Future Generation Computer Systems 155 (2024) 384–401

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making

Rubén González-Sendino, Emilio Serrano [*], Javier Bajo

Ontology Engineering Group, ETSI Informáticos, Universidad Politécnica de Madrid, 28660 Madrid, Spain

ARTICLE INFO

Keywords:
Causal model
Bias mitigation
Fairness
Responsible artificial intelligence
Bayes

ABSTRACT

In the evolving field of Artificial Intelligence, concerns have arisen about the opacity of certain models and their potential biases. This study aims to improve fairness and explainability in AI decision making. Existing bias mitigation strategies are classified as pre-training, training, and post-training approaches. This paper proposes a novel technique to create a mitigated bias dataset. This is achieved using a mitigated causal model that adjusts cause-and-effect relationships and probabilities within a Bayesian network. Contributions of this work include (1) the introduction of a novel mitigation training algorithm for causal model; (2) a pioneering pretraining methodology for producing a fair dataset for Artificial Intelligence model training; (3) the diligent maintenance of sensitive features in the dataset, ensuring that these vital attributes are not overlooked during analysis and model training; (4) the enhancement of explainability and transparency around biases; and finally (5) the development of an interactive demonstration that vividly displays experimental results and provides the code for facilitating replication of the work.

1. Introduction

Artificial intelligence (AI) can potentially transform our world, but it can also perpetuate societal inequities if not properly designed. The opacity of the models can obscure the reasoning behind the decisions and unfairly impact people. In particular, cognitive bias introduces discrimination in the data set, extending from data generation to model deployment [1]. The data utilization process is a critical issue because characteristics are typically selected based on correlation, ignoring the fact that correlation does not imply causality, although causality indicates correlation [2].

Exploiting sensitive features can result in disparate impacts, leading to discrimination when privileged and underprivileged groups are treated unequally [2]. Therefore, the objective of fairness is to develop an algorithm that satisfies fairness and performance metrics [3]. It is generally acknowledged that fairness disparities can be addressed at the individual or group level, and bias mitigation can occur during pre-training, training, and post-training. [4,5].

This study aims to create a discrimination-free model that makes decisions without affecting personal characteristics. Thus, the dataset used to train the algorithm should be fair and unbiased. To achieve this, conventional pretraining algorithms typically transform existing samples or adjust the weights of instances directly connected to the label. Two primary limitations have been identified in the existing literature: (1) the difficulty of mitigating more than one biased variable simultaneously; and (2) the tendency of mitigation efforts to focus solely on the label, often disregarding potential correlations with other variables [1,6,7].

The algorithm presented is a novel technique to generate fair and discrimination-free datasets based on a causal model. This algorithm addresses the limitations explained above, allowing the mitigation of multiple variables simultaneously and their relationships with other features.

The Fairness Learning process proposed in this paper is categorized into training and pre-training techniques to mitigate bias in structured data. During training, the biases learned by a causal model are mitigated. The algorithm modifies relationships and alters probabilities to ensure a fair impact among selected groups. Mitigation can be implemented by considering one or more sensitive features simultaneously. In the pre-training stage, a fair dataset is generated using the mitigated causal model. This dataset is used to train the model and ensure that the AI algorithm does not perpetuate discrimination by being trained without bias.

The use of a causal model in the proposed approach is motivated by its ability to represent causality and its inherent interpretability. This improves the understanding of the results [8]. In addition, these models are adaptable, permitting modifications to relations and probabilities. Finally, Bayesian models learn the data distribution, giving them the characteristic to generate data [9].

[*] Corresponding author.
E-mail addresses: ruben.gonzalez.sendino@alumnos.upm.es (R. González-Sendino), emilio.serrano@upm.es (E. Serrano).

https://doi.org/10.1016/j.future.2024.02.023
Received 7 September 2023; Received in revised form 21 February 2024; Accepted 23 February 2024
Available online 24 February 2024

Experimental findings suggest that the mitigated causal models in the proposed approach can yield equal results. Moreover, these outcomes are devoid of bias when the model is trained using the fairly generated dataset, as they are not influenced by sensitive features. This enables the use of the derived dataset to train more appropriate algorithms for the problem at hand.

Sensitive features are preserved throughout the entire process of the presented approach. This: (1) improves auditing capabilities and understanding of their relationships with other attributes; (2) ensures their inclusion for analysis when new features are incorporated; and (3) facilitates the creation of a fair dataset that includes these features without influencing the decisions.

This work addresses a fundamental issue in data utilization: inherent biases that can lead to discrimination against specific groups, culminating in the establishment of privileged and underprivileged classes. The significant repercussions of data bias extend beyond fairness and discrimination concerns, impacting various use cases. In any engineering system or process that relies on data for decision making, biases can distort the results, leading to inefficiencies, unfairness, or even harm [10]. Consequently, our bias mitigation technique has the potential to benefit engineers in diverse fields, helping them achieve more accurate and equitable data-driven mitigation.

An implementation of the approach detailed in this study, along with the code used to generate our experimental results, is publicly accessible. Researchers interested in reproducing, using, or expanding our work can find these resources in the Fairness Learning Artificial Intelligence (FLAI) library.[1] Additionally, a demo[2] has been created to interact with the generated graph (original and mitigated) and visualize the experimental results.

The paper is structured as follows: Section 2 provides related work, while Section 3 includes the necessary background to understand the presented approach, including AI Fairness and Causal Models. Section 4 delves into the algorithm proposed to mitigate bias and its three main phases: mitigating relations, mitigating conditional probability distribution, and fair data generation. Section 5 presents the experimental results of the evaluation of the approach, and Section 6 discusses them. Section 7 concludes and provides an overview of future work. Additionally, Appendices have been included to explain the demonstration and use of the Python library, along with additional results.

2. Related work

The works that try to increase fairness are those that mitigate bias. The mitigation process can be divided into three steps: pre-training, training, and post-training. The most practical steps are pretraining and training; the last option should be post-training [11]. The bias reduction methods presented in this section are applicable to structured data and align with the approach with which this work contributes.

In the pre-training phase, the objective is to enhance the dataset to mitigate its inherent biases. Prominent techniques include resampling, fair representation, optimize preprocessing, and reweighting. The approach presented in this study will be evaluated in comparison with following methodologies:

- Re-weighting is more widely used to transform data by modifying the weights in the dataset [7,12]. It involves assigning lower weights to instances from a privileged group that is more likely to have a favorable outcome. In comparison, instances of an underprivileged group receive higher weights [13].
- Resampling is used to change the size of the dataset, which affects the distribution without transforming the data [11,14]. Re-sampling methods are divided into under-sampling and oversampling. Various algorithms for fairness mitigation are tested for data augmentation, while techniques for undersampling are less popular [15,16]. The two dominant approaches to oversampling are the "Synthetic Minority Oversampling Technique" (SMOTE) and "Generative Adversarial Networks" (GANs) [14,17].
- Learning Fair Representation (LFR) is a popular algorithm for finding a latent representation that encodes data while maintaining fairness. However, this encoding complicates the explanation [13].
- Optimized Pre-Processing (OPP) makes use of instance reweighting to adjust training sample weights based on differences in feature distribution between privileged and underprivileged groups. The method significantly improves fairness [18].

Secondly, the training step aims to reduce the discrimination that the model could learn from the data. Popular techniques at this stage include Regularization and Adversarial Training, with multiple approaches emerging.

- Regularization is a well-known technique in machine learning. It is used to correct under- or over-fitting when training the model [12]. For example, regularization methods in the loss function of deep neural networks can help reduce the difference in the prediction disparity between different groups [19]. Regularization can also penalize high correlations between sensitive attributes and results. However, the addition of regularization methods to a machine learning model can complicate the explanation and interpretation of its results [13].
- Adversarial de-biasing involves training two neural networks, where one network learns to predict the outcome, and the other network identifies and removes any biases in the training data that could affect the prediction of the first network. The second network, known as the "adversary", attempts to find and exploit weaknesses in the first predictions. Scores between different demographic groups can be balanced, promoting demographic parity, equality of chances, and equality of opportunity [20,21].

Finally, the last option is to correct the discrimination learned by modifying the output. The three algorithms used in this step are:

- *Equalized odds* add a post-learning step to determine the optimal probabilities of changing the output labels. Equalized odds enforce fairness and precision [5,7,12–14,22].
- *Calibrated equalized odds* optimize the probabilities to change the output with an equalized odds objective, starting from the scoring output of a calibrated classifier [5,7,12,14,22].
- *Reject option classification* provides favorable outcomes for protected underprivileged groups and unfavorable outcomes for privileged ones. This method uses a confidence band around the decision boundary with the highest uncertainty [5,7,12–14,22].

The techniques presented in this section have been demonstrated using only one single sensitive variable. This represents a limitation, as datasets may contain multiple sensitive variables, as is the case with the examples used in this paper. Consequently, this may not allow for a comprehensive mitigation and correction of fairness across all characteristics. Additionally, these algorithms focus on mitigating bias regarding model labels or outputs, which poses a limitation to achieving complete mitigation [23].

In conclusion, mitigating algorithm bias requires careful consideration. Various techniques can be used at each stage to address discrimination, improve fairness, and maintain the accuracy of the model. It is crucial to balance these factors while ensuring that the model results are realistic. Using the right combination of techniques, it is possible to create more equitable and fair machine learning models that can be reliably used across diverse applications.

[1] https://github.com/rugonzs/FLAI

[2] https://www.rubengonzalez.ai/demo

3. Background

This section aims to cover the concepts necessary to understand the remainder of the paper. Section 3.1 on fairness explains the key metrics used to assess fairness. Section 3.2 on causal models describes the process of creating a causal graph and its use to generate samples. Finally, Section 3.3 explains the importance of explainability and how causal models help with it.

3.1. AI fairness

Data used to train AI models often contain biases that can lead to discrimination. The main point regarding this aspect is that biases can be inherited or introduced [19]. Three groups could be identified when relating bias to the life-cycle model: data bias, learning bias, and deployment bias [1]. The metrics commonly used to measure fairness are associated with the comparison of privileged (PG) and underprivileged (UG), but there are also metrics to compare individuals, although they are less popular [11]. The metrics have a valid range that has previously been defined in the literature [5].

These fairness metrics operate by juxtaposing two distinct groups. Converting sensitive features into binary variables is a prevalent approach, aligning various categories within these features with privileged or underprivileged groups [5]. In addition, an automated algorithm is available to identify the privileged group, enabling a comprehensive comparison among all groups, and selecting the optimal group as a reference point [24]. The most common metrics used to measure fairness are:

- Equal Opportunity Difference [13]. Measures the difference in true positive rates (TPR) between the underprivileged and privileged groups. The ideal value is 0; in this study, the interval between −0.1 and 0.1 will be fair.

$$\mathrm{TPR} = \Big[\frac{\text{True Positive (TP)}}{\text{TP} + \text{False Negative (FN)}}\Big] \tag{1}$$

$$\mathrm{EOD} = \mathrm{TPR}_{\mathrm{UG}} - \mathrm{TPR}_{\mathrm{PG}} \tag{2}$$

- Odds Difference [25]. Calculates the difference between the underprivileged and privileged groups in false positive rates (FPR) and true positive rates (TPR). The ideal value is 0; in this study, the interval between −0.1 and 0.1 will be fair.

$$\mathrm{FPR} = \Big[\frac{\text{False Positive (FP)}}{\text{FP} + \text{True Negative (TN)}}\Big] \tag{3}$$

$$\mathrm{OD} = \big(\mathrm{FPR}_{\mathrm{UG}} - \mathrm{FPR}_{\mathrm{PG}}\big) + \big(\mathrm{TPR}_{\mathrm{UG}} - \mathrm{TPR}_{\mathrm{PG}}\big) \tag{4}$$

- Statistical Parity Difference [16]. It calculates the difference in the probability of favorable results (Predicted as Positive (PPP)) between the underprivileged and privileged groups. The ideal value is 0; in this study, the interval between −0.1 and 0.1 will be fair.

$$\mathrm{PPP} = \Big[\frac{\text{TP} + \text{FP}}{\text{Total Population (N)}}\Big] \tag{5}$$

$$\mathrm{SPD} = \mathrm{PPP}_{\mathrm{UG}} - \mathrm{PPP}_{\mathrm{PG}} \tag{6}$$

- Disparate Impact [22]. Compares the proportion of individuals who receive a positive output for two groups: Underprivileged and privileged. The ideal value is 1. In this study, the interval between 0.8 and 1.2 will be fair.

$$\mathrm{DI} = \frac{\mathrm{PPP}_{\mathrm{UG}}}{\mathrm{PPP}_{\mathrm{PG}}} \tag{7}$$

- Theil Index [26]. Subclass of the generalized entropy index (using alpha = 1). The entropy index measures the inequality in a group or individual concerning the fairness of the algorithm outcome. The ideal value is 0; in this study, values lower than 0.1 will be fair.

$$\mathrm{TI} = \frac{1}{N}\sum_{i=1}^{N}\frac{b_i}{\mu}\ln\frac{b_i}{\mu}, b = predicted - labeled \tag{8}$$

Previous metrics are collected in different tools such as AIF360 [5], FairLearn [27,28], LFIT [29], Aequitas [24], LimeOut [30], MAML [11], the What-If Toolkit (WIT) [31], Audit AI [32], and Dalex [33]. Libraries established to address this problem are AIF360, Dalex, and Aequitas.

In the field of causal models, the concept of counterfactual fairness can be introduced to assess the fairness of a model. Counterfactual fairness, as defined by Kusner et al. [34], encapsulates the idea that a decision is considered fair for an individual if it remains consistent both in the actual world and in the counterfactual world where the individual is part of a different demographic group, thus ensuring equal treatment.

3.2. Causal models

Causal models are specifically designed to uncover cause-and-effect relationships within datasets, addressing the challenge of determining whether changes in one variable can lead to changes in another. A Causal Bayesian Network (CBN) is a causal model that extends the Bayesian network framework. Bayesian networks themselves are probabilistic graphical models that efficiently encode the joint distribution of a collection of random variables [35]. However, the critical distinction lies in the fact that Causal Bayesian Networks explicitly incorporate assumptions of causality among the variables. Thus, the interpretation of a Directed Acyclic Graph (DAG) in a causal model implies that the parent nodes of a certain variable correspond to its direct causes [36].

Definition 3.1 (*Structural Causal Model*)**.** A Structural Causal Model (SCM) [37–39] is defined as a 4-tuple $M = (X, U, F, P_U)$, where: X is a finite set of endogenous variables, that are determined by other variables in the model; U denotes a finite set of exogenous variables, that are determined by factors outside the model; F is a set of functions $\{f_1, f_2, \ldots, f_n\}$, where each function f_i represents a causal mechanism such that $\forall x_i \in X$, $x_i = f_i(\mathrm{Pa}(x_i), u_i)$, with $\mathrm{Pa}(x_i)$ being a subset of $X \setminus \{x_i\} \cup U$; and, P_U is a probability distribution over U. A SCM is also known as a Structural Equation Model or Functional Causal Model.

Definition 3.2 (*Causal Bayesian Network*)**.** A Causal Bayesian Network (CBN) [38,40] represents an SCM $M = (X, U, F, P_U)$ using a directed graphical model $G = (V, E)$, where: V corresponds to the set of endogenous variables X; E represents the causal mechanisms, indicating that for each causal mechanism $x_i = f_i(\mathrm{Pa}(x_i), u_i)$, there is a directed edge from each node in the parent set $\mathrm{Pa}(x_i)$ to x_i.

The causal modeling process comprises two fundamental steps: constructing the Directed Acyclic Graph (DAG) and calculating the probabilities associated with the relationships represented in the graph. The most common methods for learning structures include:

- The exhaustive search method involves searching for the best DAG by evaluating all possible graphs, considering their structure and data [41]. Due to its computational complexity, this approach is typically limited to small datasets and graphs with few nodes.
- Hill Climb Search is a heuristic search algorithm that starts with an initial DAG and iteratively refines the structure by adding, deleting, or reversing edges to improve the model score [42]. This technique is more computationally efficient than exhaustive search and can be applied to larger datasets.
- Naive Bayes is a simple probabilistic classifier that applies Bayes' theorem with strong independence assumptions between features. Despite its simplicity and the fact that it often does not accurately model feature dependencies, Naive Bayes can perform surprisingly well in many practical applications [43].

- Tree-augmented Naive Bayes (TAN) is an extension of the Naive Bayes algorithm that allows for limited dependencies between features. TAN constructs a tree-structured Bayesian network, which retains the simplicity of Naive Bayes while providing improved performance when feature dependencies exist [44].
- Chow-Liu algorithm constructs a tree-structured Bayesian network by maximizing mutual information between variables [45]. It provides a simple and efficient method for finding the best tree structure for a given dataset, though it may not capture more complex relationships between variables
- Constrain Based methods to identify the structure of a causal model by examining conditional independence relationships between variables. These methods use statistical tests to iteratively identify and add edges to the graph based on the presence or absence of conditional independence relationships [46].
- Score-based methods involve searching for the best network structure by optimizing a scoring function that measures the goodness of fit between the network and the data. These methods typically use search algorithms, such as genetic algorithms or simulated annealing, to efficiently explore the space of possible network structures [47].

After the creation of the DAG, it could be evaluated; the two options are: (1) using a score of Bayesian scoring functions: The Bayesian Dirichlet equivalent uniform (BDeu) [48] and K2 [42]; or (2) information-theoretic scoring functions: The Minimum Description Length (MDL) [49], The Bayesian Information Criterion (BIC) [50], The Akaike Information Criterion (AIC) [51], The Normalized Maximum Likelihood (NML) [52] and The Minimum Information Theoretic (MIT) [53]. These scoring functions are used to assess the quality of network structures in terms of their ability to fit the data and represent the underlying causal relationships. Bayesian scoring functions assess network structures using a combination of the likelihood of the data given the structure and a prior distribution over the structures. On the other hand, information-theoretic scoring functions aim to minimize the amount of information needed to describe data and network structures, which helps to find models that balance complexity and accuracy in data fitting [54].

The second step is the probability calculation. This process estimates the values of Conditional Probability Distributions (CPDs) in a Bayesian network. This estimation is based on the observed data and is essential for understanding the relationships between variables. Two widely used methods for parameter learning are Maximum Likelihood Estimation (MLE) and Bayesian Estimation [55].

MLE calculates CPDs using the relative frequencies of variable states observed in the data. However, MLE can overfit the data, especially when the sample size is small or the data are not representative of the underlying distribution. Bayesian Parameter Estimation mitigates the overfitting issue of MLE by incorporating previous CPDs, representing initial beliefs about variables before observing the data [42,47].

Bayesian networks are good generative algorithms, this task is called the sampling technique, some popular algorithms are: forward sampling [9], weighted sampling [9], rejection sampling [56], Gibbs Sampling [57,58], Metropolis–Hastings [59,60], and Importance Sampling [61,62]. Each of these techniques has advantages and disadvantages, making the selection of the appropriate method highly dependent on the problem being addressed and the characteristics of the Bayesian network.

Some algorithms may be more accurate or efficient depending on the situation, such as the size and complexity of the network, the presence of evidence and the specific distribution of interest. Comparing these techniques according to accuracy, efficiency, and computational requirements is crucial. Considering these considerations and the problem details will ensure that the most suitable sampling method is chosen for the task at hand.

3.3. *Explainable Artificial Intelligence*

The field of Explainable Artificial Intelligence (XAI) has gained significant importance in artificial intelligence research due to the growing need to understand and explain the decisions made by machine learning models [8]. In this context, several methods have emerged in the field of Explainability in Artificial Intelligence (XAI) to provide interpretations of machine learning models. Among them, two of the most prominent ones are LIME (Local Interpretable Model-agnostic Explanations) [63] and SHAP (SHapley Additive Explanations) [64]. LIME specializes in generating explanations by introducing perturbations into the data, while SHAP values have been designed to allocate the contribution of each input feature to the difference between the prediction of the model and an average reference value.

Causal algorithms are gaining importance as explanatory tools [65]. These algorithms are considered transparent and have the potential to improve the understanding of the underlying causes of each prediction, thereby increasing confidence in these predictions [66]. Integrating counterfactual analysis in explanation provides an additional level of detail by allowing comparisons with imaginary or unobserved situations [38].

Two of the most widely used explanation methods in AI, LIME and SHAP, have been enhanced through the integration of causal models into their frameworks. CALIME, an extension of LIME, enhances the stability and accuracy of explanations by more faithfully replicating the behavior of black-box models [67]. Unlike LIME, which relies on random perturbations, CALIME uses GENCDA to generate synthetic datasets for tabular data [68]. Furthermore, Causal SHAP values augment SHAP by applying do-calculus [69] to distinguish the total causal effect into direct and indirect contributions [70].

Additionally, other algorithms are emerging to explain deep neural networks (DNNs) by using causal inference based on do-calculus. For example, Narendra et al. [71] focus on creating a causal model based on the structure of a DNN, where interventions are used to identify significant elements within the network. In another study [72], interventions are used to generate new images, thus determining the effects of such interventions.

4. Fair causal mitigation

This study aims to decrease the influence of sensitive variables on the results generated by AI algorithms. Mitigation is achieved using a causal model, and the technique can be categorized as pretraining and training mitigation.

Fig. 1 illustrates the architecture sequence, beginning with the creation of a causal model using the methods described in Section 3. Following this, Fair Causal Mitigation takes place, requiring sensitive variables as input. This self-contained process makes the workflow entirely automated and functional.

Fair Causal Mitigation consists of three steps: (1) Mitigate Relations, (2) Mitigate Probabilities, and (3) Generate Fair Data. Execution of the first and second stages can be considered as a causal mitigation training technique. As a result, an equitable classification model is produced. The third stage is a pretraining technique. The generated dataset can be used to train an algorithm.

Bayesian Networks serve as excellent models for representing causes and effects. Their interpretability and transparency contribute to a better understanding of the data. In this study, the causal graph has been inferred using a hill-climb search algorithm, along with the parameters employing a Bayesian Estimator. In contrast to other pre-training methods that may lack transparency, causal models offer the advantage of customization based on rules or expert knowledge to achieve the desired outcome.

The results obtained in this experiment are grounded in the principles of counterfactual fairness, serving as a foundational point. Positive outcomes facilitate a broader application of this mitigation technique to more complex processes.

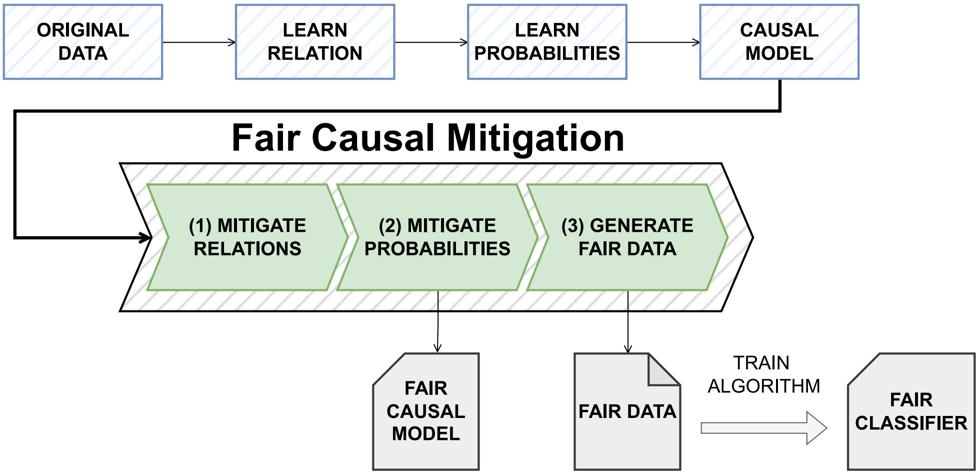
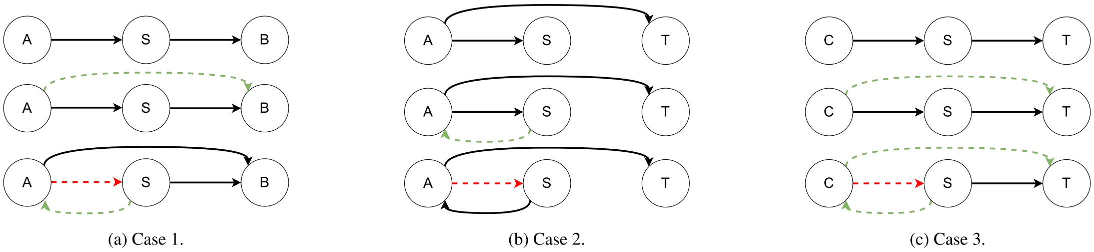


Fig. 1. The figure depicts the flow of the proposed solution. (1) The first step would be to generate a causal model. (2) Apply the mitigation algorithm to the generated model. Finally, the mitigated model and the fair dataset are obtained.



(a) Case 1.

(b) Case 2.

(c) Case 3.

Fig. 2. DAGs are used to explain the mitigation rules. Green edges represent added relations, and red edges represent removed relations.

4.1. Mitigate relation

This step aims to mitigate the influence of sensitive features on relationships while preserving cause-and-effect connections. The efficacy of this process significantly impacts equitable data generation outcomes. Ensuring accurate node relationships is crucial to minimize label influence or bias transmitted through proxy features. This approach might be bypassed if graphs are expertly crafted or if the Naive Bayes algorithm is utilized.

To achieve mitigation, three rules will be employed: transitivity, reversal, and removal. The primary objective is to isolate sensitive variables and prevent other variables from exerting influence on them. Eliminating dependencies on biased variables requires the unaffected status of these variables by any other variable.

In a Bayesian network graph following the relationship $X \rightarrow Y \rightarrow Z$, X indirectly influences Z through Y. Changes in X affect Y, thus affecting Z. Transitivity operates on the conditional probability product rule, as depicted in Eq. (9), illustrating the indirect influence of X on Z.

Taking into account Y as a sensitive characteristic, it is imperative that changes in Y do not directly affect Z; this will be addressed in Section 4.2. Preserving the indirect relationship learned within the inferred graph involves introducing a direct link between X and Z. This ensures consistency in cause-and-effect relationships, aligning with specific problem requirements.

$$P(Z|X) = \sum_{Y} P(Z|Y)P(Y|X) \tag{9}$$

The reversion technique is rooted in maintaining statistical independence between sensitive and other features. Hence, if any feature influences a sensitive one, the relationship is reversed to mitigate influence on the sensitive feature, and vice versa. Reversion is facilitated by the inverse conditional probability equation, as represented in Eq. (10).

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \tag{10}$$

Fig. 2 illustrates three scenarios to elucidate the mitigation rules explained below. Nodes A, B, and C are normal, S are sensitive, and T is the target. Green means that a new relation is added; red means that this relation is removed.

1.1 Transitivity. Variables with a relationship between themselves towards a sensitive variable will automatically impact the nodes with which the sensitive variable is directed. Fig. 2(a) represents the relation: $A \rightarrow S \rightarrow B$ and in Fig. 2(c) represents the relation: $C \rightarrow S \rightarrow T$. Applying transitivity, the new relations will be $A \rightarrow B$ and $C \rightarrow T$, respectively.

1.2 Reversal: After applying transitivity, the relationship from feature to sensitivity needs to be inverted. The resulting reversals are shown in Fig. 2(a) as $S \rightarrow A$, in Fig. 2(b) as $S \rightarrow A$, and in Fig. 2(c) as $S \rightarrow C$".

2. Removal. After transitivity and reversal are applied, the relation between the first feature and the sensitivity should be removed.

The primary objective has been accomplished; no variable now directly impacts the sensitive variable. The relationships involving the sensitive variables have been redirected to their ancestor nodes. Consequently, the graph now encapsulates relationships in which discrimination can be mitigated. Additionally, data generation is going to be based on these relations.

4.2. Mitigate conditional probability distribution

In this phase, the impact is removed and the previous step is necessary to generate a graph predisposed to mitigation. The theory applied in this process will be counterfactual equality. The core objective of this approach is to ensure that distinct groups receive equal treatment such that $P(X|S_{\text{privileged}}) = P(X|S_{\text{Underprivileged}})$. In doing so, the algorithm prevents discrimination based on sensitive features (S). The probability

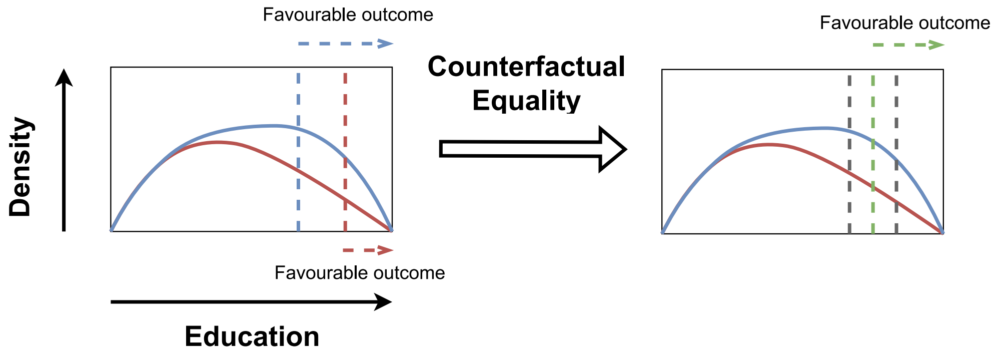


Fig. 3. The figure illustrates two distributions, where the red represents the non-privileged group and the blue represents the privileged group. The X-axis reflects the level of education.

that an instance belongs to a particular group can be denoted as $P(S_{\text{privileged}}) = P(S_{\text{Underprivileged}})$, leading to $P = 0.5$.

Fig. 3 illustrates an example of equality in which the level of education matters whether to be hired or not. The criteria for these decisions might be influenced by cognitive biases, leading to discrimination against less privileged groups [1]. On the left side of Fig. 3, it is evident that the less privileged group (red) requires a higher level of education compared to the privileged group (blue). The concept of equality aims to ensure that this decision boundary is the same for both groups, necessitating the alignment of the boundary at a common point. On the right-hand side, it is shown that both groups will share the same education threshold to achieve a favorable outcome.

The graph structures have been modified in the previous section. Therefore, the conditional distribution of the variables needs to be learned again from the data. This process is executed by Algorithm 1. The objective of this algorithm is to achieve equality for all specified sensitive variables.

Algorithm 1 learns the conditional distribution using Maximum Likelihood or Bayesian parameter estimation. The initial lines of the algorithm prepare the variables to achieve the desired results. An important step is when the algorithm checks if a parent node is included in the list of sensitive features (line 9). In line 13, when no sensitive features are present in the list of parent nodes, the conditional distribution is calculated as usual following Maximum Likelihood or Bayesian parameter estimation.

In contrast, in line 10, when one or more sensitive features are present in the parent nodes, the formulas have been adapted to calculate the conditional distribution in two steps. In the first step, a usual conditional distribution calculation is performed, but sensitive features are excluded. In the second step, the resulting probability is uniformly divided for each of the unique values in each sensitive feature.

To illustrate the previous calculation, consider the following features: the target variable Y, sensitive feature S, and normal feature X, where each feature could be 0 or 1. The result of step 1 excluding S is: $P(Y = 1|X = 1) = 0.6$. The second step calculates the probability uniformly for each sensitive value, resulting in $P(Y = 1|X = 1, S = 1) = 0.3$ and $P(Y = 1|X = 1, S = 0) = 0.3$. This results in complete counterfactual equality, since each value for the sensitive features will have the same probability.

At this stage, the graph results maintain fairness, adhering to the principle that any modification in sensitive attributes does not alter the outcome. The causal algorithm is well-prepared to deduce unbiased results, demonstrating a robust approach to evade discrimination. Moreover, it is equipped proficiently to generate fair data, ensuring that data utilization adheres to equitable principles.

Algorithm 1 Mitigate CPD

```
 1: procedure MitigateCPD($sensitive\_feature$)
 2:     $list\_cpd$ ← empty list
 3:     for each $node$ in $graph.nodes$ do
 4:         Gather $node\_value$ and $evidence$
 5:         Create all $evidence\_combination$
 6:         $list\_probas$ ← empty list
 7:         for each $ec$ in $evidence\_combination$ do
 8:             Filter dataset based on current $ec$
 9:             if $node$ in $sensitive\_feature$ then
10:                 Calculate probabilities uniformly
11:                 Add to $list\_probas$
12:             else
13:                 Calculate probabilities based on $data$
14:                 Add to $list\_probas$
15:             end if
16:         end for
17:         Create a new TabularCPD and append to $list\_cpd$
18:     end for
19:     Update $graph$ with a new DAG based on $list\_cpd$
20: end procedure
```

4.3. Fair data generation

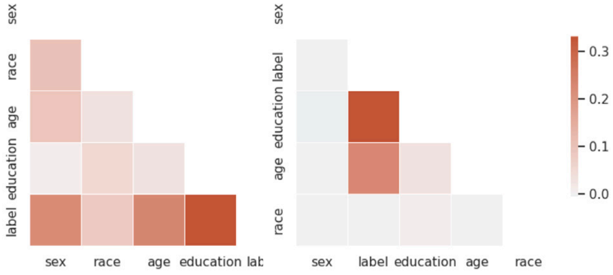
This section generates a fair dataset using the mitigated causal model as a basis. To achieve this, a sampling technique based on the joint distribution of the Bayesian Network is employed [73]. The precision of the resultant dataset correlates directly with the number of generated samples; as the number of samples approaches infinity, the probabilities converge towards their true values. Ensuring the most accurate results, generating a sufficiently large number of samples is essential while considering the trade-off between accuracy and computational efficiency.

The forward sampling method is used to create a fair dataset in this study [9]. The forward sampling algorithm 2 begins with an empty sample. The network variables are then ordered in a topological sequence, ensuring that each variable is considered only after all its dependencies have been evaluated. If the variable X_i does not have ancestors, its value is sampled from the marginal distribution of the variable (see line 5). If the variable X_i has ancestors, its value is sampled from a conditional distribution using values already determined for its ancestors (see line 7). This process is repeated for all variables to generate a complete sample of the joint distribution. Repeating this process allows for an empirical representation of the joint distribution of all variables in the Bayesian Network.

Fig. 4 shows the correlation between the characteristics of the original and mitigated dataset. The label correlates with the four features in the original dataset. In the mitigated data, the label is significantly correlated only with age and education. Furthermore, sex and race no longer correlate with the label and the other variables (Education and Age).

Algorithm 2 Forward Sampling for Bayesian Networks

```
Initialize empty sample $S \leftarrow \{\}$
Order $X_1, X_2, \dots, X_n$ in topological order
for each variable $X_i$ in $X_1, X_2, \dots, X_n$ do
    if $X_i$ has no ancestors then
        Sample value $x_i$ from $P(X_i)$
    else
        Sample value $x_i$ from $P(X_i|\text{Ancestors}(X_i))$
        using values of Ancestors($X_i$) in $S$
    end if
    Add $x_i$ to $S$
end for
Return $S$
```



(a) Original correlation.

(b) Mitigated correlation.

Fig. 4. Compare the correlation between features before and after the mitigation process. Gray values indicate low correlation, while dark orange indicates high correlation.

This could be used as a first indicator to validate that the impact of the features on the fair data set has been mitigated. The following indicators will be explained in Section 5, showing the results of the fairness metrics and important characteristics.

5. Results

This section covers the examination of the results obtained in this research. Three analyses will be performed: (1) Accuracy and fairness metrics for the pre-training and training mitigation technique in Section 5.1; (2) Importance evaluation of sensitive features using shap value in Section 5.2; (3) Compare the results with other pre-training metrics to evaluate the fair data generated in Section 5.3.

To understand the results, the following information indicates the meaning of each model used in the evaluation.

- Training mitigation
    1. Causal Model (CM). An original model without mitigation.
    2. Causal Model Mitigated (CMM). Causal model after applying the mitigation techniques in Sections 4.1 and 4.2.
- Pretraining mitigation
    1. XGBoost (XGB). The algorithm was trained using the original dataset.
    2. XGBoost Mitigated (XGBM). The algorithm was trained with the fair data set generated in Section 4.3 for the training.

In this section, the data used to evaluate the models for accuracy and fairness metrics will be Original Dataset (OD) and Fair Dataset (FD). The original data set was the data set without transformations, and the fair data set is the data set generated in Section 4.3.

To guarantee the applicability of this algorithm, the results will be evaluated with respect to both fairness and performance. Fairness will be evaluated through metrics such as Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). Additionally, the performance of the model will be measured by accuracy, True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and Predicted as Positive (PPP).

The dataset tested in this work has previously been used in the literature and presents biases. These data sets have been audited and both the privileged and the underprivileged groups have been identified. The three datasets used will be Compas [74,75], German [76], and Adult [77]. In the Adult data set, the label = 1 indicates a salary greater than $50K/year. The sensitive features are sex and race. The variables education and age (which could be related to experience) are usually considered to influence the salary. The results for other datasets are available in Appendix B.

5.1. Bias mitigation evaluation

This section will evaluate the results of the mitigation techniques presented. Performance metrics seek to ensure the usability of the model and fairness metrics aim to ensure the freedom of discrimination in the results. The objective is to maintain or improve the original accuracy and increase the fairness metrics. Fair will be considered if the values fall within the acceptable range defined by Bellamy et al. [5]. The ideal values for the fairness metrics were discussed in Section 3.

The results are in Tables 1 and 2. The first collects the results according to the performance, and the second is the evaluation of the fairness metrics. The models are described in Section 5. The dataset determines which data are used to evaluate the algorithm: Original (OD) or Fair Data (FD).

The causal mitigation technique (CGM) gives acceptable results. In terms of performance, the metrics are almost balanced between both groups and maintain the results that no mitigated model provides (CG). The Fairness metrics show good results and reduce the disparities obtained by the original model (CG).

The Fair Data evaluation also evaluates the causal mitigation cause the fair data is generated through this model. Therefore, in some manner, this also validates the mitigated causal. The results show that Fair Data gives perfect results in terms of fairness and also balances the performance between both groups. This indicates that it is possible to generate Fair Data without discrimination to train an algorithm or use it as a reference.

The tables show that both mitigation techniques adapt perfectly when tested by original data and Fair Data. On the contrary, when algorithms learn bias, the results give the same deficient metrics for both datasets. This complete mitigation pipeline ends in fair data that are perfectly usable to allow a model to learn objective patterns to make free-of-discrimination decisions.

5.2. Importance evaluation

To complete the study of the mitigation technique, the final examination focuses on the importance of the feature in the decisions. The importance will be illustrated using the SHAP library [64]. The objective is to analyze SHAP values to identify if sensitive features influence the decision making of the model. The libraries Lime [63] and Dalex [33] have been tested and the results obtained reflect the same conclusions.

Figs. 5(a) and 5(b) show the importance of the model trained with the original data, and the importance is tested for the original and fair data, respectively. Figs. 5(c) and 5(d) show the importance of the model mitigated trained with fair data, tested the importance for original and fair data, respectively.

The results indicate that no matter the data used to test the algorithm. If the model learns discrimination in the training phase, this will not be avoided, although you test with fair data. This is similar to the previous results in Section 5.1. The model that was trained without

Table 1
Evaluation results for each model using performance metrics: Accuracy (A), True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and Predicted as Positive (PPP). The dataset used is the Original and Fair Data generated by mitigated Causal Model. Best accuracy values highlighted in bold.

Algorithm	Dataset	Feature	Group	A	TPR	FPR	FNR	PPP
Causal Model	Original	Sex	Privilege	76%	0.46	0.11	0.54	0.21
		Sex	Underprivileged	**89%**	0.0	0.0	1.0	0.0
		Race	Privilege	**80%**	0.41	0.07	0.6	0.16
		Race	Underprivileged	**86%**	0.24	0.03	0.76	0.06
Causal Model	Fair Data	Sex	Privilege	**79%**	0.46	0.11	0.54	0.19
		Sex	Underprivileged	76%	0.0	0.0	1.0	0.0
		Race	Privilege	78%	0.32	0.08	0.68	0.14
		Race	Underprivileged	78%	0.21	0.05	0.79	0.09
Causal Model Mitigated	Original	Sex	Privilege	76%	0.46	0.11	0.54	0.21
		Sex	Underprivileged	84%	0.47	0.11	0.53	0.15
		Race	Privilege	78%	0.46	0.11	0.54	0.20
		Race	Underprivileged	83%	0.47	0.10	0.53	0.16
Causal Model Mitigated	Fair Data	Sex	Privilege	**79%**	0.47	0.11	0.53	0.19
		Sex	Underprivileged	79%	0.46	0.11	0.53	0.19
		Race	Privilege	79%	0.47	0.11	0.53	0.19
		Race	Underprivileged	79%	0.47	0.11	0.53	0.19
XGBoost	Original	Sex	Privilege	76%	0.46	0.11	0.54	0.21
		Sex	Underprivileged	**89%**	0.0	0.0	1.0	0.0
		Race	Privilege	**80%**	0.41	0.07	0.6	0.16
		Race	Underprivileged	**86%**	0.24	0.03	0.76	0.06
XGBoost	Fair Data	Sex	Privilege	**79%**	0.46	0.11	0.54	0.19
		Sex	Underprivileged	76%	0.0	0.0	1.0	0.0
		Race	Privilege	78%	0.32	0.07	0.68	0.14
		Race	Underprivileged	77%	0.21	0.05	0.79	0.08
XGBoost Mitigated	Original	Sex	Privilege	76%	0.46	0.11	0.54	0.22
		Sex	Underprivileged	84%	0.47	0.11	0.53	0.15
		Race	Privilege	78%	0.46	0.11	0.54	0.2
		Race	Underprivileged	83%	0.48	0.11	0.52	0.17
XGBoost Mitigated	Fair Data	Sex	Privilege	**79%**	0.47	0.11	0.53	0.2
		Sex	Underprivileged	80%	0.48	0.1	0.53	0.2
		Race	Privilege	79%	0.47	0.11	0.53	0.19
		Race	Underprivileged	79%	0.48	0.11	0.52	0.2

Table 2
Evaluation results for each model using fairness metrics: Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). The dataset used is the Original and Fair Data generated by mitigated Causal Model. The metrics highlighted in bold indicate the best results. Ideal values are: EOD = 0; DI = 1; SPD = 0; and OD = 0.

Algorithm	Dataset	Feature	EOD	DI	SPD	OD
Causal Model	Original	Sex	-0.46	0.0	-0.21	-0.57
		Race	-0.17	0.37	-0.1	-0.21
Causal Model	Fair Data	Sex	-0.46	0.0	-0.19	-0.57
		Race	-0.11	0.63	-0.05	-0.14
Causal Model Mitigated	Original	Sex	**0.01**	0.63	-0.06	**0.01**
		Race	**0.0**	0.7	-0.04	**0.0**
Causal Model Mitigated	Fair Data	Sex	**0.0**	**1.0**	**0.0**	**0.0**
		Race	**0.0**	**1.0**	**0.0**	**0.0**
XGBoost	Original	Sex	-0.46	0.0	-0.21	-0.57
		Race	-0.16	0.38	-0.1	-0.21
XGBoost	Fair Data	Sex	-0.46	0.0	-0.19	-0.57
		Race	-0.11	0.64	-0.05	-0.14
XGBoost Mitigated	Original	Sex	**0.0**	0.7	**0.0**	**0.01**
		Race	**0.0**	0.83	**0.0**	**0.0**
XGBoost Mitigated	Fair Data	Sex	**0.0**	**0.99**	**0.0**	**0.0**
		Race	**0.0**	**1.0**	**0.0**	**0.0**

discrimination gives almost zero importance to the sensitive features while maintaining the importance of the rest.

The results of the shap analysis, correlation, and fairness metrics indicated that the fair data generated in this process are entirely unbiased. These data produce fair results based on objective characteristics and avoid discrimination.

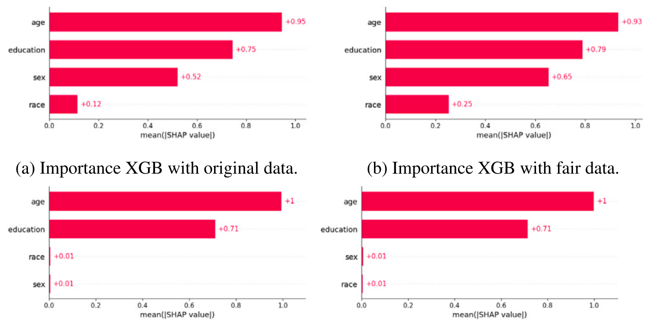
5.3. Algorithm comparison

This section compares the proposed solution to other mitigation algorithms considering two sensitive features: sex and race. In our approach, the mitigation of both sex and race has been carried out simultaneously. However, rival approaches have not been designed to work with multiple features.

This approach will be compared with other pre-processing techniques: Re-sample, Re-weighting, Learning Fair Representation (LFR), and Optimized Pre-Processing (OPP). The goal is to determine which method better transforms the input data to produce discrimination-free results. These techniques were selected because they are most similar to the Fair Data generation proposed in this research.

The comparison will be obtained using transformed data to determine which algorithm can produce more equitable training data. Three tests will be performed in this section: analyzing the correlation, obtaining SHAP values, and evaluating metrics. The algorithm used for training and testing will be XGBoost.

Fig. 6 displays the correlation values for the different techniques and sensitive features. The resample and reweighting techniques present



(a) Importance XGB with original data.

(b) Importance XGB with fair data.

(c) Importance XGBM with original data.

(d) Importance XGBM with fair data.

Fig. 5. The importance get using shap for normal (XGB) and mitigated (XGBM) XGBoost models.

results without correlation of the sensitive features with the label. However, in both cases, this sensitive feature correlates with other features, which could indicate its final impact on classification. The LFR method presents a negative correlation for sensitive features, suggesting that it may attempt to reverse the impact to favor the underprivileged group.

In general, none of the techniques achieves the ideal value obtained in Fig. 4 by the algorithm presented in this paper, which can eliminate the correlation for both features simultaneously with respect to the rest of the features.

As expected, the results of the feature importance analysis in Fig. 7 are related to the previous correlation. Techniques that reduce correlation also mitigate the impact of this feature on classifications. None of the methods completely mitigates the effects of the feature in each prediction. For better or worse, these features are useful for classification.

Fig. 5 shows that the Fair Data generated in this study produces almost zero impact on each feature of the prediction. This means that all groups are treated similarly. Techniques that favor one group over another to reduce discrimination do not treat all groups equally.

Tables 3 and 4 present the algorithm evaluations of the accuracy and fairness metrics. Table 3 shows the review when the impact of sex is mitigated. The algorithms cannot balance the accuracy between the privileged and underprivileged groups. Unlike the previous results, the EOD, SPD, and OD are within the valid range [5]. The algorithm with better performance is OPP. The OPP algorithm correlation and importance are worse than those of the rest. Therefore, the algorithm is likely to produce positive discrimination to achieve these results.

Table 4 evaluates the mitigation of race impact. In this evaluation, the same issue arises as with the sex feature. No algorithm can balance the accuracy between both groups. However, unlike sex, the algorithm with better performance regarding fairness is the re-sample technique, which also performs well when analyzing correlation and SHAP values. The results are consistent and, in this case, discrimination could be mitigated correctly.

In Appendix B, the other two datasets have been used to test the proposed algorithm and compare it with the others. The results show that the dataset generated with the causal algorithm performs consistently, regardless of the use case. This is not the case with the other pre-training techniques, which vary their results depending on the data used.

The Fair Data generated in this study provides the best-performing training data for all tests conducted. The correlation between the target and the rest of the features is reduced to a minimum value. Additionally, the SHAP values show that the importance of these characteristics in classification is minimal. As a result, this dataset balances accuracy between both groups and achieves a perfect evaluation of fairness metrics, producing unbiased results.

6. Discussion

The mitigation algorithm presented in this study aims to achieve equality between groups. Equality entails ensuring that everyone has the same rights, resources, and opportunities, regardless of origin, gender, race, religion, the differently abled or other personal characteristics. A new perspective on fairness focuses on equity. Equity acknowledges people's different needs and circumstances and advocates for providing each individual with what they need to succeed and achieve a fair outcome. Equity involves identifying and addressing the systematic and structural inequalities that prevent certain groups from having equal opportunities and outcomes [78,79].

The aforementioned topic explains why the original dataset did not achieve perfect fairness in Table 2. Due to social disparities, not all groups have identical opportunities to access the same education. An estimated 80% variation in educational opportunities is determined by circumstances, mainly by family origin [80]. Therefore, if the underprivileged groups have a poorer education than the privileged group, this will result in a disparate impact and unfair outcome. The results obtained from the disparate impact metric are interesting because they suggest that this metric can detect social differences, delving deeper into equity rather than equality.

The algorithm proposed in this investigation achieves perfect equality between both groups. However, it is also possible to achieve equity if desired. Instead of assigning equal probability to both groups in Section 4.2, positive discrimination can be given to underprivileged groups to achieve equality with privileged groups. This approach should be guided by an expert, indicating in which cases they might want to favor an Underprivileged group. The advantage of using causal models for mitigation techniques is that they allow for the modification of each circumstance individually.

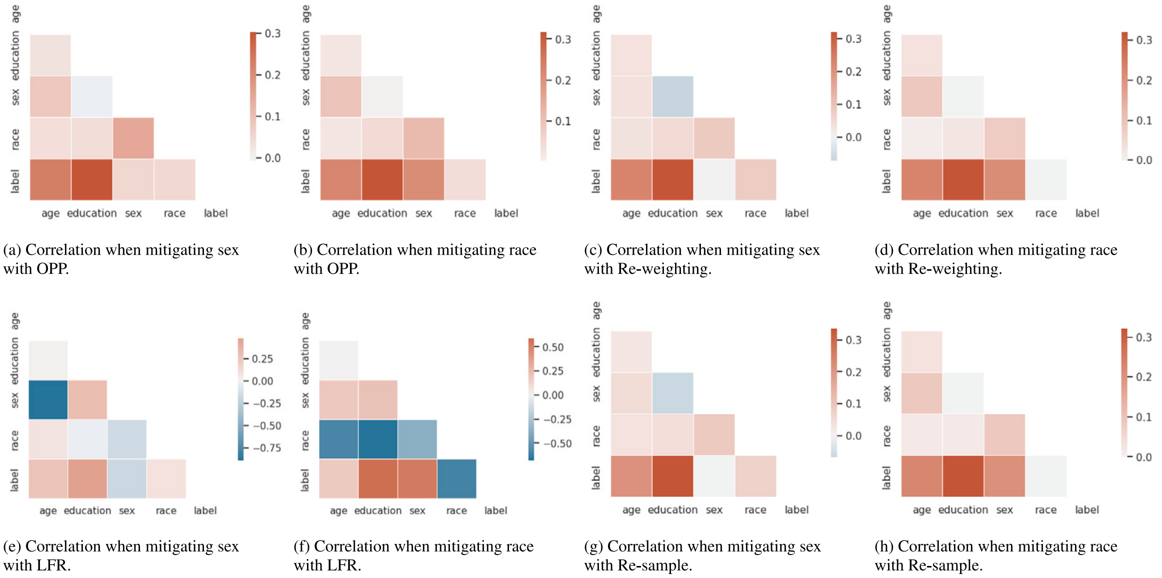
The results demonstrate that the algorithm is devoid of discrimination and that any perceived bias is attributed to social disparities. This shifts the focus from the notion that AI is inherently unfair, emphasizing instead the significance of data and social differences; it recognizes that not all individuals begin their lives from the same standpoint.

The algorithm introduced in this paper effectively mitigates biases between sensitive variables and others. Comparing Fig. 4 with Fig. 6, it is observed that our proposal reduces the correlation not only with the label but also with other features, such as Education and Age. This outcome is distinctly different from what is observed in Fig. 6 where the reweighting and resample methods are used. Although the datasets employed in this study are cases of binary classification, our algorithm is also capable of addressing biases in multilabel datasets. This is achieved by applying mitigation for each state of the target variable. As illustrated in Fig. 4, the presented proposal successfully mitigates biases in non-binary variables such as education and age.

The final algorithms selected for the presented approach include: Hill-Climb Search to infer the graph structure, Bayes Estimator to learn parameters, and Forward Sampling to generate samples. Other algorithms were tested, including Exhaustive Search for graph creation, Naive Bayes, Gibbs Sampling, and Weighted Sampling for sample generation. All these are valid alternatives to use within the presented proposal, although effectiveness and efficiency may differ. For instance, an Exhaustive Search exhibits higher complexity and requires more computational resources, limiting scalability. Furthermore, the choice of a specific sampling technique should be optimized for the specific data set along with other hyperparameters.

7. Conclusions and future work

Employing mitigated causal models that ensure fair impact across various groups enables the creation of nondiscriminatory, well-trained algorithms. Our research indicates that under such models, both privileged and underprivileged groups are treated similarly, with the output being influenced solely by relevant features rather than group attributes.



(a) Correlation when mitigating sex with OPP.

(b) Correlation when mitigating race with OPP.

(c) Correlation when mitigating sex with Re-weighting.

(d) Correlation when mitigating race with Re-weighting.

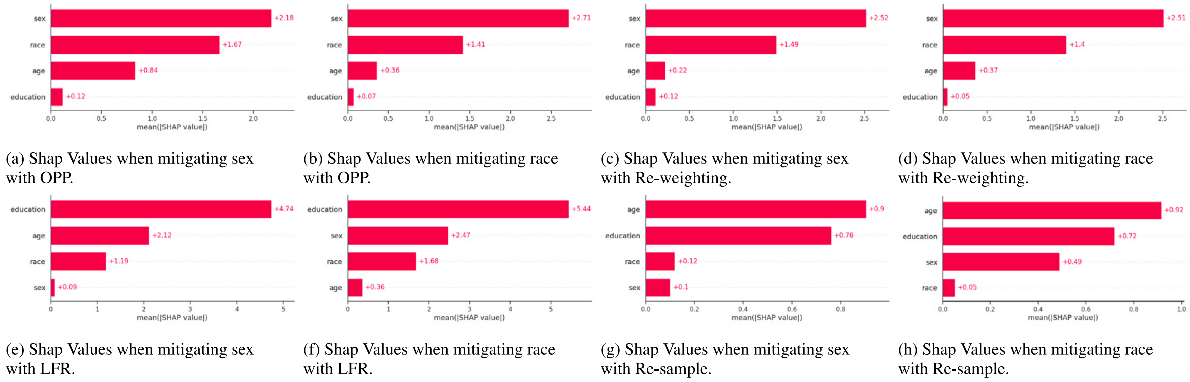
(e) Correlation when mitigating sex with LFR.

(f) Correlation when mitigating race with LFR.

(g) Correlation when mitigating sex with Re-sample.

(h) Correlation when mitigating race with Re-sample.

Fig. 6. Comparing Correlation for other pre-processing mitigation techniques: Re-sample, Re-weighting, Learning Fair Representation (LFR), and Optimized Pre-Processing (OPP).



(a) Shap Values when mitigating sex with OPP.

(b) Shap Values when mitigating race with OPP.

(c) Shap Values when mitigating sex with Re-weighting.

(d) Shap Values when mitigating race with Re-weighting.

(e) Shap Values when mitigating sex with LFR.

(f) Shap Values when mitigating race with LFR.

(g) Shap Values when mitigating sex with Re-sample.

(h) Shap Values when mitigating race with Re-sample.

Fig. 7. Comparing Shap Values for other pre-processing mitigation techniques: Re-sample, Re-weighting, Learning Fair Representation (LFR), and Optimized Pre-Processing (OPP).

Table 3
Evaluation results when the variables Sex and Race are mitigated simultaneously (Fair Data proposal) vs. mitigating only Sex (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). Best results are highlighted in bold, the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	**79%**	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	79%				
	Race	Privileged	**79%**	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	79%				
Optimized Pre-Processing	Sex	Privileged	77%	**−0.01**	0.82	−0.04	−0.03
		Underprivileged	80%				
	Race	Privileged	77%	−0.07	0.67	−0.07	−0.11
		Underprivileged	82%				
Re-weighting	Sex	Privileged	76%	**0.0**	0.68	−0.07	**0.01**
		Underprivileged	85%				
	Race	Privileged	78%	−0.25	0.33	−0.14	−0.32
		Underprivileged	**86%**				
Learning Fair Representation	Sex	Privileged	74%	−0.04	0.57	−0.06	−0.04
		Underprivileged	**87%**				
	Race	Privileged	78%	−0.03	0.68	−0.04	−0.05
		Underprivileged	84%				
Re-sample	Sex	Privileged	76%	**0.0**	0.69	−0.06	**0.01**
		Underprivileged	85%				
	Race	Privileged	78%	−0.25	0.29	−0.15	−0.34
		Underprivileged	**86%**				

Table 4
Evaluation results when the variables Sex and Race are mitigated simultaneously (Fair Data proposal) vs. mitigating only Race (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). Best results are highlighted in bold; the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	**79%**	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	79%				
	Race	Privileged	**79%**	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	79%				
Optimized Pre-Processing	Sex	Privileged	76%	−0.41	0.05	−0.23	−0.53
		Underprivileged	**90%**				
	Race	Privileged	80%	**0.0**	0.71	−0.05	−0.02
		Underprivileged	**86%**				
Re-weighting	Sex	Privileged	76%	−0.48	0.0	−0.23	−0.6
		Underprivileged	**90%**				
	Race	Privileged	80%	0.04	0.86	−0.02	0.05
		Underprivileged	85%				
Learning Fair Representation	Sex	Privileged	76%	−0.46	0.0	−0.21	−0.57
		Underprivileged	88%				
	Race	Privileged	79%	−0.05	0.69	−0.05	−0.06
		Underprivileged	85%				
Re-sample	Sex	Privileged	76%	−0.48	0.0	−0.23	−0.6
		Underprivileged	89%				
	Race	Privileged	79%	0.07	0.96	**0.0**	0.09
		Underprivileged	83%				

Fair data brings us closer to creating more robust, fair, and reliable machine learning models. Our findings indicate that algorithms act primarily as interpreters, processing patterns, and determining outputs based on these inputs. Consequently, data, rather than the algorithm itself, hold the key to unbiased and equitable outcomes.

This study contributes to the literature by offering a pretraining and training mitigation technique that can be applied simultaneously for multiple sensitive features. This approach helps prevent the feedback loop generated in production by using unfair algorithms and also creates fair data for reference. The outcomes of this study include:

- Mitigation Algorithm for Causal Model: This technique is considered during training and can mitigate various biases simultaneously. This model can be used for classification.
- Fair Data: An ideal data set is obtained by applying a sampling technique based on the joint distribution of the mitigated causal model. This technique could be included in the pre-training mitigation algorithm.
- Discrimination-Free Algorithm: The fair data generated in step two can be used to train an algorithm free of discrimination.

Employing a causal model and retaining sensitive features improves explainability and transparency, increasing the trustworthiness of the models for end users. Integrating sensitive features into the loop ensures that they are maintained in the data process. Fairness can be checked and analyzed when a new feature is added to the dataset.

This study placed particular emphasis on explaining and understanding the behavior of mitigation algorithms. The results were detailed and obtained using various algorithms in Section 5.3, contrasting fairness metrics, correlations, and important features between them. A closer look reveals considerable differences in how each achieves these results. In particular, some approaches align more closely with the principles of fairness than others, despite achieving similar metrics. This underlines the importance of explainability and understandability in enhancing the trustworthiness of AI-generated results.

The extension of the proposal presented to unstructured data, such as texts and images, is an important line of future work. Another future direction is to test positive discrimination to achieve equity in real-world cases. This would allow developers, analysts, and policy makers to obtain desired results for different characteristics, guiding learning processes without relying on historical patterns. Besides, this paper, as the revised related works, considers the mapping of sensitive features into only two categories: privileged or underprivileged. The exploration of more challenging mappings, such as discrete privilege level, would be novel in the literature.

A limitation identified in this study is the absence of a method to detect sensitive features automatically. This is because not all features result in discrimination, and fairness metrics are related to machine learning results. Nowadays, experts identify sensitive features, and automating this process is currently a challenge. Future work should focus on improving the detection of potential bias features to trigger fairness metrics.

CRediT authorship contribution statement

Rubén González-Sendino: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Emilio Serrano:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Conceptualization. **Javier Bajo:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work has been granted by the "EICACS (European Initiative for Collaborative Air Combat Standardisation)" project of the Horizon Europe programme of the European Commission, under grant agreement No. 101103669.

Appendix A. Library

FLAI is a Python library designed with two main features: the construction of a causal algorithm and the mitigation of biases within this algorithm.

1. Causal Algorithm Creation: This library simplifies the development of a reliable causal algorithm, setting the foundation for impartial data analysis.
2. Bias Mitigation: Bias mitigation is carried out in two crucial areas - In-Training and Pre-Training.

The library includes features that allow users to mitigate bias in causal algorithms (In-Training Mitigation) in two significant ways:

- Graph Relationship Modification: Relationships within the graph can be altered to establish a more balanced structure.
- Probability Table Modification: The probability table can be fine-tuned to prevent the propagation or amplification of existing biases.

With the mitigated causal algorithm, a mitigated bias dataset can be generated. This dataset can then be utilized for training other algorithms (Pre-Training Mitigation), enabling the bias mitigation process to extend to the initial stages of new model development.

A.1. Installation

FLAI can be installed easily using pip, Python's package installer. This process is typically done in the terminal or command prompt using the following command:

```
pip install flai-causal
```

A.2. Causal creation

The library allows users to create a causal graph using the data and causal_graph modules. The user can then plot the graph to visualize it; Fig. A.1 shows the result.

```
from FLAI import data
from FLAI import causal_graph
import pandas as pd

df = pd.read_pickle('../Data/adult.pickle')
flai_dataset = data.Data(df, transform=True)
flai_graph = causal_graph.CausalGraph(
    flai_dataset, target = 'label')
flai_graph.plot(directed = True)
```

A.3. Causal mitigation

Two main mitigation strategies are available in the library relations and probabilities; Fig. A.2 shows the result.

- Relations Mitigation: The library allows users to mitigate the relationships in the causal graph.

```
flai_graph.mitigate_edge_relation(
     sensible_feature=['sex','age'])
```

- Table Probabilities Mitigation: The library allows users to mitigate the calculation of conditional probability distributions.

```
flai_graph.mitigate_calculation_cpd(
    sensible_feature = ['age','sex'])
```

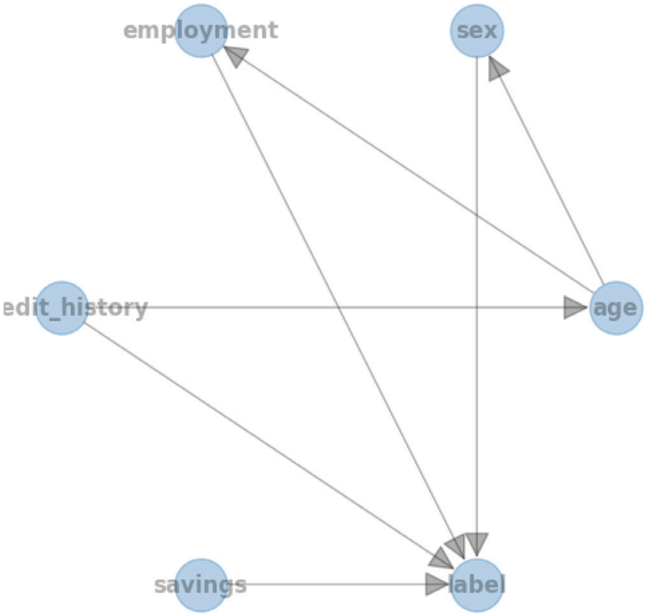


Fig. A.1. Causal graph created.

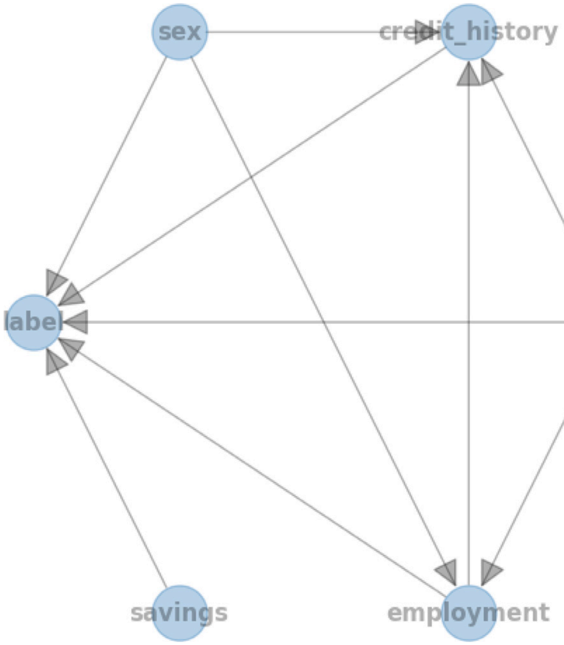


Fig. A.2. Causal graph mitigated.

A.4. Causal inference

The library implements a method to infer the output. This allows us to compare the results with the original and mitigated causal graphs. Table A.1 shows the results of the original graph, exposing the impact of the bias features. Table A.2 shows the results for the mitigated graph, and the bias features have no impact.

```
flai_graph.inference(
    variables=['sex','label'],evidence={})
mitigated_graph.inference(
    variables=['sex','label'],evidence={})
```

Table A.1
Original impact of sex.

sex	label	p
0	0	0.1047
0	**1**	**0.2053**
1	0	0.1925
1	**1**	**0.4975**

Table A.2
Original impact of sex.

sex	label	p
0	0	0.1498
0	**1**	**0.3502**
1	0	0.1498
1	**1**	**0.3502**

Table A.3
Metrics Performance after mitigation. Metrics are: Accuracy (ACC), True Positive Rate (TPR), False Positive Rate (FPR), False Negative Rate (FNR), and Predicted as Positive (PPP).

	ACC	TPR	FPR	FNR	PPP
model	0.7034	0.97995	0.94494	0.02005	0.96948
sex_privileged	0.7024	0.97902	0.94363	0.02098	0.96841
sex_Underprivileged	0.7044	0.98087	0.94626	0.01913	0.97055
age_privileged	0.7042	0.97881	0.94118	0.02119	0.96758
age_Underprivileged	0.7026	0.98109	0.94872	0.01891	0.97139

Table A.4
Metrics Fairness after mitigation. Metrics are: Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD).

	EOD	DI	SPD	OD
sex	0.00185	1.00221	0.00214	0.00448
age	0.00228	1.00394	0.00382	0.00981

A.5. Fair data

A mitigated bias dataset can be generated using the mitigated causal graph. This dataset can train other algorithms, promoting fairness in model outputs.

```
fair_data = flai_graph.generate_dataset(
    n_samples = 1000,methodtype = ’bayes’)
```

A.6. Train algorithm with fair data

Training an algorithm with fair data can be achieved using common libraries such as XGBoost and scikit-learn. Once trained, the model performance and fairness metrics can be calculated. Table A.3 shows the performance results and Table A.4 shows the fairness metrics.

```
model_mitigated = XGBClassifier().fit(
  mitigated_X_train,mitigated_y_train)
metrics = mitigated_dataset.fairness_metrics(
   target_column=’label’,
   predicted_column = ’Predicted’,
   columns_fair = {’sex’ :
      {’privileged’ : 1,’Underprivileged’ : 0},
            ’age’ :
      {’privileged’ : 1,’Underprivileged’ : 0}})
```

The explanation results could be calculated using the Shap library. This reflects the importance of each feature in the decision-making process. Fig. A.3 shows the importance of the features in the original model; Fig. A.4 shows the none impact of the bias features in the mitigated model.

Appendix B. Additional result

This section presents additional results for the Compas and German datasets. The Compas dataset includes sensitive variables of sex and race. The results for sex are in Table B.1, and the results for race are in Table B.2. In both cases, the algorithm proposed in this study yields satisfactory results in terms of fairness metrics. Additionally, the resample metric would be the one that best mitigates gender, and reweighting would be the one that best mitigates race. However, none of them matches the dataset generated in this study.

The German dataset includes sensitive variables of sex and age. The results for sex are in Table B.3, and the results for age are in Table B.4. Similar to the previous case, the fair dataset produces satisfactory results. Among the other algorithms, Learning Fair Representation would perform best.

The results demonstrate the variability of the other algorithms depending on the use case, whereas the one presented in this study produces consistent results regardless of the dataset and variables.

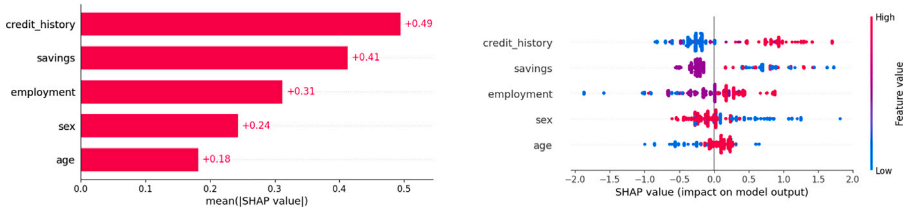


Fig. A.3. Shap results for original XGBoost Algorithm.

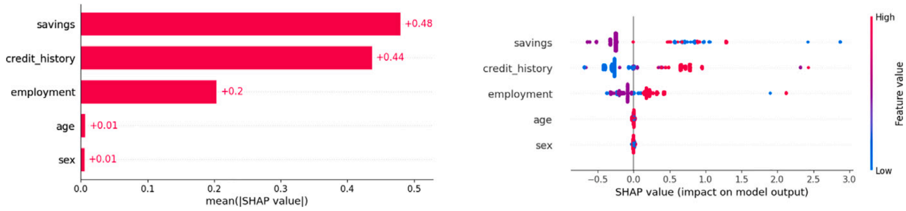


Fig. A.4. Shap results for mitigated XGBoost Algorithm.

Table B.1
Compas Evaluation results when the variables Sex and Race are mitigated simultaneously (Fair Data proposal) vs. mitigating only Sex (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). In bold are highlighted best results, the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	63%	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	63%				
	Race	Privileged	63%	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	63%				
Optimized Pre-Processing	Sex	Privileged	**69%**	−0.1	0.84	−0.11	−0.15
		Underprivileged	65%				
	Race	Privileged	68%	−0.2	0.69	−0.22	−0.35
		Underprivileged	64%				
Re-weighting	Sex	Privileged	**67%**	−0.05	0.81	−0.13	−0.17
		Underprivileged	**67%**				
	Race	Privileged	67%	−0.13	0.69	−0.21	−0.35
		Underprivileged	**68%**				
Learning Fair Representation	Sex	Privileged	64%	0.05	0.98	−0.02	0.03
		Underprivileged	66%				
	Race	Privileged	66%	−0.25	0.57	−0.33	−0.6
		Underprivileged	66%				
Re-sample	Sex	Privileged	66%	**0.0**	0.92	−0.05	**0.0**
		Underprivileged	64%				
	Race	Privileged	65%	−0.1	0.72	−0.19	−0.32
		Underprivileged	65%				

Table B.2
Compas Evaluation results when the variables Sex and Race are mitigated simultaneously (Fair Data proposal) vs. mitigating only Race (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). In bold are highlighted best results, the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	63%	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	63%				
	Race	Privileged	63%	**0.0**	**1.0**	**0.0**	**0.0**
		Underprivileged	63%				
Optimized Pre-Processing	Sex	Privileged	**69%**	−0.06	0.82	−0.12	−0.13
		Underprivileged	**66%**				
	Race	Privileged	**66%**	−0.06	0.85	−0.09	−0.14
		Underprivileged	66%				
Re-weighting	Sex	Privileged	67%	−0.13	0.72	−0.18	−0.29
		Underprivileged	66%				
	Race	Privileged	64%	0.04	0.93	−0.03	**0.01**
		Underprivileged	**68%**				
Learning Fair Representation	Sex	Privileged	**69%**	−0.26	0.60	−0.33	−0.60
		Underprivileged	64%				
	Race	Privileged	64%	0.13	1.11	0.05	0.20
		Underprivileged	66%				
Re-sample	Sex	Privileged	68%	−0.02	0.90	−0.06	**−0.01**
		Underprivileged	64%				
	Race	Privileged	**66%**	−0.11	0.74	−0.19	−0.30
		Underprivileged	64%				

Appendix C. Demo

A dedicated user interface (UI) has been constructed to improve understanding and foster user interaction with the causal models developed. This digital platform, accessed at,[3] offers a comprehensive view of the key attributes and functionalities of the original and mitigated causal models. Alongside the illustrative diagrams, the interface presents a set of meticulously calculated metrics corresponding to each model, providing a more holistic understanding of their operational characteristics.

The UI design emphasizes usability, ensuring that complex operations can be performed easily. For instance, transitioning between the original and mitigated algorithms has been simplified with the click of the 'Mitigated' button. Moreover, the interface integrates counterfactual analysis capabilities. By employing the various controls corresponding to different variables, users can adjust the parameters of interest, and consequently, the system dynamically recalculates the corresponding probabilities.

A particularly noteworthy aspect of the UI demonstration is its ability to highlight the differences in the predictive performance of the original and mitigated models. For example, as demonstrated between Figs. C.1(a) and C.1(b), also between Figs. C.1(c) and C.1(d) given the same selection of variables. The original models predict different probabilities for a favorable outcome (i.e., a label equal to 1), while the mitigated graph returns the same result. This discrepancy underscores the inherent biases in the original model and the corrective action taken by the mitigated model, reaffirming the necessity and effectiveness of our bias mitigation approach.

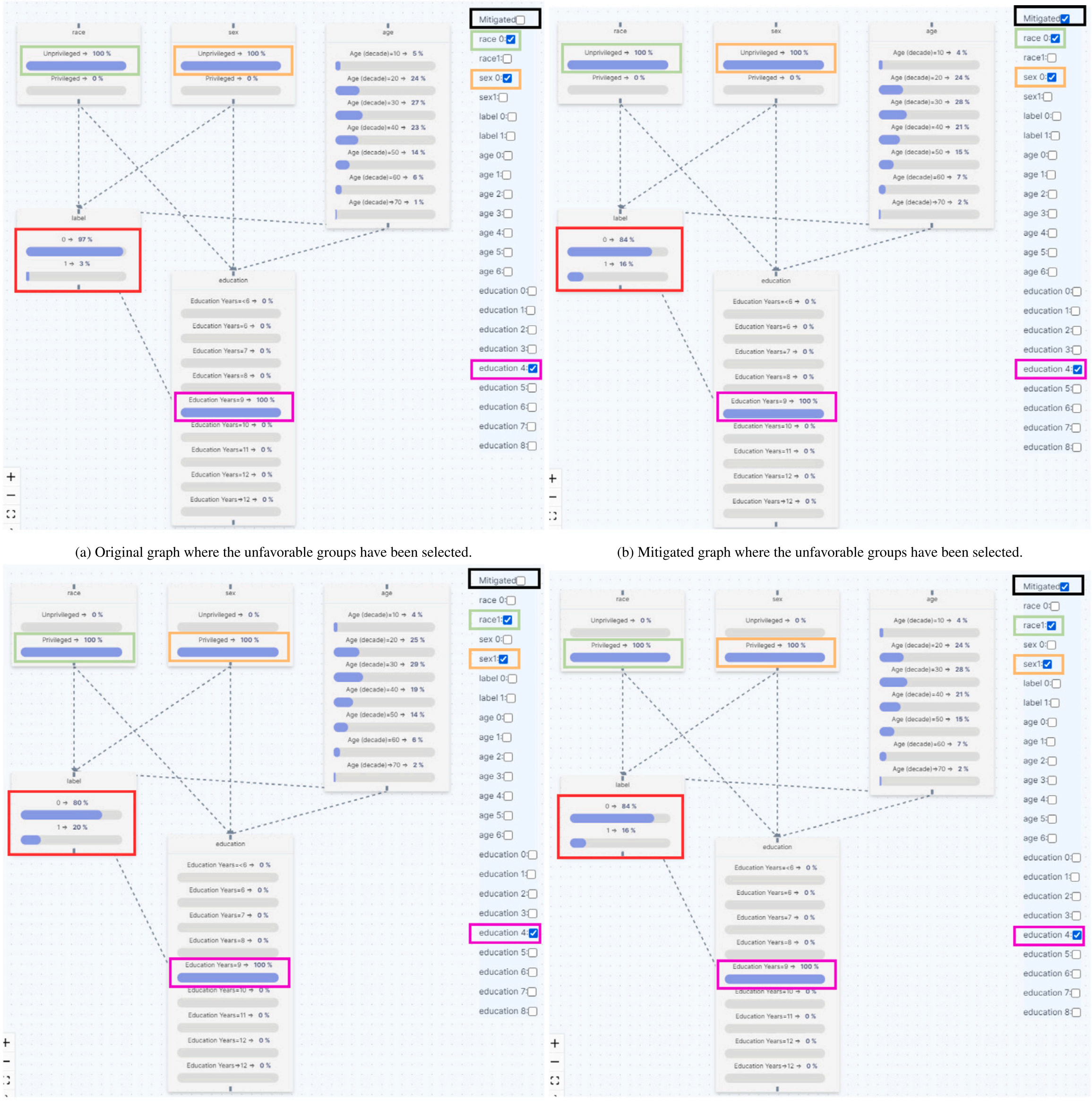
[3] https://www.rubengonzalez.ai/demo

Table B.3
German Evaluation results when the variables Sex and Age are mitigated simultaneously (Fair Data proposal) vs. mitigating only Sex (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). Best results are highlighted in bold; the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	70%	**0.0**	**0.99**	**−0.01**	**−0.02**
		Underprivileged	70%				
	Age	Privileged	70%	**−0.01**	**0.99**	**−0.01**	**−0.01**
		Underprivileged	**70%**				
Optimized Pre-Processing	Sex	Privileged	69%	−0.06	0.81	−0.18	−0.49
		Underprivileged	**73%**				
	Age	Privileged	72%	−0.17	0.71	−0.27	−0.57
		Underprivileged	63%				
Re-weighting	Sex	Privileged	**73%**	−0.15	0.80	−0.19	−0.43
		Underprivileged	69%				
	Age	Privileged	**73%**	−0.27	0.59	−0.4	−0.90
		Underprivileged	68%				
Learning Fair Representation	Sex	Privileged	50%	−0.13	0.92	−0.05	−0.12
		Underprivileged	36%				
	Age	Privileged	44%	0.27	1.30	0.19	0.22
		Underprivileged	55%				
Re-sample	Sex	Privileged	71%	−0.10	0.87	−0.12	−0.26
		Underprivileged	65%				
	Age	Privileged	71%	−0.31	0.61	−0.38	−0.81
		Underprivileged	62%				

Table B.4
German Evaluation results when the variables Sex and Age are mitigated simultaneously (Fair Data proposal) vs. mitigating only Age (rival approaches unable to address more than one variable). Metrics evaluated: Accuracy, Equal Opportunity Difference (EOD), Disparate Impact (DI), Statistical Parity Difference (SPD), and Odds Difference (OD). In bold are highlighted best results, the ideal values are: Accuracy = 100; EOD = 0; DI = 1; SPD = 0; and OD = 0.

Method	Feature	Group	Accuracy	EOD	DI	SPD	OD
Fair Data	Sex	Privileged	70%	**0.0**	**0.99**	**−0.01**	**−0.02**
		Underprivileged	**70%**				
	Age	Privileged	70%	**−0.01**	**0.99**	**−0.01**	**−0.01**
		Underprivileged	**70%**				
Optimized Pre-Processing	Sex	Privileged	**72%**	−0.09	0.81	−017	−0.42
		Underprivileged	68%				
	Age	Privileged	71%	−0.21	0.65	−0.33	−0.74
		Underprivileged	68%				
Re-weighting	Sex	Privileged	**72%**	−0.54	0.45	−0.55	−1.12
		Underprivileged	50%				
	Age	Privileged	66%	−0.07	0.84	−0.14	−0.34
		Underprivileged	63%				
Learning Fair Representation	Sex	Privileged	67%	−0.27	0.75	−0.21	−0.35
		Underprivileged	47%				
	Age	Privileged	63%	0.02	**1.01**	**0.01**	**0.0**
		Underprivileged	53%				
Re-sample	Sex	Privileged	**72%**	−0.07	0.85	−0.15	−0.35
		Underprivileged	**70%**				
	Age	Privileged	**72%**	−0.16	0.71	−0.28	−0.60
		Underprivileged	67%				



(a) Original graph where the unfavorable groups have been selected.

(b) Mitigated graph where the unfavorable groups have been selected.

(c) Original graph where the favorable groups have been selected.

(d) Mitigated graph where the favorable groups have been selected.

Fig. C.1. The Demo interface usages.

References

[1] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (2022) http://dx.doi.org/10.1145/3494672.

[2] R.S. Baker, A. Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. (2021) http://dx.doi.org/10.1007/s40593-021-00285-9.

[3] S. Park, H. Ko, Machine learning and law and economics: A preliminary overview, 2020, http://dx.doi.org/10.1515/ajle-2020-0034.

[4] D. Pessach, E. Shmueli, Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings, Expert Syst. Appl. 185 (2021) 115667, http://dx.doi.org/10.1016/j.eswa.2021.115667, https://www.sciencedirect.com/science/article/pii/S0957417421010575.

[5] R.K.E. Bellamy, K. Dey, M. Hind, S.C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K.N. Ramamurthy, J.T. Richards, D. Saha, P. Sattigeri, M. Singh, K.R. Varshney, Y. Zhang, AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018, CoRR abs/1810.01943. http://arxiv.org/abs/1810.01943, arXiv:1810.01943.

[6] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab, Bias in data-driven artificial intelligence systems—An introductory survey, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 10 (2020) http://dx.doi.org/10.1002/widm.1356.

[7] Y. Zhang, A. Ramesh, Learning fairness-aware relational structures, IOS Press BV, 2020, pp. 2543–2550, http://dx.doi.org/10.3233/FAIA200389.

[8] A.B. Arrieta, N.D. Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[9] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, MIT Press, 2009.

[10] J. Johnson, T. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (2019) 27, http://dx.doi.org/10.1186/s40537-019-0192-5.

[11] Y. Zhang, J. Sang, Towards Accuracy-Fairness Paradox: Adversarial Example-Based Data Augmentation for Visual Debiasing, Association for Computing Machinery, Inc., 2020, pp. 4346–4354, http://dx.doi.org/10.1145/3394171.3413772.

[12] C. Harris, Mitigating cognitive biases in machine learning algorithms for decision making, in: The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020, 2020, pp. 775–781, http://dx.doi.org/10.1145/3366424.3383562.

[13] A. Stevens, P. Deruyck, Z.V. Veldhoven, J. Vanthienen, Explainability and Fairness in Machine Learning: Improve Fair End-To-End Lending for Kiva, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 1241–1248, http://dx.doi.org/10.1109/SSCI47803.2020.9308371.

[14] E. Puyol-Antón, B. Ruijsink, S.K. Piechnik, S. Neubauer, S.E. Petersen, R. Razavi, A.P. King, Fairness in Cardiac Mr Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation, Springer Science and Business Media Deutschland GmbH, 2021, pp. 413–423, http://dx.doi.org/10.1007/978-3-030-87199-4_39.

[15] P. Smith, K. Ricanek, Mitigating algorithmic bias: evolving an augmentation policy that is non-biasing; mitigating algorithmic bias: evolving an augmentation policy that is non-biasing, 2020, https://github.com/yeephycho/tensorflow-.

[16] S. Sharma, Y. Zhang, J.M. Aliaga, D. Bouneffouf, V. Muthusamy, K.R. Varshney, Data Augmentation for Discrimination Prevention and Bias Disambiguation, Association for Computing Machinery, Inc., 2020, pp. 358–364, http://dx.doi.org/10.1145/3375627.3375865.

[17] A. Rajabi, O.O. Garibay, Towards Fairness in AI: Addressing Bias in Data Using Gans, Springer Science and Business Media Deutschland GmbH, 2021, pp. 509–518, http://dx.doi.org/10.1007/978-3-030-90963-5_39.

[18] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K.R. Varshney, Optimized pre-processing for discrimination prevention, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf.

[19] Y. Zheng, S. Wang, J. Zhao, Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models, Transp. Res. C 132 (2021) 103410, http://dx.doi.org/10.1016/j.trc.2021.103410, https://www.sciencedirect.com/science/article/pii/S0968090X21004058.

[20] S. Abbasi-Sureshjani, R. Raumanns, B.E. Michels, G. Schouten, V. Cheplygina, Risk of training diagnostic algorithms on data with demographic bias, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12446 LNCS, 2020, pp. 183–192, http://dx.doi.org/10.1007/978-3-030-61166-8_20.

[21] B.H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340, http://dx.doi.org/10.1145/3278721.3278779, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85058764331&doi=10.1145%2f3278721.3278779&partnerID=40&md5=f6a562b532d746a10314f7a173948b7f, cited by: 425; All Open Access, Bronze Open Access, Green Open Access.

[22] M.A.U. Alam, Ai-fairness towards activity recognition of older adults, Assoc. Comput. Mach. (2020) 108–117, http://dx.doi.org/10.1145/3448891.3448943.

[23] R. González-Sendino, E. Serrano, J. Bajo, P. Novais, A review of bias and fairness in artificial intelligence, Int. J. Interact. Multimed. Artif. Intell. (2024) 1–13, http://dx.doi.org/10.9781/ijimai.2023.11.001, https://www.ijimai.org/journal/sites/default/files/2023-11/ip2023_11_001.pdf.

[24] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K.T. Rodolfa, R. Ghani, Aequitas: A bias and fairness audit toolkit, 2018, http://dx.doi.org/10.48550/ARXIV.1811.05577, arXiv. https://arxiv.org/abs/1811.05577.

[25] S. Ahmed, S.A. Athyaab, S.A. Muqtadeer, Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 557–563, http://dx.doi.org/10.1109/ICICT50816.2021.9358507.

[26] T. Speicher, H. Heidari, N. Grgic-Hlaca, K.P. Gummadi, A. Singla, A. Weller, M.B. Zafar, A unified approach to quantifying algorithmic unfairness, 2018, http://dx.doi.org/10.1145/3219819.3220046.

[27] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, K. Walker, Fairlearn: A Toolkit for Assessing and Improving Fairness in AI, Technical Report MSR-TR-2020-32, Microsoft, 2020.

[28] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, D. Pedreschi, Fairlens: Auditing black-box clinical decision support systems, Inf. Process. Manage. 58 (2021) 102657, http://dx.doi.org/10.1016/j.ipm.2021.102657, https://www.sciencedirect.com/science/article/pii/S030645732100145X.

[29] A. Ortega, J. Fierrez, A. Morales, Z. Wang, M. de la Cruz, C.L. Alonso, T. Ribeiro, Symbolic ai for xai: Evaluating lfit inductive programming for explaining biases in machine learning, Computers 10 (2021) http://dx.doi.org/10.3390/computers10110154.

[30] V. Bhargava, M. Couceiro, A. Napoli, Limeout: An Ensemble Approach to Improve Process Fairness, Springer Science and Business Media Deutschland GmbH, 2020, pp. 475–491, http://dx.doi.org/10.1007/978-3-030-65965-3_32.

[31] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE Trans. Vis. Comput. Graphics 26 (2020) 56–65, http://dx.doi.org/10.1109/TVCG.2019.2934619.

[32] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, F. Polli, Building and auditing fair algorithms: A case study in candidate screening, 2021, pp. 666–677, http://dx.doi.org/10.1145/3442188.3445928.

[33] H. Baniecki, W. Kretowicz, P. Piatyszek, J. Wisniewski, P. Biecek, dalex: Responsible machine learning with interactive explainability and fairness in python, J. Mach. Learn. Res. 22 (2021) 1–7, http://jmlr.org/papers/v22/20-1473.html.

[34] M.J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, 2017, p. 30, https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[35] S.L. Lauritzen, N.A. Sheehan, Graphical models for genetic analyses, Statist. Sci. 48 (2003) 9–514, http://dx.doi.org/10.1214/ss/1081443232.

[36] J. Kaddour, A. Lynch, Q. Liu, M.J. Kusner, R. Silva, Causal machine learning: A survey and open problems, 2022, arXiv:2206.15475.

[37] E. Bareinboim, J.D. Correa, D. Ibeling, T. Icard, On pearl's hierarchy and the foundations of causal inference, in: H. Geffner, R. Dechter, J.Y. Halpern (Eds.), Probabilistic and Causal Inference: The Works of Judea Pearl, in: ACM Books, vol. 36, ACM, 2022, pp. 507–556, http://dx.doi.org/10.1145/3501714.3501743.

[38] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning - problems, methods and evaluation, SIGKDD Explor. 22 (2020) 18–33, http://dx.doi.org/10.1145/3400051.3400058.

[39] J. Pearl, Causality, Cambridge University Press, 2009b.

[40] R. Guo, L. Cheng, J. Li, P.R. Hahn, H. Liu, A survey of learning causality with data: Problems and methods, ACM Comput. Surv. 53 (2021) 75:1–75:37, http://dx.doi.org/10.1145/3397269.

[41] D.M. Chickering, Learning Bayesian networks is np-complete, 1996, pp. 121–130.

[42] G.F. Cooper, E. Herskovits, A Bayesian method for the induction of probabilistic networks from data, Mach. Learn. 9 (1992) 309–347.

[43] D.D. Lewis, Naive (bayes) at forty: the independence assumption in information retrieval, 1998, pp. 4–15.

[44] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.

[45] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inform. Theory 14 (1968) 462–467.

[46] P. Spirtes, C. Glymour, R. Scheines, Causation, prediction, and search, 2000.

[47] D. Heckerman, D. Geiger, D.M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Mach. Learn. 20 (1995) 197–243.

[48] D. Heckerman, D. Geiger, Learning Bayesian networks: A unification for discrete and gaussian domains, 1994, pp. 274–284.

[49] W. Lam, F. Bacchus, Learning Bayesian belief networks: An approach based on the mdl principle, Comput. Intell. 10 (1994) 269–293.

[50] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.

[51] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (1974) 716–723.

[52] J. Rissanen, Strong optimality of the normalized ml models as universal codes and information in data, IEEE Trans. Inform. Theory 47 (2001) 1712–1717.

[53] C.J. Needham, J.R. Bradford, A.J. Bulpitt, D.R. Westhead, A primer on learning in Bayesian networks for computational biology, PLoS Comput. Biol. 2 (2006) e98.

[54] A. Ankan, A. Panda, pgmpy: Probabilistic graphical models using python, 2015.

[55] J. Pearl, Causal inference in statistics: An overview, Stat. Surv. 3 (2009a) 96–146, http://dx.doi.org/10.1214/09-SS057.

[56] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 3 ed., Prentice Hall, 2010.

[57] S. Geman, D. Geman, Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. (1984) 721–741.

[58] G. Casella, E.I. George, Explaining the gibbs sampler, Amer. Statist. 46 (1992) 167–174.

[59] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.

[60] W.K. Hastings, Monte Carlo sampling methods using markov chains and their applications, Biometrika 57 (1970) 97–109.

[61] J.S. Liu, Monte Carlo Strategies in Scientific Computing, in: Springer Series in Statistics, Springer, 2001.

[62] A. Doucet, N. De Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice, Springer, 2001.

[63] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August (2016) 13-17, 2016, pp. 1135–1144.

[64] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[65] Y. Chou, C. Moreira, P. Bruza, C. Ouyang, J.A. Jorge, Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications, 2021, CoRR abs/2103.04244. https://arxiv.org/abs/2103.04244, arXiv:2103.04244.

[66] S. Beckers, Causal explanations and XAI, in: B. Schölkopf, C. Uhler, K. Zhang (Eds.), Proceedings of the First Conference on Causal Learning and Reasoning, PMLR, 2022, pp. 90–109, https://proceedings.mlr.press/v177/beckers22a.html.

[67] M. Cinquini, R. Guidotti, Calime: causality-aware local interpretable model-agnostic explanations, 2022, http://dx.doi.org/10.48550/arXiv.2212.05256, CoRR abs/2212.05256, arXiv:2212.05256.

[68] M. Cinquini, F. Giannotti, R. Guidotti, Boosting synthetic data generation with effective nonlinear causal discovery, 2023, http://dx.doi.org/10.48550/arXiv.2301.07427, CoRR abs/2301.07427, arXiv:2301.07427.

[69] J. Pearl, The do-calculus revisited, 2012, CoRR abs/1210.4852. http://arxiv.org/abs/1210.4852, arXiv:1210.4852.

[70] T. Heskes, E. Sijben, I.G. Bucur, T. Claassen, Causal shapley values: exploiting causal knowledge to explain individual predictions of complex models, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), in: Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 4778–4789, https://proceedings.neurips.cc/paper_files/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

[71] T. Narendra, A. Sankaran, D. Vijaykeerthy, S. Mani, Explaining deep learning models using causal inference, 2018, arXiv:1811.04376.

[72] Álvaro Parafita, J. Vitrià, Explaining visual models by causal attribution, 2019, arXiv:1909.08891.

[73] E. Taskesen, Learning Bayesian networks with the bnlearn python package, 2020, https://erdogant.github.io/bnlearn.

[74] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Sci. Adv. 4 (2018) eaao5580, http://dx.doi.org/10.1126/sciadv.aao5580.

[75] A.W. Flores, K. Bechtel, C. Lowenkamp, False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., Federal Probation 80 (2016).

[76] D. Dua, C. Graff, UCI machine learning repository, 2017, http://dx.doi.org/10.24432/C5X89F, http://archive.ics.uci.edu/ml.

[77] R. Kohavi, Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1996, pp. 202–207.

[78] Y. Wang, When artificial intelligence meets educational leaders' data-informed decision-making: A cautionary tale, Stud. Educ. Eval. 69 (2021) 100872, http://dx.doi.org/10.1016/j.stueduc.2020.100872, https://www.sciencedirect.com/science/article/pii/S0191491X20301206.

[79] C.E. Kontokosta, B. Hong, Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions, Sustainable Cities Soc. 64 (2021) 102503, http://dx.doi.org/10.1016/j.scs.2020.102503, https://www.sciencedirect.com/science/article/pii/S2210670720307216.

[80] M.C. Sentís, S. de la Rica, L. Gorjón, Opportunity bias in spain: Empirical evidence, drivers and trends.

Rubén González is a Machine Learning Engineer with a Master's degree in Artificial Intelligence from Universidad Politécnica de Madrid completed in 2017. He is currently pursuing his Ph.D. at the same institution. Over the years, he has gained extensive experience in innovation departments, actively collaborating on European Horizon 2020 projects. His expertise is rooted in deep neural networks, natural language processing, causal algorithms, and computer vision. Beyond technical intricacies, he is deeply committed to explicability and responsible artificial intelligence.

Emilio Serrano is an Associate Professor in the Department of Artificial Intelligence at Universidad Politécnica de Madrid (UPM). He has also been a Visiting Researcher with the University of Edinburgh, the University of Oxford, and the National Institute of Informatics in Tokyo. His main research line is Social and Explainable Artificial Intelligence for Smart Cities. His scientific production includes more than 80 publications (more than 30 JCR papers). He has been principal investigator in six educational innovation projects in data science, participated in several European and National funding programs (6 European projects), and supervised two Ph.D. theses.

Dr. Javier Bajo, full professor at the Department of Artificial Intelligence, Computer Science School at Universidad Politécnica de Madrid (UPM) and Director of the UPM AI.nnovation Space Research Center in Artificial Intelligence. His main lines of research are Social Computing and Artificial and Hybrid Societies; Intelligent Agents and Multi-agent Systems, Ambient Intelligence, Machine Learning. He has supervised 15 Ph.D. thesis, participated in more than 50 research projects (in most of them as principal investigator) and published more than 300 articles in recognized journals (81 JCR papers) and conferences.