COUNTERFACTUAL REASONING FOR RESPONSIBLE AI ASSESSMENT

Politecnico di Bari

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, and Eugenio Di Sciascio

Azzurra Ragone

azzurra.ragone@uniba.it

Scenario: financial domain

The decision to **approve** or **deny credit** is regulated with precise and detailed regulatory compliance requirements (i.e., *Equal Credit Opportunity Act*, *Federal Fair Lending Act*, *Consumer Credit Directive for EU Community*).

These rules aim to **prevent discrimination** in human decision-making processes.

What about AI-based decision-making systems?

Starter point

Current regulations require **discarding sensitive features** (e.g., *gender*, *race*, *religion*) in the algorithm's decision-making process to prevent unfair outcomes

Fairness under unawareness

Even without sensitive features in the training set, algorithms can persist in discrimination.

When sensitive features are omitted (*fairness under unawareness*), they could be inferred through non-linear relations with the so-called **proxy features**

OUR RESEARCH GOAL

To reveal the **potential hidden bias** of a machine learning model even when **sensitive features** are discarded

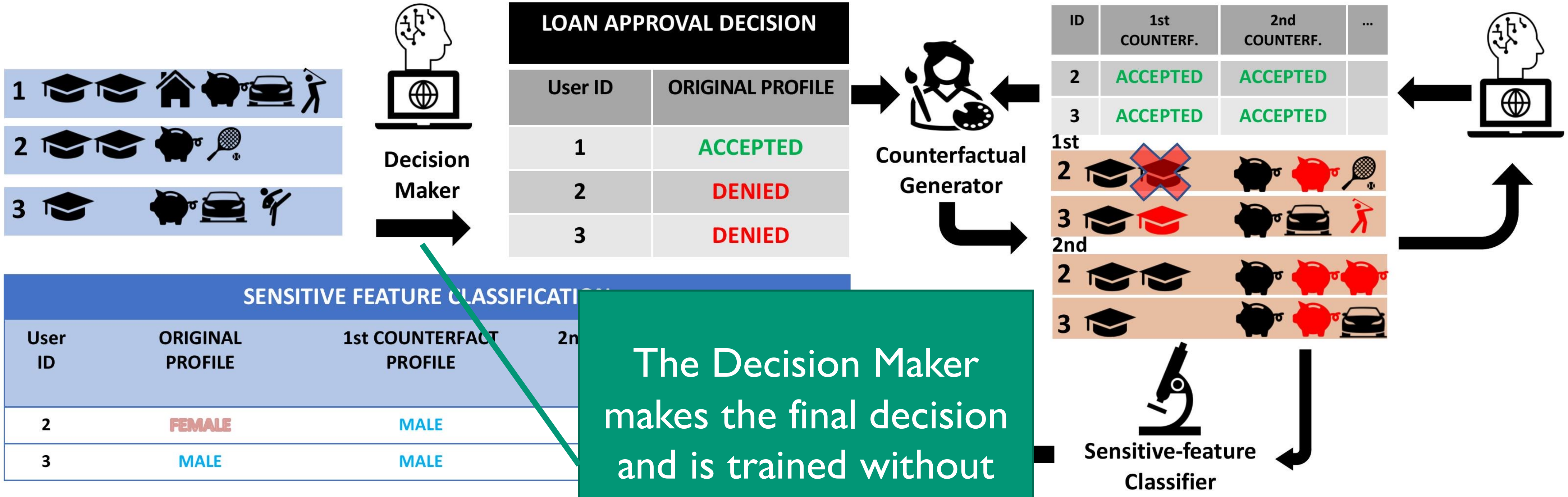
Our study

We study how to unveil whether a black-box predictor is biased in *fairness under unawareness* setting by exploiting **counterfactual reasoning**

Research Questions

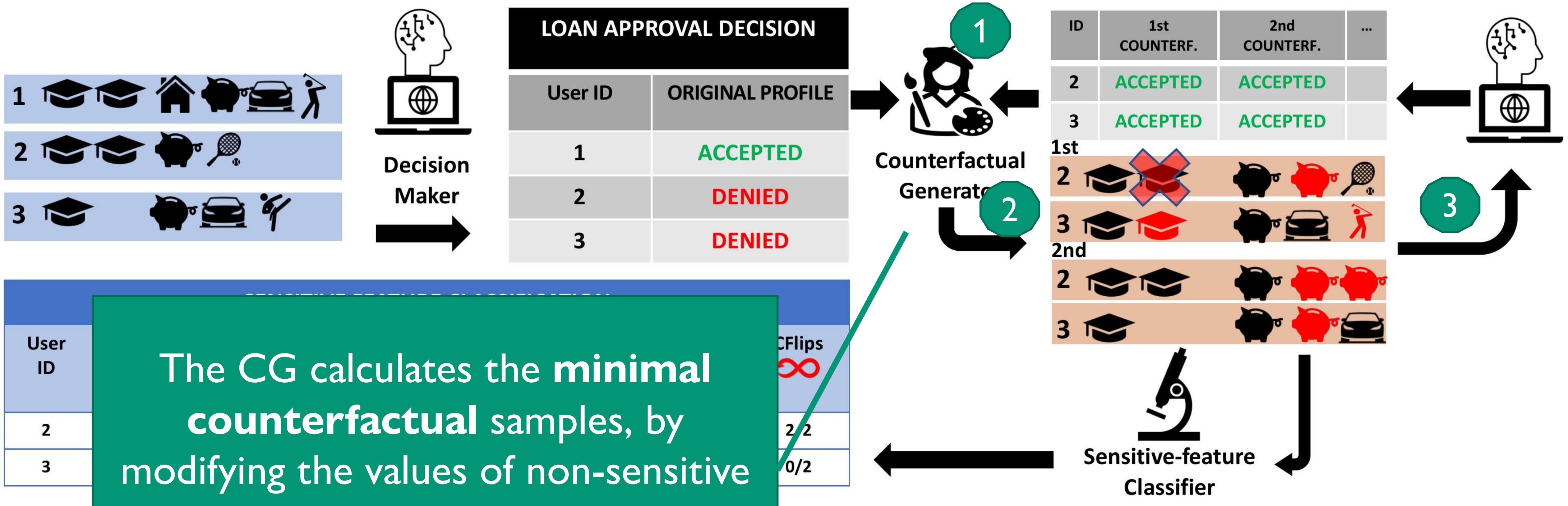
- **RQ1**: Is there a method for determining whether a dataset **contains proxy features** or not?
- **RQ2**: Does the **Fairness Under Unawareness** setting ensure that decision biases are avoided?
- **RQ3**: Is **counterfactual reasoning** effective for discovering decision biases?
- **RQ4**: Is it possible to define a strategy for **identifying the proxy features**?

Example: loan application



The Decision Maker makes the final decision and is trained without using the sensitive features

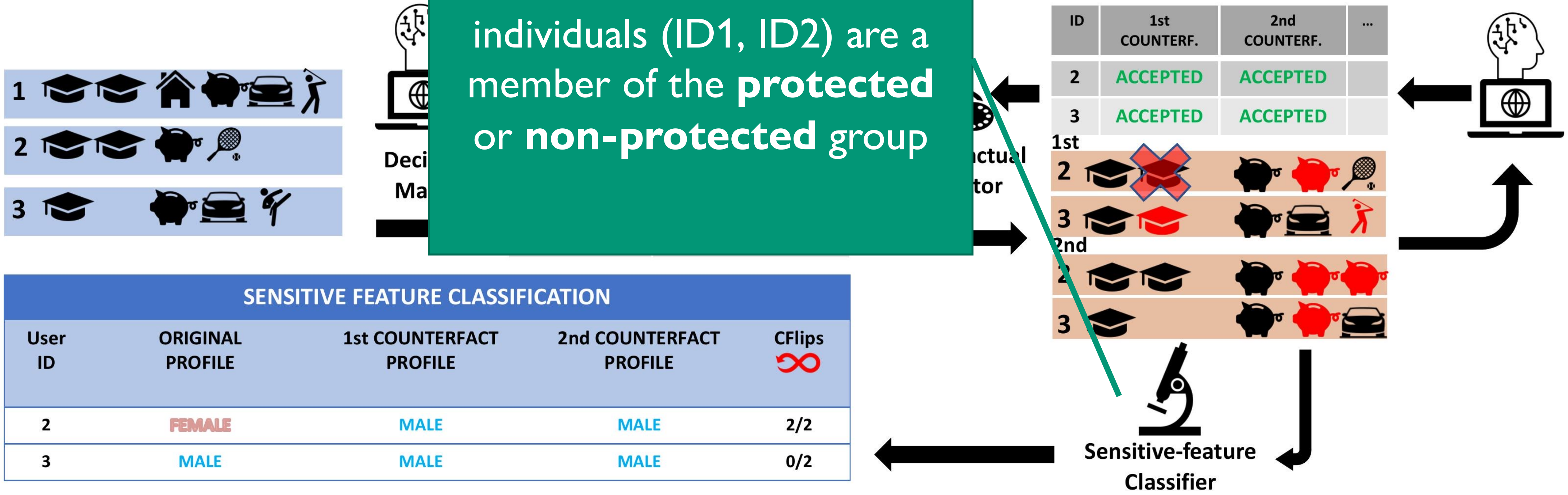
Example: loan application



The CG calculates the **minimal counterfactual** samples, by modifying the values of non-sensitive features, to obtain the desired outcome (e.g., loan approved).

Example: loan application

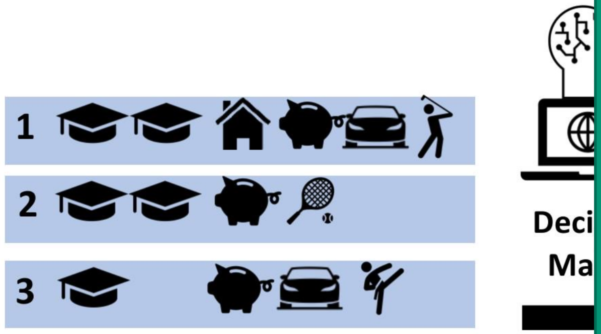
The SfC classifies if the individuals (ID1, ID2) are a member of the **protected** or **non-protected** group

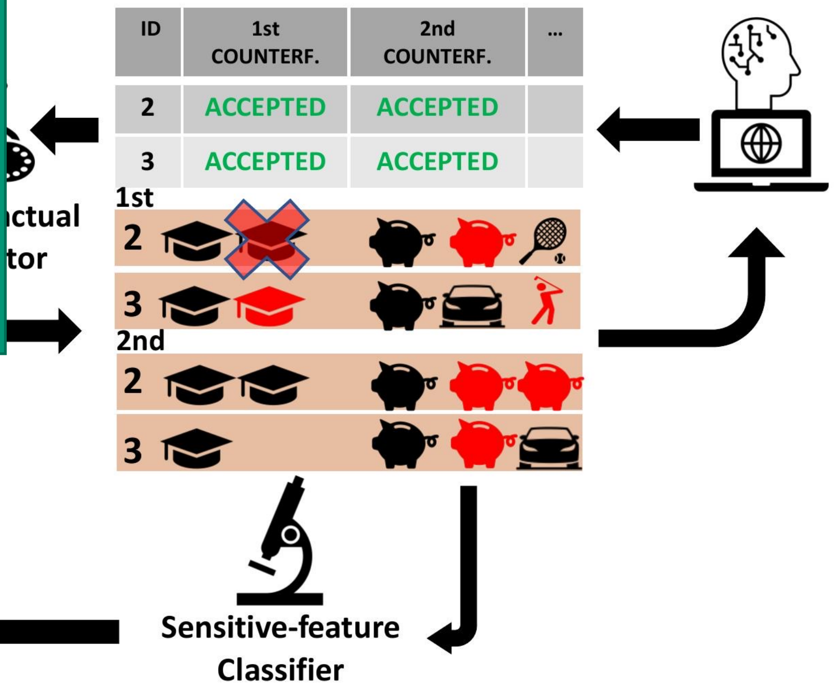


SENSITIVE FEATURE CLASSIFICATION				
User ID	ORIGINAL PROFILE	1st COUNTERFACT PROFILE	2nd COUNTERFACT PROFILE	CFlips
2	FEMALE	MALE	MALE	2/2
3	MALE	MALE	MALE	0/2

Example: loan application

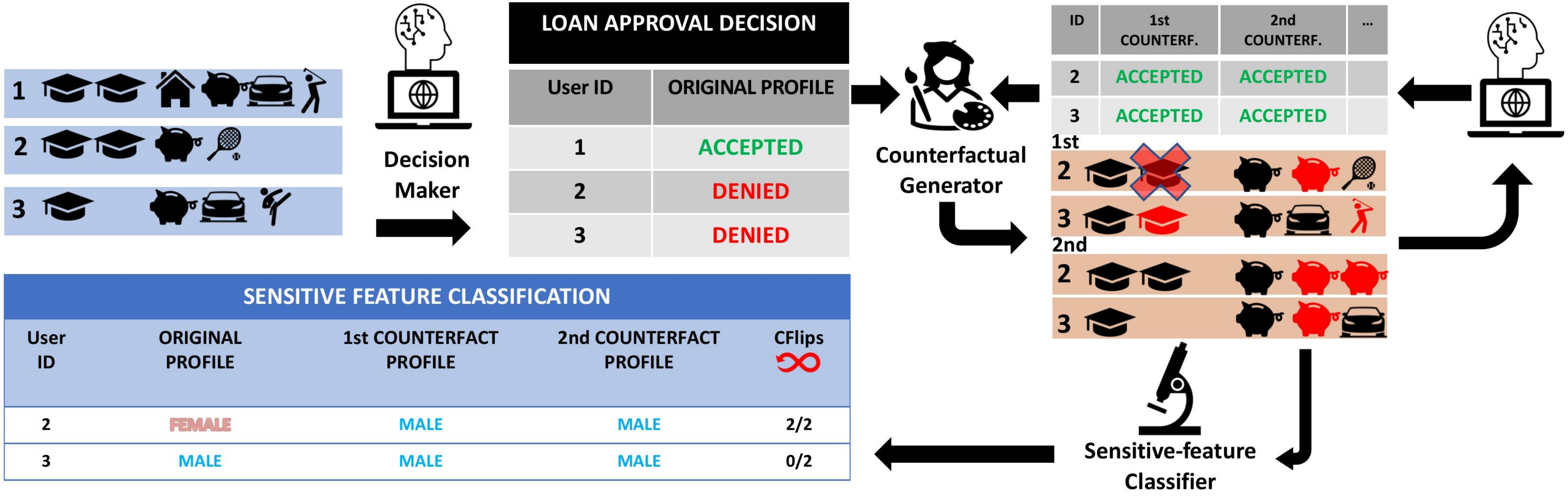
The SfC shows if the new counterfactual profile obtaining the loan is classified now as male, (opposite to the original class)





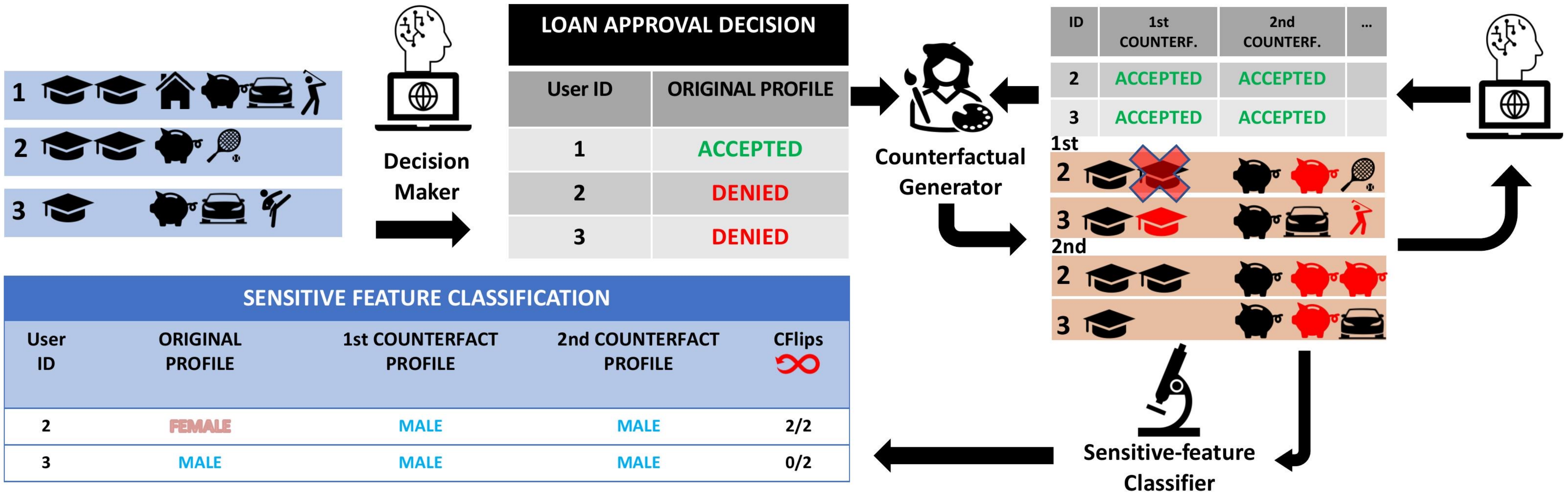
SENSITIVE FEATURE CLASSIFICATION				
User ID	ORIGINAL PROFILE	1st COUNTERFACT PROFILE	2nd COUNTERFACT PROFILE	CFlips
2	FEMALE	MALE	MALE	2/2
3	MALE	MALE	MALE	0/2

Example: loan application



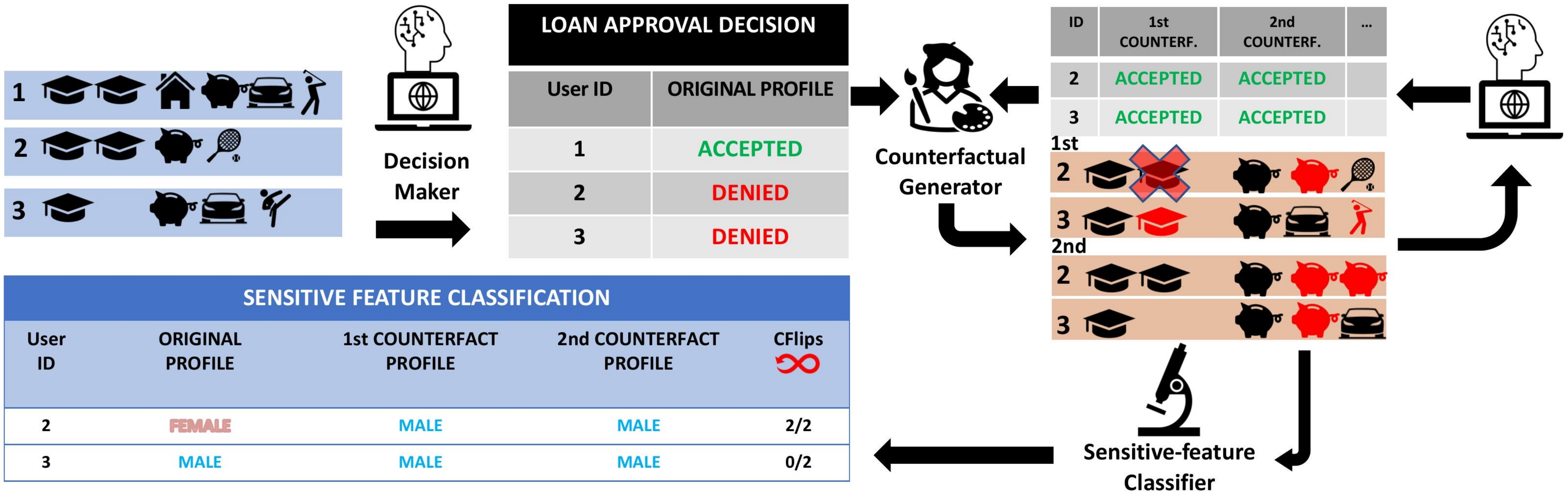
The decision is biased: even though the system does not exploit sensitive features and does not the ID2 gender, it classifies ID2's counterfactual profile (who gets the loan) as belonging to the (privileged) male class.

Example: loan application



To quantify the bias, we compute the number of **Counterfactual Flips**: the number of counterfactual samples belonging to another demographic group

Example: loan application



LOAN APPROVAL DECISION	
User ID	ORIGINAL PROFILE
1	ACCEPTED
2	DENIED
3	DENIED

ID	1st COUNTERF.	2nd COUNTERF.	...
2	ACCEPTED	ACCEPTED	
3	ACCEPTED	ACCEPTED	

SENSITIVE FEATURE CLASSIFICATION				
User ID	ORIGINAL PROFILE	1st COUNTERFACT PROFILE	2nd COUNTERFACT PROFILE	CFlips
2	FEMALE	MALE	MALE	2/2
3	MALE	MALE	MALE	0/2

IDEA: The bigger the CFlips value is,
the stronger the biases and the discrimination the model suffers from

Datasets

Dataset	s	privileged (s^+)
Adult	*gender* *maritalStatus*	*male* *married*
Adult-deb.	*gender* *maritalStatus*	*male* *married*
Crime	*race*	*white*
German	*gender* *age*	*male* > 25 *year*

Adult(*): dataset used for income prediction

German: dataset for default prediction

Crime: dataset for violent states prediction

Decision Makers

We used **seven** largely adopted **learning models** to handle the classification task:

- Logistic Regression (LR), Decision Tree (DT), Support-Vector Machines (SVM), LightGBM (LGBM), XGBoost (XGB), Random Forest (RF), and Multi-Layer Perceptron (MLP).

Plus, **three** in-processing **debiasing algorithms:**

- Linear Fair Empirical Risk Minimization (LFERM), Adversarial Debiasing (Adv), and Fair Classification (FairC).

Counterfactual Generator

For the sake of reproducibility and reliability, the counterfactuals are generated by a third-party counter-factual framework: **DiCE**, an open-source framework developed by Microsoft.

DiCE not only offers several strategies for generating counterfactual samples but also is a **model-agnostic** approach.

Sensitive feature classifier

We exploited **three learning models** (RF, MLP, and XGB) for implementing this component.

Research Question Q1

RQ1: Is there a method for determining whether a dataset contains proxy features or not?

How well the sensitive-feature classifier can identify if a subject **belongs** to the **privileged** or **unprivileged** group, without exploiting sensitive features in the training phase.

Research Question Q1

RQ1: Is there a method for determining whether a dataset contains proxy features or not?

Results show that, due to proxy features, it is possible to learn a classifier able to predict sensitive characteristics.
Even when only low correlated features with the sensitive information are available (i.e., Adult-debiased)

Research Question Q2

RQ2: Does the Fairness Under Unawareness setting ensure that decision biases are avoided?

Fairness is evaluated computing the Difference in Equal Opportunity (DEO). Removing the sensitive information (i.e., gender and race) do not improve model equity.

Research Question Q2

RQ2: Does the Fairness Under Unawareness setting ensure that decision biases are avoided?

The classifiers seem to be affected by discrimination even when the sensitive information is omitted (since the model can implicitly learn them). Accordingly, imposing Fairness Under Unawareness setting is not sufficient to avoid decision biases and discrimination.

Research Question Q2

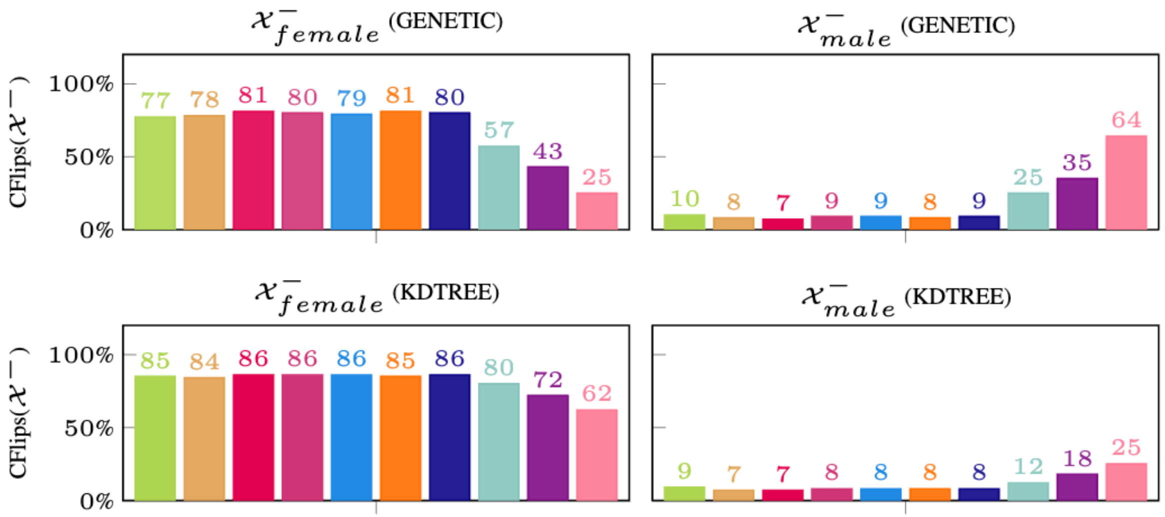
RQ2: Does the Fairness Under Unawareness setting ensure that decision biases are avoided?

For the Adult-debiased dataset some degree of discrimination is still present due to non-linear proxy features

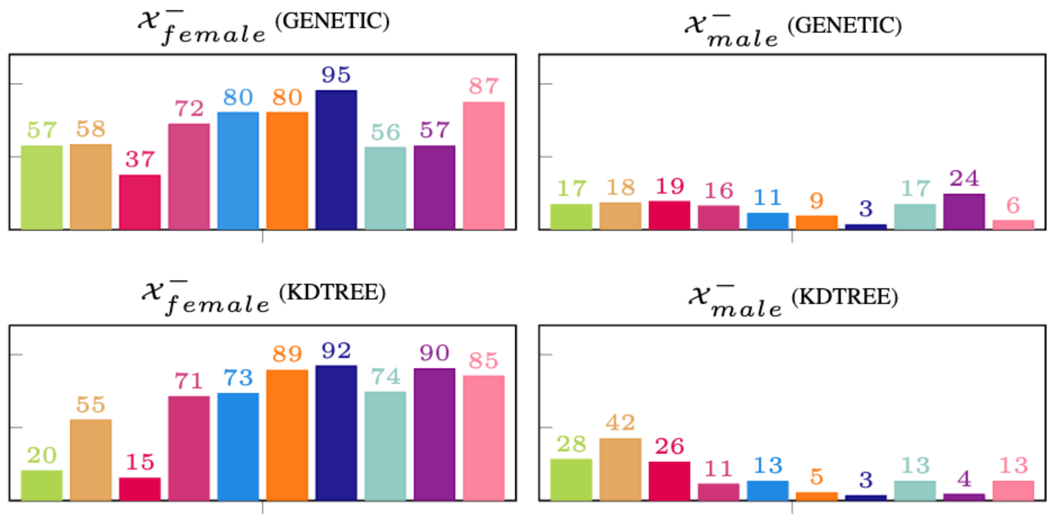
Research Question Q3

RQ3: Is counterfactual reasoning effective for discovering decision biases?

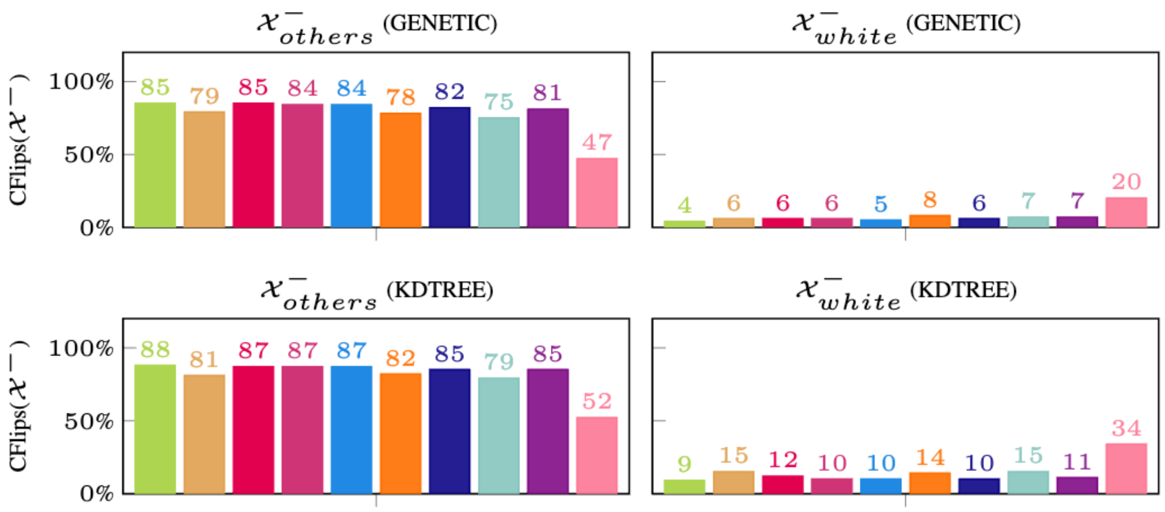
*The metric we used tells us how frequently **a change in the decision** (from negative to positive) for a sample is followed by a **change in the sensitive-feature** classification (e.g., from female to male and vice versa)*



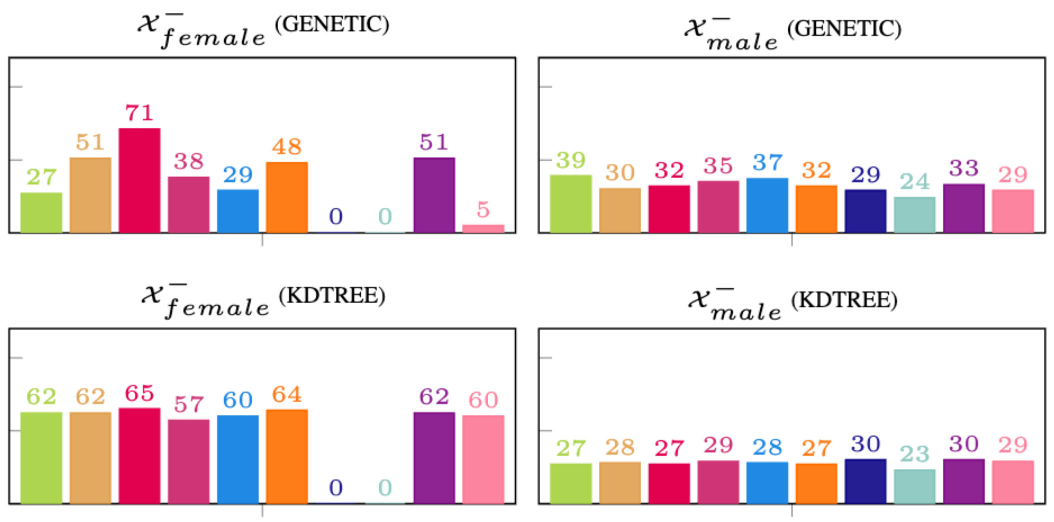
(a) CFlips for the Adult dataset



(b) CFlips for the Adult-debiased dataset



(c) CFlips for the Crime dataset



(d) CFlips for the German dataset

LR DT SVM LGBM XGB RF MLP LFERM ADV FairC

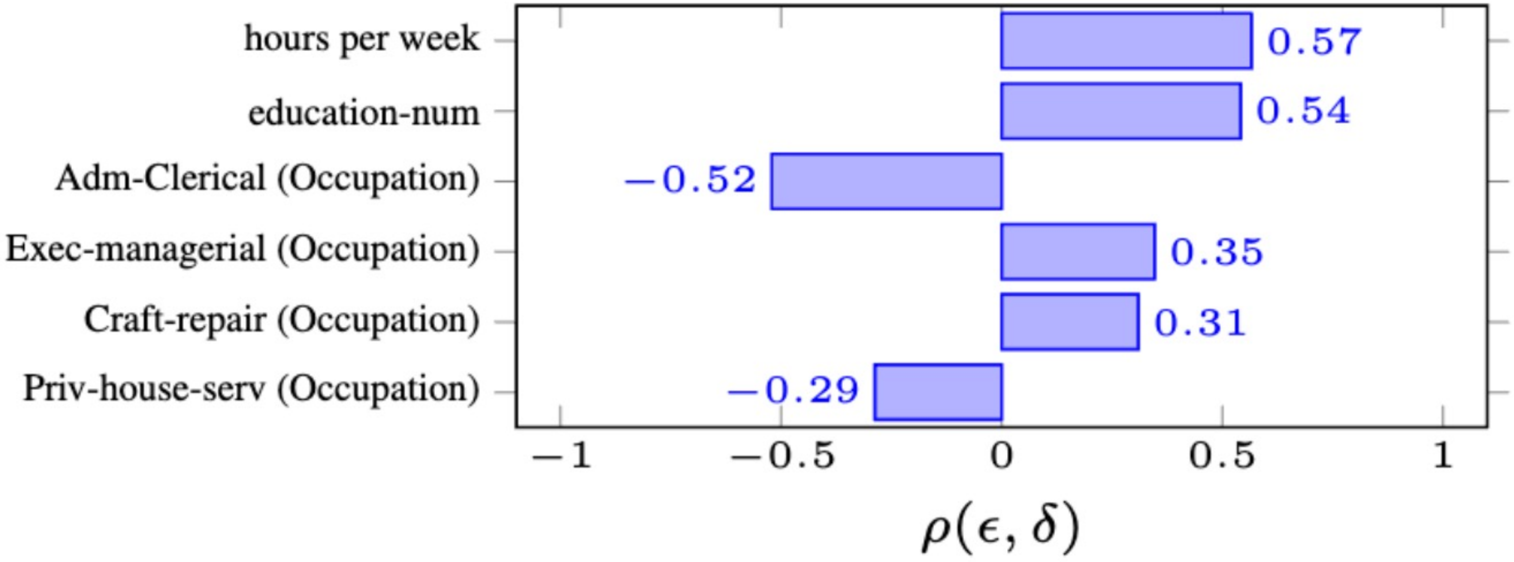
Research Question Q3

RQ3: Is counterfactual reasoning effective for discovering decision biases?

*In the plots emerges that the **unprivileged samples**, to achieve favorable decisions, must take on the characteristics of privileged samples. The results demonstrate that **counterfactual reasoning** effectively **discovers decision biases** and complements SOTA fairness metrics*

Research Question Q4

RQ4: Is it possible to define a strategy for identifying the proxy features?



Contributions

we demonstrate that fairness under unawareness assumption is **not sufficient to mitigate bias**

we propose a **methodology** for the **bias auditing** task

we show that counterfactual **reasoning** is an effective methodology to unveil the bias

we define a procedure to **identify proxy features** leveraging counterfactual reasoning

Bibliography

1. **Auditing fairness under unawareness through counterfactual reasoning**. Giandomenico Cornacchia, Vito Walter Anelli, Giovanni Maria Biancofiore, Fedelucio Narducci, Claudio Pomo, Azzurra Ragone, Eugenio Di Sciascio. Inf. Process. Manag. 60(2): 103224 (2023)
2. **Counterfactual Reasoning for Decision Model Fairness Assessment**. Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, Azzurra Ragone, Eugenio Di Sciascio. The Web Conference (WWW) 2023: 229-233

COUNTERFACTUAL REASONING FOR RESPONSIBLE AI ASSESSMENT

Giandomenico Cornacchia, Vito Walter Anelli, Fedelucio Narducci, **Azzurra Ragone** and Eugenio Di Sciascio