The Matrix Reloaded: A Counterfactual Perspective on Bias in Machine Learning

André V. Carreiro [*] [1] Mariana Pinto [*] [1 2] Pedro Madeira [1] Alberto López [3 4] Hugo Gamboa [1 2]

Abstract

This paper introduces a novel data-centric framework for bias analysis in machine learning, leveraging the power of counterfactual reasoning. We propose a Counterfactual Confusion Matrix, from which we derive a suite of metrics that provide a comprehensive view of a model's behaviour under counterfactual conditions. These metrics offer unique insights into the model's resilience and susceptibility to changes in sensitive attributes such as sex or race. We demonstrate their utility and complementarity with standard fairness metrics through experiments on synthetic data and known real-world datasets. Our results show that our metrics can reveal subtle biases that traditional bias evaluation strategies may overlook, providing a more nuanced understanding of potential model bias.

1. Introduction

Every time we train a Machine Learning (ML) model, we are not just fitting statistical patterns; we are also fitting the biases present in the data. ML models have revolutionized decision-making processes across numerous domains. However, these models also mirror or even amplify bias present in the training data, raising concerns about potentially unfair or discriminatory outcomes. Detecting and mitigating biases is critical for equitable and trustworthy ML systems. Studies often focus on identifying discrimination based on a sensitive feature (e.g., race, sex, age), also known as a protected attribute in specific applications.

There is no consensus on an unequivocal definition of a fair decision in ML, despite numerous philosophical streams emerging over time, including egalitarianism and utilitarianism (Beretta, 2019). Group fairness, a prevalent methodology for bias detection, advocates for equal treatment for all individuals (Dwork et al., 2012). The most common metrics are based on three different criteria (Barocas et al., 2019):

- **Independence**: The outcome should not be influenced by the sensitive feature input and therefore, the portion of favorable outcomes should be equal. Demographic Parity (DemP) translates this into requiring equal predicted prevalence among subgroups (Feldman et al., 2015). This notion may be applied to combat historical bias, when the labels can't be trusted.
- **Separation**: A system is deemed fair if there is equality of errors. This relates to Equalized Odds (EOdds) which requires False Positive Rate (FPR) and False Negative Rate (FNR) parity among subgroups (Hardt et al., 2016). Each of this conditions refer to relaxed versions of this metric, respectively, Predictive Equality (PredEq) and Equality of Opportunity (EOpp).
- **Sufficiency**: Samples given the same outcome should have the same error rate, regardless of subgroup. Predictive Parity (PredP) encapsulates this idea by requiring equal precision among subgroups (Verma & Rubin, 2018).

It is crucial to highlight the non-complementary nature of metrics, which implies that there is no single optimal solution. Therefore, it is essential to establish metrics and their respective trade-offs beforehand. The work of Friedler et al. (Friedler et al., 2016) delves into the mathematical limitations and often unachievable criteria for conjugating different notions. While simple to apply, this approach often disregards the underlying model decision process, focusing solely on the ground truth, the predicted outcome, and the sensitive feature.

To address this limitation, individual fairness notions have been put forward, based on the similarity between samples.

[*]Equal contribution [1]Associação Fraunhofer Portugal Research - AICOS, Porto, Portugal [2]Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys-UNL), Physics Department, NOVA School of Science and Technology, Caparica, Portugal [3]INCMLab, Imprensa Nacional – Casa da Moeda, Lisbon, Portugal [4]Department of Mathematics and CEMAPRE, Instituto Superior de Economia e Gestão (ISEG), University of Lisbon, Portugal. Correspondence to: André V. Carreiro <andre.carreiro@aicos.fraunhofer.pt>.

Data-centric Machine Learning Research (DMLR) Workshop at the 40^{th} International Conference on Machine Learning ICML, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

In this context, we highlight counterfactual fairness, employing the concept of causality. Our work aligns with this perspective of fairness, thus it will be further discussed in the following sections.

In this paper we redesign the Traditional Confusion Matrix (TCM) to adapt it to the counterfactual setting. We introduce the Counterfactual Confusion Matrix (CCM), the Extended Counterfactual Confusion Matrix (ECCM), and a suite of associated metrics, providing a comprehensive approach to evaluate bias and fairness in ML models.

1.1. Related Work

This subsection highlights the related work of applying counterfactual reasoning to fairness and related applications, since this is the approach on which this work heavily relies.

1.1.1. Counterfactual Explanations

Counterfactual Explanations (CFE) is an prominent framework for explainability in ML aiming to improve interpretability through "what if" scenarios. By identifying the minimal set of features that need to be changed to obtain a different outcome, counterfactuals can be used to explain individual predictions (Wachter et al., 2017). Moreover, different authors have explored how the counterfactual generation process can be constrained to ensure plausibility, robustness and meaningfulness of the CFEs (Mothilal et al., 2020; Artelt et al., 2021).

Stepin et al. report that, albeit counterfactuals vast applications in model-agnostic and model-specific settings are being employed in fields of numerical, visual and linguistic data, there is still a need for a standardized evaluation methodology (Stepin et al., 2021).

1.1.2. Counterfactual Fairness

The concept of counterfactual fairness was first introduced by Kusner et al. (Kusner et al., 2018), and defines a predictor as counterfactually fair if the output distribution remains the same in a counterfactual scenario where a sensitive feature is changed. Simply speaking, counterfactual fairness holds when the model predictions are independent from any changes in a sensitive attribute such as sex or race. This approach was shown to be aligned with demographic parity (Rosenblatt & Witter, 2023), which also implies that the predictor is independent of sensitive attributes.

There are two main approaches to achieve counterfactual fairness. The first, with which this work aligns, considers counterfactuals of sensitive features, while the second focuses on counterfactual outcomes. The former, and most common approach, involves changing a sensitive feature (e.g., sex or race) and observe whether the model's distribution is changed (Kusner et al., 2018; Cornacchia et al., 2023a;b; Russell et al., 2017). In contrast, the latter involves estimating a change in the outcome, and observing the effect on the predictor's distribution. This approach is often studied in Risk Assessment Instruments (Coston et al., 2020; 2021; Mishler et al., 2021; Mishler & Kennedy, 2021).

Counterfactual fairness is inherently associated with causal reasoning, as we try to understand how changing a specific variable (sensitive feature or alternative outcome) causally affects another variable (the predictor). However, defining a good causal model is challenging, especially in real-world scenarios with multiple confounding factors (Russell et al., 2017; Kusner et al., 2018; Cornacchia et al., 2023a;b). As removing the sensitive feature has been shown to not solve the bias problem due to correlated proxy features, Chen et al. (Chen et al., 2022) propose a data preprocessing method to remove confounding variables.

Regarding bias evaluation, Coston et al. (Coston et al., 2020) propose counterfactual analogues of common performance and fairness metrics, introducing doubly robust estimation for calculating them. They use counterfactual outcomes, which would have been observed under a different decision policy, and two different models: an outcome model, predicting the outcome based on the features and a treatment/decision; and a treatment/decision model that predicts the treatment or decision based on the features. Mishler et al. (Mishler et al., 2021) propose a post-processing method to achieve counterfactual equalized odds in Risk Assessment Instruments, also computed using doubly robust estimators. These last approaches fall under the category of counterfactual outcomes, while our proposed framework focuses on counterfactual (sensitive) features.

The work of Cornacchia et al. (Cornacchia et al., 2023a) proposes a counterfactual generation tool to study implicit bias in predictive models even when sensitive features are removed. Their approach allows to identify proxy features, defining a metric called Counterfactual Flips, representing the percentage of the generated counterfactuals that belong to different demographic groups as predicted by a sensitive feature predictor. The research group further presented a novel set of fairness metrics including the Counterfactual Fair Opportunity (CFO), a Discounted Cumulative Counterfactual Fairness, and its normalized version (Cornacchia et al., 2023b).

1.2. Contributions

In this work, our contributions are two-fold. Firstly, we present a redesign of the TCM, adapting it to the counterfactual setting. The CCM (and its extended version) emerges as a highly convenient and visual tool for assessing model bias in practical ML. Secondly, we propose a set of metrics derived directly from the CCM and the ECCM, offering a quantifiable and comprehensive assessment of bias in ML models. These metrics are shown to complement the established bias evaluation metrics, allowing us to gain valuable insights into the model's fairness, stability, and susceptibility to underlying biases.

Following this introductory section, Section 2 delves into the methodology underlying the CCM, ECCM, and proposed evaluation metrics. Section 3 presents case studies involving synthetic scenarios and real-world datasets, illustrating the application of our approach in detecting and comprehending bias within ML models. Finally, in Section 4 we summarize the key findings and contributions of our study, discussing potential avenues for future research.

2. Methodology

We propose a novel approach to evaluating bias in ML using the counterfactual framework. Our methodology relies on the comparison between the original model predictions against the predictions made for the counterfactual samples (with the sensitive feature flipped, in the binary case). Analog to the TCM, we propose the CCM and its extended version ECCM, allowing an easy observation of how the model's predictions are influenced by the sensitive attribute (e.g., sex or race).

2.1. Counterfactual Predictions

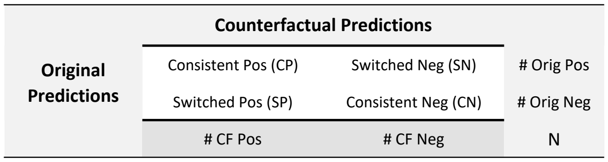
In this work, we generate the counterfactual instances by simply flipping the values of the protected attribute in the data, while keeping the remaining features' values the same. Other, more sophisticated methods are available for counterfactual generation. Nevertheless, this process is beyond the scope of this paper, and we leave their exploration for future work.

As an illustrative example, using *sex* as the sensitive attribute, we create a counterfactual for each sample in the test set by flipping the *sex* from male to female or vice versa. This study focuses on binary attributes, although we could generalize for categorical attributes with m cardinality by analysing one-versus-all or generating $m-1$ counterfactuals. Moreover, one possible approach to handle numerical attributes would be to discretize them into two or more categories.

Although our focus is to evaluate the model bias, this counterfactual framework is fundamentally data-centric. Upon generating the counterfactual dataset, we retain the use of the same model that was initially trained, thereby enabling us to make predictions based on this new, counterfactually enriched data set. This set of counterfactual predictions can then be compared with the original ones to study the model's bias. While this approach assumes that the model's predictions are a function of the input features, without depending on hidden or unobserved variables, we believe it is reasonable for the purposes of this study, as it is a common assumption in the ML literature.

2.2. Counterfactual Confusion Matrix

In a classification task with n classes, the CCM is a $n \times n$ matrix with the rows representing the counts for the original predictions and the columns representing the counterfactual predictions. Figure 2.2 defines the CCM for a binary outcome (positive vs. negative). Although the CCM may be calculated for the whole population, interesting insights can be drawn from observing the CCMs based on individual subgroups of the sensitive feature. The diagonal consists of the Consistent Positives (CP) and the Consistent Negatives (CN), the instances where the original prediction is maintained after the counterfactual flip. The anti-diagonal comprises the Switched Negatives (SN) and the Switched Positives (SP), corresponding to the number of samples where the counterfactual switches the outcome from positive to negative, and negative to positive, respectively.



	Counterfactual Predictions		
Original Predictions	Consistent Pos (CP)	Switched Neg (SN)	# Orig Pos
	Switched Pos (SP)	Consistent Neg (CN)	# Orig Neg
	# CF Pos	# CF Neg	N

Figure 1. **The Counterfactual Confusion Matrix.** A schematic representation showcasing the evaluation of decisions across original and counterfactual samples.

2.2.1. Derived Metrics

Several metrics can be derived from the CCM, some in direct analogy from usual metrics computed from the TCM. We will define the proposed metrics, provide the analog in the TCM, if it exists, and describe its meaning when applied to bias analysis. When appropriate, we further define the complement metric ($comp\_metric = 1 - metric$).

- **Consistency Rate (CR)**: The analog to Accuracy in the TCM, this measures the proportion of instances where the original prediction remained the same after flipping the sensitive feature. It is calculated as $(CP + CN)/(CP + CN + SP + SN)$. The complement of this rate can be defined as the **Switch Rate (SR)**, which is calculated as $(SP+SN)/(CP+CN+SP+SN)$. Although measuring the consistency of counterfactual predictions is not new (Cornacchia et al., 2023a), we formalize it based on the proposed CCM.
- **Positive Switch Rate (PSR)**: Measures the proportion of instances originally predicted as negative that switched to a positive outcome after flipping the sensi-

tive feature. It is calculated as $SP/(SP+CN)$ and stands as the equivalent to the FPR in the TCM. Its complement can be defined as **Negative Consistency Rate (NCR)**, whose equivalent in the TCM is the Specificity.

- **Negative Switch Rate (NSR)**: Inversely, the NSR captures the fraction of original positive predictions that switch from a positive to negative prediction after the counterfactual flip. It is calculated as $SN/(SN+CP)$. The analog in the TCM is the FNR. Its complement may be defined as **Positive Consistency Rate (PCR)** whose TCM analog is the Sensitivity or Recall.

- **Positive Consistent Precision (PCP)**: This metric is equivalent to the Precision in the TCM, and is computed as $CP/(CP+SP)$. It can be interpreted as the proportion of positive counterfactual predictions that remain consistent after flipping the sensitive feature. Its complement is defined as the **Positive Switch Discovery Rate (PSDR)**, whose analog in the TCM is the False Discovery Rate (FDR).

- **Positive-to-Negative Ratio (P2NR)**: The ratio between PSR and NSR, this metric provides a measure of the relative susceptibility of the model's predictions to changes in the sensitive feature, considering both directions: positive-to-negative, and vice-versa. A P2NR > 1 suggests the model is more prone to switch from negative to positive (favouring more positive outcomes). Contrarily, a P2NR < 1 indicates that the model's predictions are more susceptible to switch from positive to negative, implying that the model is biased against the group represented by the flipped attribute, as it is more likely to deny opportunity (the positive outcome).

- **Counterfactual Matthew's Correlation Coefficient (CMCC)**: The counterfactual version of the Matthew's Correlation Coefficient (MCC), measures the alignment between the original and counterfactual predictions. It keeps the favourable properties of the original MCC, like being robust to unbalanced data (in this case the predictions, not the ground truth). It can be computed as $\frac{CP \times CN - SP \times SN}{\sqrt{(CP+SP) \times (CP+SN) \times (CN+SP) \times (CN+SN)}}$.

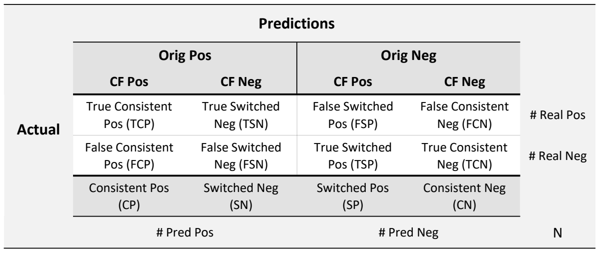
These metrics provide a comprehensive view of the model's performance under counterfactual conditions. To better assess the potential bias, one should compute these metrics for the different subgroups of the sensitive attribute (e.g., males and females when *sex* is the attribute) and compare the obtained results. One could use the absolute difference to get a measure of group disparity, or use a ratio for a relative disparity, especially if higher sensitivity to small changes is desired.

For some of the metrics, higher values for one of the subgroups may suggest that the model is biased against that subgroup, like the PSR, PCP, and P2NR. For metrics like the NSR, higher values for a subgroup may indicate that the predictions are biased in favour of that subgroup. High values of the SR may indicate the presence of bias, but it's difficult to assess in which direction. Higher values for the CR, NCR, PCR, and PCP align with a counterfactually fair model.

However, these metrics do not account for the ground truth labels. This allows to study model bias independent from knowing the real labels of the population. On the oher hand, it may miss important context such as the prevalence or model correctness in specific subgroups in the training set, to better evaluate if the model minimizes, maintains, or amplifies bias.

2.3. Extended Counterfactual Confusion Matrix

We propose an ECCM to allow the visualization of the relationship between the ground truth labels, or actual outcome prevalence, with both the original and counterfactual predictions. In the binary scenario, the ECCM is a 2×4 matrix (with aggregated totals as extra cells), defined in Figure 2. Its structure expands the TCM, dividing each of the original matrix cells into two (consistent and switched). We note that summing each 2-cell column results in the four constituent cells of the CCM: CP, SN, SP, and CN. For instance, the first two cells of the first row (True Consistent Positives (TCP) and True Switched Negatives (TSN)) sum up to the known True Positives (TP), and the two cells below (False Consistent Positives (FCP) and False Switched Negatives (FSN)) the False Positives (FP).



	Predictions				
	Orig Pos		Orig Neg		
	CF Pos	CF Neg	CF Pos	CF Neg	
Actual	True Consistent Pos (TCP)	True Switched Neg (TSN)	False Switched Pos (FSP)	False Consistent Neg (FCN)	# Real Pos
	False Consistent Pos (FCP)	False Switched Neg (FSN)	True Switched Pos (TSP)	True Consistent Neg (TCN)	# Real Neg
	Consistent Pos (CP)	Switched Neg (SN)	Switched Pos (SP)	Consistent Neg (CN)	
	# Pred Pos		# Pred Neg		N

Figure 2. **The Extended Counterfactual Confusion Matrix.** A more granular tool for evaluating subgroup biases, expanding the CCM to analyze the relationship between ground truth labels and both original and counterfactual predictions.

Considering the additional information about the actual outcomes, a new set of metrics can be defined based on the ECCM.

- **True Switch Negative Rate (TSNR)**: Calculated as $TSN/(TSN+FSN)$, it measures the proportion of instances switching from positive to negative that were originally correctly predicted as positives. The complement is the **False Switch Negative Rate (FSNR)**. Analogously, we can define them for SP: the **True Switch Positive Rate (TSPR)** and the **False Switch Positive Rate (FSPR)**.

These metrics allow us to investigate in more detail whether the counterfactual switches are more prevalent when the model makes correct or wrong predictions (e.g., SN can originate from true or false positives). However, we argue that the previous metrics are insufficient when studied by themselves.

- **True Positive Switch Rate (TPSR)**: This measures the proportion of TP that end up switching (to negative) in the counterfactual setting. It's computed as TSN/TP. Similarly, we can define **False Positive Switch Rate (FPSR)** as the proportion of FP that switch to negative when the sensitive feature is flipped, as computed by FSN/FP. Additionally, we can define the equivalent metrics for the negative predictions: the **True Negative Switch Rate (TNSR)** and the **False Negative Switch Rate (FNSR)**.

Based on the metrics above, by comparing their values within a specific subgroup or between different subgroups, we can get a more comprehensive understanding whether the model is more biased when returning correct or incorrect predictions. As an example, if the TPSR is significantly higher than the FPSR, it may suggest that the model's predictions are more robust when it is correct, compared to when it mistakenly predicts negatives as positives. This could help focus the mitigation efforts on the instances where the model currently fails, by collecting more diverse training data in these samples' neighborhood, adjusting the model's parameters, or other targeted bias mitigation techniques.

2.4. Probability-based Counterfactual Analysis

In the context of classification, the model output is not usually a binary outcome. The final prediction results from applying a threshold to a score returned by the model, which can be thought of as the confidence of the model regarding that outcome, or an uncalibrated probability. The metrics we have analysed thus far result from selecting a threshold on these scores (usually a default threshold value of 0.5 is used). However, even when the original and counterfactuals are identical (CR $= 1$), important insights may be hiding beneath the surface. As an example, and assuming the usual decision threshold of 0.5, consider that, for a specific subgroup, all (actual) positive samples were predicted with a score of 0.9, whereas in the counterfactual scenario this score dropped to 0.55. The final binary decision is the same, but we can still observe an impact of flipping the sensitive attribute in the model, suggesting a potential bias. In this context, we present additional metrics to consider the underlying probability scores.

2.4.1. Mean Counterfactual Differences

Let $s_i = f(X_i)$ be the output of model f for sample X_i, or its confidence score. Additionally, X_i^{CF} is the counterfactual version of sample X_i, which in the current work means a similar set of feature values, except for the sensitive attribute, whose value is flipped. We can define a metric that is an analog to the Root Mean Squared Rrror (RMSE) used in error analysis:

- **Root Mean Squared Counterfactual Differences (RMSCD)**: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i^{CF}) - f(X_i))^2}$

We chose especifically the RMSE version, since in this case we're dealing with values in the interval $[0, 1]$. In this context, the differences, in percentage, are more easily interpreted when using the root mean squared differences.

2.4.2. Distribution Shifts

Another approach to compare how the confidence scores change in the original versus the counterfactual scenarios, is based on analysing how the score distributions deviate. In case the outcome is entirely independent from the sensitive feature, both distributions should be identical even after flipping (minus some possibly negligible random factor). One way to achieve this goal is to use the Kullback–Leibler divergence (KLD) to compare both distributions. Also known as relative entropy, the KLD, also defined as $D_{KL}(P||Q)$, measures how a probability distribution P deviates from a second distribution Q.

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(i) \log \frac{P(i)}{Q(i)} \tag{1}$$

In our counterfactual context, we can define P as the distribution of original prediction scores, and Q the corresponding distribution for the counterfactual predictions. D_{KL} is not a metric, in the sense that it is not symmetric, and we choose this ordering, since usually P represents the observations, while Q might represent a theory or a model of P. As expected, a high value for D_{KL} as computed for a specific subgroup suggests the presence of bias in the model. If a proper (symmetric) metric is desirable, we can take the average of $D_{KL}(P||M)$ and $D_{KL}(Q||M)$, where $M = \frac{1}{2}(P + Q)$ (the average distribution), also known as the Jensen-Shannon divergence (JSD).

3. Case Studies

To substantiate the efficiency and applicability of our proposed bias evaluation framework leveraging both the CCM and the ECCM, we put forth a series of case studies that incorporate data from synthetic and real-world settings.

A salient aspect of our research lies in the capability of the CCM, ECCM, and their associated metrics to enhance model evaluation by adding a new layer that unveils biases potentially missed by traditional approaches. While standard metrics provide valuable insights into fairness concerns, they may not capture nuanced biases under the counterfactual setting. In our case studies, we aim to offer a comprehensive fairness evaluation in ML models, reveal hidden biases, and empirically demonstrate our proposal's effectiveness in highlighting unnoticed inequalities in complex real-world scenarios. In the forthcoming scenarios, we will focus on the most relevant metrics to present our empirical findings. Due to space constraints, the results are condensed in a 4×4 matrix format, merging the ECCM for each sensitive subgroup under study.

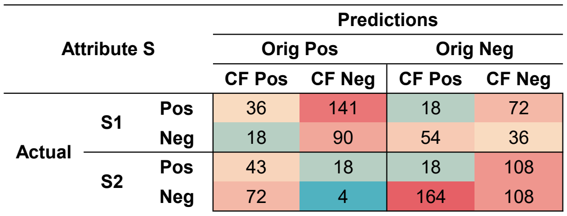
3.1. Synthetic Scenarios

In an endeavor to illustrate the complementarity between the traditional and our proposed counterfactual setup, we have designed two synthetic scenarios. These involve the hypothetical decision of taking a red or a blue pill, representing the positive and negative outcomes in a binary classification problem, respectively. The synthetic scenarios are designed to simulate bias patterns linked to a sensitive attribute denoted as **S**, allowing a systematic analysis on how the ECCM captures and quantifies biases and fairness disparities. This attribute comprises two distinct subgroups as possible values, namely, **S1** and **S2**. Both synthetic scenarios comprise 1000 samples, with balanced representation of **S** subgroups. For the first scenario, we set a ratio $S1/S2$ of 0.87 (S1: 465, S2: 535), while in the second scenario, the ratio is 0.78 (S1: 439, S2: 561). We designed a first scenario in which both traditional and our proposed approach demonstrate concordant and complementary findings and a second scenario in which counterfactual-based metrics reveal biases that are not captured by the traditional methods.

3.1.1. Scenario 1 - Parallel Findings

In this scenario, both traditional and counterfactual methods provide a consistent evaluation of the predicting model for the crafted data, which was designed to manifest a **bias towards the S1 subgroup**, as is shown in Figure 3.

This intentional skew is mirrored in the model's higher rate of classification errors for the **S1** subgroup, reflected by an EOdds value of 0.33, the maximum value of the True Positive Rate (TPR) and FPR differences, with the latter shown in Table 1 together with a subset of the proposed metrics. This divergence is further exemplified by a NSR difference of 0.65, indicating a high probability of a negative outcome when an original **S1** observation switches to **S2**. Examining the P2NR, we observe a value of 2.85 for **S2**, underscoring a bias favouring positive outcomes when this subgroup is flipped to **S1**. Conversely, the P2NR value for **S1** is substantially lower at 0.49, reinforcing the prevailing bias against **S2**. The FPSR difference is also interesting to discuss, as it suggests that, for **S1**, it has a significant proportion of FP switching to negative, compared to **S2** (0.83 vs. 0.05). This suggests that bias is especially prevalent when the model wrongly predicts a positive for **S1**, an insight that could guide further mitigation strategies.



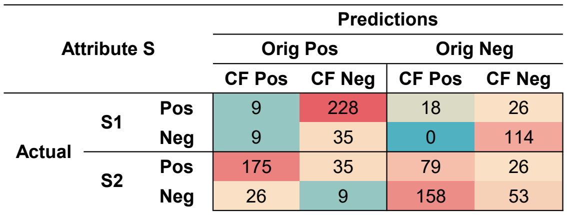
Attribute S			Predictions			
			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	S1	Pos	36	141	18	72
		Neg	18	90	54	36
	S2	Pos	43	18	18	108
		Neg	72	4	164	108

Figure 3. The ECCM generated for the Synthetic Scenario 1, where the proposed framework aligns with the traditional metrics.

We note that, as we generate directly the model decisions, we do not have the confidence scores needed to compute metrics like the RMSCD.

3.1.2. Scenario 2 - Unveiling Hidden Biases

In stark contrast to the first scenario, this setup illustrates a situation where the traditional and counterfactual methodologies diverge significantly in their model evaluation. Here, while the traditional metrics suggest a fair and unbiased model (differences of FNR and FPR under 0.20), the ECCM (Figure 4) and associated metrics reveal an undercurrent of **bias against subgroup S2**.



Attribute S			Predictions			
			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	S1	Pos	9	228	18	26
		Neg	9	35	0	114
	S2	Pos	175	35	79	26
		Neg	26	9	158	53

Figure 4. The ECCM generated for the Synthetic Scenario 2, where the proposed counterfactual metrics unveil hidden biases.

From the results in Table 2, both the PSR and NSR reveal a preference for the **S1** subgroup, as made clear by a high proportion of **S2** switching to a positive outcome when they are flipped to **S1** (PSR$=0.75$). On the other hand, the NSR is very high (0.94) for **S1** indicating that most **S1**

Table 1. Classic and Counterfactual metrics for Synthetic Scenario 1.

	SR	PSR	NSR	P2NR	TPSR	FPSR	TNSR	FNSR	TSNR	TSPR	CMCC	FNR	FPR
Total	0.51	0.44	0.60	0.73	0.67	0.51	0.60	0.17	0.63	0.86	-0.04	0.48	0.34
S1	0.65	0.40	**0.81**	**0.49**	0.80	0.83	0.60	0.20	0.61	0.75	-0.23	0.34	0.55
S2	0.38	0.46	0.16	**2.85**	0.30	0.05	0.60	0.14	0.82	0.90	0.34	0.67	0.22
Diff	0.27	-0.06	**0.65**	-2.36	0.50	**0.78**	0.00	0.06	-0.21	-0.15	-0.57	-0.33	0.33

Table 2. Classic and Counterfactual metrics for Synthetic Scenario 2.

	SR	PSR	NSR	P2NR	TPSR	FPSR	TNSR	FNSR	TSNR	TSPR	CMCC	FNR	FPR
Total	0.56	0.54	0.58	0.92	0.49	0.65	0.56	0.41	0.62	0.14	-0.12	0.25	0.20
S1	0.64	0.11	**0.94**	0.12	0.00	0.41	**0.80**	0.04	0.00	0.13	**-0.09**	0.16	0.28
S2	0.50	**0.75**	0.18	4.18	0.75	0.75	0.26	**0.83**	0.67	0.20	**0.08**	0.33	0.14
Diff	0.14	-0.64	0.76	**-4.06**	**-0.75**	-0.34	0.54	**-0.79**	-0.67	-0.07	-0.17	-0.17	0.14

samples switch to negative when flipping to **S2**. This is also evident from the P2NR large difference and the low values for CMCC.

3.2. Real Scenarios

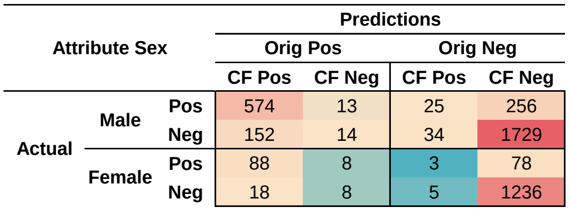
Furthermore, we conduct experiments using real-world datasets - the Adult Census Income (Kohavi & Becker, 1996) and COMPAS Recidivism (ProPublica, 2017), commonly used for bias analysis in ML. We evaluate bias disparities using both traditional and the newly proposed bias metrics derived from the ECCM, presented in Tables 3 and 4, with *sex* and *race* as sensitive features, respectively.

3.2.1. ADULT CENSUS INCOME

To better illustrate the application of the proposed metrics, consider a scenario where a bank employs AI to assess loan applications using a model trained on the Adult Census Income. The bank's AI system aims to determine whether an individual is likely to repay a loan, with a binary decision: granting (positive) or denying (negative) the loan. In a simplified scenario, the bank has set a minimum income threshold of $ 50,000 as the sole requirement for loan approval. To ensure fairness, it is crucial to prevent any disproportionate rejection of loan applications from qualified individuals based on sex. With this goal, different fairness metrics could be employed but, based on the situation, separation metrics would likely be preferred.

We trained the model using the Light GBM algorithm (Ke et al., 2017), obtaining 0.87 accuracy and 0.78 precision. While low, we will assume this would be a satisfactory result. We then calculated the traditional fairness metrics obtaining 0.14 for EOpp and 0.07 for PredEq, as shown in Table 3. The former is the difference of FNR and the latter of FPR. The results suggest a slight bias against women since men have a lower FNR and marginally higher FPR.

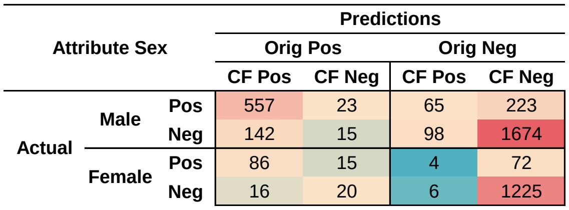
The counterfactual metrics also show only slight biases (CMCC> 0.90), although in a different perspective. We highlight a higher NSR for females (although under 0.20), which could suggest that flipping to males results in negative outcomes more often than the other way around. We note that the FPSR for women, at 0.31, may indicate that almost a third of the FP for this subgroup ends up switching to negative after flipping to male, hinting at a possible (albeit small) source of bias in this type of error.



			Predictions			
Attribute Sex			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	Male	Pos	574	13	25	256
		Neg	152	14	34	1729
	Female	Pos	88	8	3	78
		Neg	18	8	5	1236

Figure 5. The ECCM for the Light GBM model trained on the Adult Income Census data, considering *Sex* as the sensitive feature.

Recalling the original problem, let's assume the bank is subject to legislation that does not accept an EOpp over 0.10 and requires a mitigation process. In this example, we used Fair GBM (Cruz et al., 2023), a fairness-constrained algorithm derived from Light GBM, trained with the same hyperparameters and constrained on FNR.



			Predictions			
Attribute Sex			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	Male	Pos	557	23	65	223
		Neg	142	15	98	1674
	Female	Pos	86	15	4	72
		Neg	16	20	6	1225

Figure 6. The ECCM for the Fair GBM model trained on the Adult Income Census data, considering *Sex* as the sensitive feature.

As expected, we verified a reduction from 0.14 to 0.10 in EOpp, indicating a seemingly fairer model. In spite of that, our metrics reveal an even more pronounced biased behavior. The female NSR increased from 0.13 to 0.26 and the FPSR from 0.31 to 0.56, aggravating the previously detected behaviour. Furthermore, comparing the values of P2NR for male samples, the value for the Fair GBM model is 1.50 (vs. 0.81 for Light GBM). A P2NR> 1 suggests the model is more prone to switch from negative to positive, indicating that the mitigation process worsened the counterfactual bias in favour of female instances.

3.2.2. COMPAS Recidivism

The COMPAS dataset is a collection of records commonly used in the criminal justice system to predict the risk of reincidence. In our example, we employed the Adversarial Debiasing neural network model (Bellamy et al., 2018) in two scenarios: without (base model) and with debiasing, for *race* as sensitive feature, which we simplified to either white or non-white individuals, for classifying if an individual is likely to reincide (positive outcome, albeit negative for the individual) or not (negative outcome).

For this experiment we will analyse the DemP, obtained from the ratio of the predicted prevalence among subgroups (%P), and employ our metrics to unveil potential biases. The standardized threshold for DemP is set at 0.80, yet our initial test without debiasing revealed a rate of 0.53 (0.23/0.43). This indicates that non-white individuals are twice as likely of being assigned as having a high risk of reincidence.

Other classic metrics also support this discrepancy, particularly a higher FNR and lower FPR for white individuals, resulting in an EOpp of 0.21 and a PredP of 0.14. This suggests a bias in favor of white individuals. Our proposed metrics also report a tendency to benefit white individuals, with a slightly lower NSR (remember that the positive outcome is predicted reincidence here) and higher P2NR. Moreover, the value of 0.27 for FPSR suggests that a larger proportion of FP switch to negative (non-reincidence) when flipping other races to white.

Attribute Race			Predictions			
			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	White	Pos	101	11	21	156
		Neg	52	5	25	367
	Other	Pos	366	52	20	261
		Neg	146	54	18	510

Figure 7. The ECCM generated for the base model trained on the COMPAS dataset, considering *Race* as the sensitive feature.

When Adversarial Debiasing was used, it granted an increase in DemP to 0.83, surpassing the legal requirement. Nevertheless, the resulting ECCM, represented in Figure 8, displayed some hidden biases derived from the mitigation process and, as a result, a tendency to impair white individuals. First, we note a higher NSR of 0.30 for white instances, compared to 0.00 for non-whites. Additionally, inspecting the metrics including the ground truth, we observe a higher likelihood for whites to switch TP to False Negatives (FN) (0.28 vs 0.00), and FP to True Negatives (TN) (0.33 vs 0.00) when flipping *race*. On the other hand, when switching the sensitive feature to white, there is a propensity to detect previously overlooked cases (FN) for non-whites, noted by a FNSR of 0.26. These findings highlight the need for complementary evaluation frameworks for fairness in ML since optimizing towards specific criteria may introduce other types of undesirable biases.

Attribute Race			Predictions			
			Orig Pos		Orig Neg	
			CF Pos	CF Neg	CF Pos	CF Neg
Actual	White	Pos	99	39	0	151
		Neg	58	28	0	363
	Other	Pos	360	0	89	250
		Neg	147	0	62	519

Figure 8. The ECCM generated for the Adversarial Debiasing model trained on the COMPAS data, considering *Race* as the sensitive feature.

4. Conclusions

This work introduces a new take on the Confusion Matrix, tailored to a counterfactual setting. The CCM, and its extended version ECCM, provide a clear and efficient means to assess the susceptibility of a predictive model to changes in a specified sensitive attribute. The derived metrics offer valuable insights into the presence of bias and how it impacts specific subgroups. Moreover, the ECCM allows a more granular view on potential bias sources, such as the model's higher susceptibility when making Type-I errors. These insights could help targeted bias mitigation. We demonstrated the applicability and complementarity of our framework in synthetic scenarios and real-world datasets. The findings supported the need for a new perspective, as existing bias mitigation techniques focusing on specific metrics may inadvertently compromise other important fairness criteria. Future research will delve into a more formal look at the proposed metrics, and study trade-off analyses (akin to ROC curves). We also intend to study the plausability of generated counterfactuals. Ultimately, we aim to investigate mitigation strategies leveraging this framework, thereby promoting fairer outcomes in ML.

Acknowledgements

This work was partially supported by the Imprensa Nacional - Casa da Moeda, as part of the IN3+ innovation award initiative.

Table 3. Classic and Counterfactual metrics obtained for the Adult Census Income dataset before and after applying fairness constraints.

	SR	PSR	NSR	P2NR	TPSR	FPSR	TNSR	FNSR	TSNR	TSPR	CMCC	JSD	RMSCD	FNR	FPR
LightGBM															
Total	0.03	0.02	0.05	0.41	0.03	0.11	0.01	0.08	0.49	0.58	**0.92**	0.07	0.05	0.35	0.06
Male	0.03	0.03	0.04	0.81	0.02	0.08	0.02	0.09	0.48	0.58	0.92	0.06	0.05	0.32	0.09
Female	0.02	0.01	**0.13**	**0.05**	0.08	**0.31**	0.00	0.04	0.50	0.63	0.89	0.12	0.04	0.46	0.02
Diff	0.01	0.02	**-0.09**	**0.76**	-0.06	-0.23	0.02	0.05	-0.02	-0.05	**0.03**	-0.06	0.01	**-0.14**	0.07
FairGBM															
Total	0.06	0.05	0.08	0.62	0.06	0.18	0.03	0.19	0.52	0.60	**0.83**	0.18	0.08	0.35	0.06
Male	0.07	0.08	0.05	**1.53**	0.04	0.10	0.06	0.23	0.61	0.60	0.83	0.14	0.07	0.33	0.08
Female	0.03	0.01	**0.26**	**0.03**	0.15	**0.56**	0.00	0.05	0.43	0.60	0.81	0.17	0.08	0.43	0.03
Diff	0.04	0.07	**-0.21**	**1.50**	-0.11	-0.46	0.06	0.18	0.18	0.00	**0.02**	-0.03	-0.01	**-0.10**	0.05

Table 4. Classic and Counterfactual metrics obtained for the COMPAS dataset before and after applying fairness constraints.

	SR	PSR	NSR	P2NR	TPSR	FPSR	TNSR	FNSR	TSNR	TSPR	CMCC	JSD	RMSCD	FNR	FPR	%P
Base Model																
Total	0.10	0.06	0.16	0.39	0.12	0.23	0.05	0.09	0.52	0.51	0.79	0.04	0.05	0.46	0.22	0.36
White	0.08	0.08	0.09	0.85	0.10	0.09	0.06	0.12	0.69	0.54	0.78	0.04	0.05	0.61	0.13	**0.23**
Other	0.10	0.05	0.17	0.27	0.12	**0.27**	0.03	0.07	0.49	0.47	0.80	0.04	0.05	0.40	0.27	**0.43**
Diff	-0.02	0.03	-0.08	0.58	-0.03	-0.18	0.03	0.05	0.20	0.07	-0.02	0.00	0.00	0.21	-0.14	-0.20
Adversarial Debiasing																
Total	0.10	0.11	0.09	1.15	0.08	0.12	0.07	0.18	0.58	0.41	0.78	0.06	0.08	0.50	0.20	0.34
White	0.09	0.00	**0.30**	0.00	**0.28**	**0.33**	0.00	0.00	0.58	-	0.79	0.04	0.08	0.52	0.19	0.30
Other	0.11	0.16	0.00	-	0.00	0.00	0.11	**0.26**	-	0.41	0.80	0.04	0.08	0.48	0.20	0.36
Diff	-0.02	-0.16	0.30	-	0.28	0.33	-0.11	-0.26	-	-	-0.01	0.00	0.00	0.04	-0.01	-0.06

References

Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., and Hammer, B. Evaluating robustness of counterfactual explanations. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 01–09, 2021.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. URL http://www.fairmlbook.org.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

Beretta, E. The invisible power of fairness. how machine learning shapes democracy. 2019.

Chen, H., Lu, W., Song, R., and Ghosh, P. On learning and testing of counterfactual fairness through data preprocessing, 2022.

Cornacchia, G., Anelli, V. W., Narducci, F., Ragone, A., and Sciascio, E. D. Counterfactual reasoning for bias evaluation and detection in a fairness under unawareness setting, 2023a.

Cornacchia, G., Anelli, V. W., Narducci, F., Ragone, A., and Sciascio, E. D. Counterfactual fair opportunity: Measuring decision model fairness with counterfactual reasoning, 2023b.

Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. Counterfactual risk assessments, evaluation, and fairness, 2020.

Coston, A., Kennedy, E. H., and Chouldechova, A. Counterfactual predictions under runtime confounding, 2021.

Cruz, A. F., Belém, C., Jesus, S., Bravo, J., Saleiro, P., and Bizarro, P. Fairgbm: Gradient boosting with fairness constraints. In *International Conference on Learning Representations*. 2023.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kohavi, R. and Becker, B. Adult (census income) dataset. https://archive.ics.uci.edu/ml/datasets/Adult, 1996. UCI Machine Learning Repository.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness, 2018.

Mishler, A. and Kennedy, E. Fade: Fair double ensemble learning for observable and counterfactual outcomes, 2021.

Mishler, A., Kennedy, E. H., and Chouldechova, A. Fairness in risk assessment instruments. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, mar 2021. doi: 10.1145/3442188.3445902. URL https://doi.org/10.1145%2F3442188.3445902.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 607–617, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372850. URL https://doi.org/10.1145/3351095.3372850.

ProPublica. Compas recidivism risk score data and analysis. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis, 2017. ProPublica Data Store.

Rosenblatt, L. and Witter, R. T. Counterfactual fairness is basically demographic parity, 2023.

Russell, C., Kusner, M. J., Loftus, J., and Silva, R. When worlds collide: Integrating different counterfactual assumptions in fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf.

Stepin, I., Alonso, J. M., Catala, A., and Pereira-Fariña, M. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315.

Verma, S. and Rubin, J. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pp. 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *SSRN Electronic Journal*, 2017. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063289.