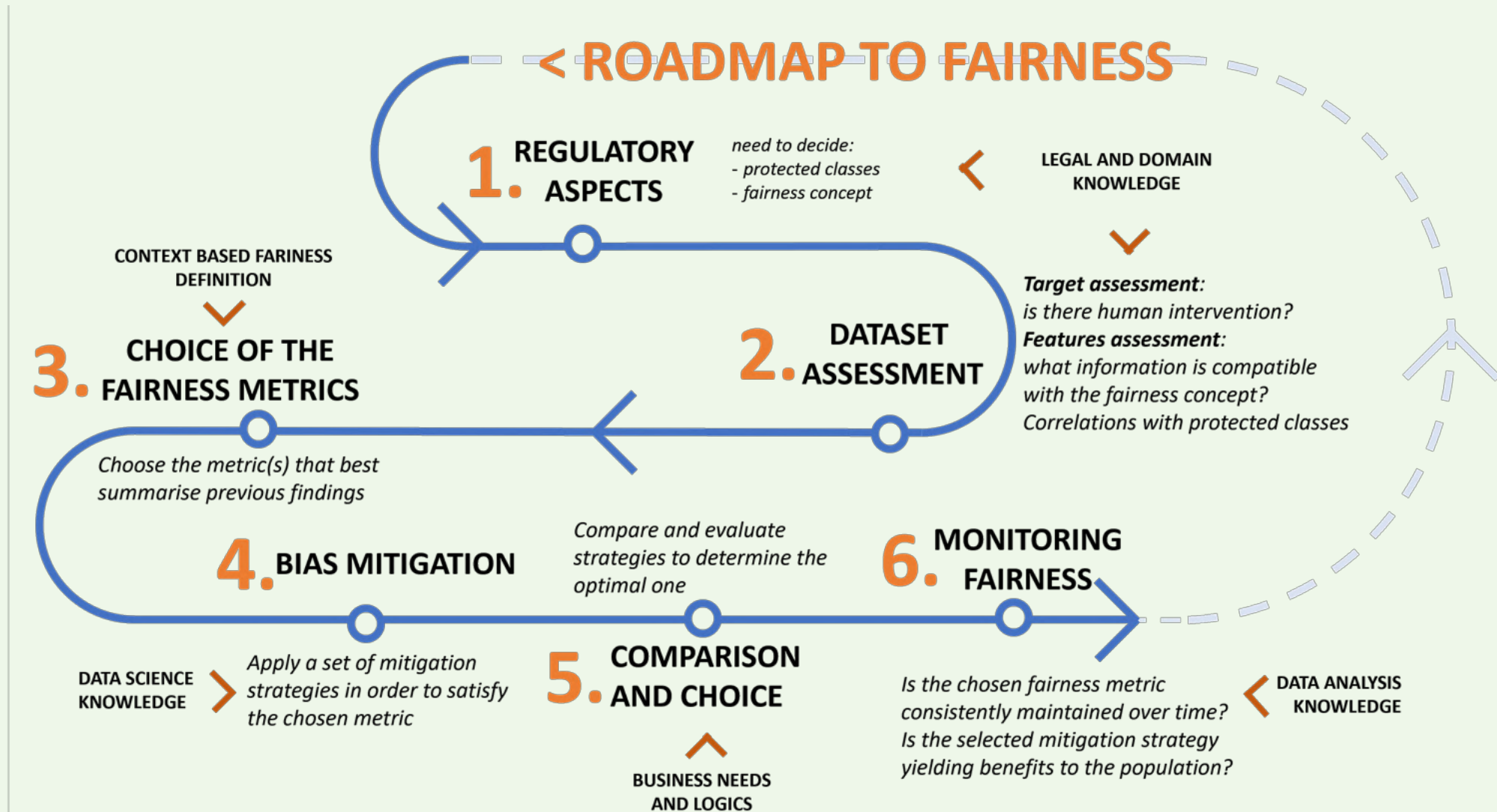
The group mitigation model has to assign a **marginal contribution** to the vulnerable class to provide demographic parity in the outcome

The marginal contribution on the sensitive variables is a **proxy** of individual discrimination

SHAP helps to observe the marginal contribution

△ **Shapley values** are reasonable to observe changes in individual discrimination between the two models

A Roadmap for Addressing Fairness in the Banking Sector



SCIENTIFIC PUBLICATIONS

- [1] **BeFair: Addressing Fairness in the Banking Sector**
Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)

- [2] **Towards Fairness Through Time**
Machine Learning and Principles and Practice of Knowledge Discovery in Databases. ECML PKDD 2021.

- [3] **Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness**
International Journal of Human-Computer Interaction 2022

- [4] **A clarification of the nuances in the Fairness metrics landscape**
Scientific Reports 2022

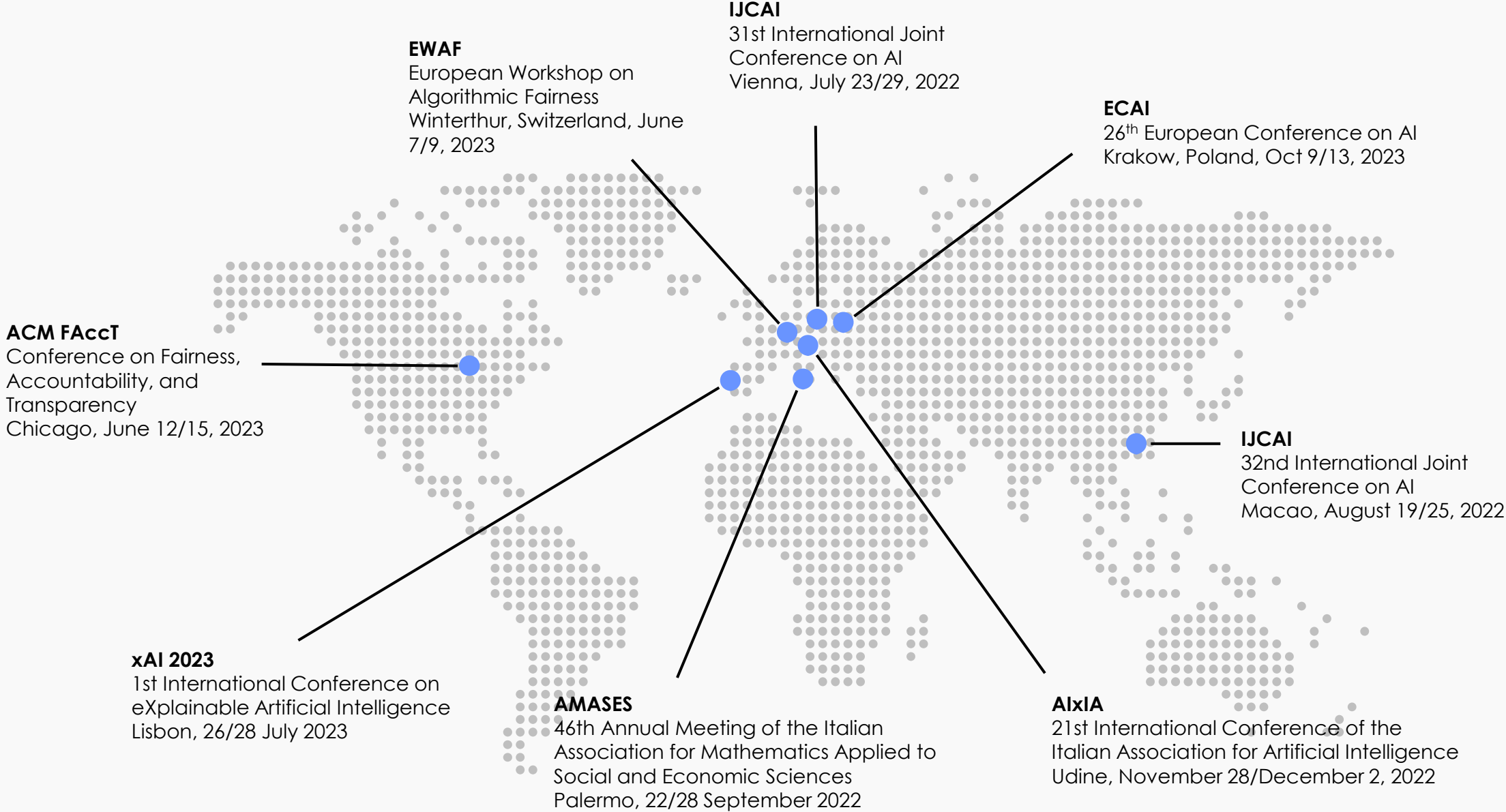
- [5] **Counterfactual Explanations as Interventions in Latent Space**
Data Mining & Knowledge Discovery 2022

- [6] **FFTree: A flexible tree to handle multiple fairness criteria**
Information Processing & Management 2022

- [7] **Bias on Demand: A Modelling Framework that Generates Synthetic Data with Bias.**
Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency

- [8] **Declarative Encoding of Fairness in Logic Tensor Networks**
Accepted for publication in the *Proceedings of the 26th European Conference of Artificial Intelligence*

SCIENTIFIC CONFERENCES



The image features a background of dense green foliage, including leaves and thin branches, which is partially obscured by a wide, horizontal white band. The text "Thank you" is centered within this white band.

Thank you