Medicine After Death: XAI and Algorithmic Fairness Under Label Bias

AKINTANDE, O.J.*, DTU Compute, Technical University of Denmark, Denmark
BIGDELI, S. A., DTU Compute, Technical University of Denmark, Denmark
FERAGEN, A., DTU Compute, Technical University of Denmark, Denmark

Trustworthy AI methods like algorithmic fairness and explainable artificial intelligence (XAI) are becoming increasingly important in the fields of machine learning and artificial intelligence (AI). Yet, while we use trustworthy AI tools to investigate the robustness of AI models, we rarely consider that the AI tools themselves are also models, that can also fail. In this paper, we present a case study highlighting how algorithmic fairness and XAI can lead to incorrect interpretation and bias mitigation when the underlying data suffers from systematic label bias.

Label bias is common in crucial application domains such as healthcare or welfare AI – a well known example being diseases that are underdiagnosed in certain demographic groups. In practice, moreover, the real labels are often inaccessible – consider e.g. mental diseases such as major depressive disorder. Without access to true labels, it becomes challenging to estimate the magnitude of the bias. Prior work has documented well how label bias can propagate into biased predictive models, but the question of how (undetected) label bias affects Trustworthy AI tools remains unexplored. We design a case study using the well-known COMPAS dataset, which actually comes with two real sets of labels: One which is known to be highly biased, and one which is a well-accepted proxy for the underlying effect. This enables us to study label bias in a realistic way. We show how label bias leads to incorrect diagnosis of algorithmic bias, as well as incorrect mitigation. Also, we show that using XAI on models trained on biased labels highlights different important features than when training the same models on unbiased labels. When the label bias is unknown to the user, this can lead to incorrect interpretation of what causes different outcomes.

In conclusion, we find that trustworthy AI in the face of label bias acts as a "medicine-after-death" (MAD) process, that addresses symptoms rather than the root causes of bias and is therefore ineffective at solving the problem.

Reference Format:
Akintande, O.J., Bigdeli, S. A., and Feragen, A.. 2025. Medicine After Death: XAI and Algorithmic Fairness Under Label Bias. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 20 pages.

1 Introduction

Label bias refers to a particular type of data contamination where, given input variables x and labels (or outcomes) y, the labels are *systematically biased*. In this paper we are particularly interested in settings where the label bias follows a demographic pattern. This is a very common scenario: Medicine is rich with examples of diseases that

[*] All authors contributed equally to this research.

Authors' Contact Information: Akintande, O.J., DTU Compute, Technical University of Denmark, Lyngby, Copenhagen, Denmark, ojoak@dtu.dk; Bigdeli, S. A., DTU Compute, Technical University of Denmark, Lyngby, Copenhagen, Denmark, sarbi@dtu.dk; Feragen, A., DTU Compute, Technical University of Denmark, Lyngby, Copenhagen, Denmark, afhar@dtu.dk.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

are systematically underdiagnosed in certain groups, e.g. skin diseases are known to be highly underdiagnosed among patients with dark skin [1], [32] , and mental disorders are believed to be underdiagnosed depending on gender, age and socioeconomic status [27], [22]. For job application screening processes where historically, women or older applicants are systematically undervalued, historical training data will carry the same systematic label bias. In this paper, we shall consider the well-known COMPAS dataset [2], whose *decile scores*, that estimate the risk of re-offense among defendants awaiting trial, are known to carry a strong demographic bias: The risk of re-offense among African-American subjects is generally overestimated, whereas it is underestimated among Caucasians.

Unfortunately, label bias is often not easy to measure. Indeed, if unbiased labels were available, we would use those to train and test our algorithms in the first place. But for applications such as mental health risk prediction or job application screening, we simply do not have access to the true labels. We may suspect that our labels are biased, but we do not know how. As a result, one might easily just decide to "work with what we have" and move along training, testing, bias mitigating and interpreting one's algorithms with imperfect labels – after all, perfect labels do not exist.

However, as Trustworthy AI methods such as algorithmic bias detection and mitigation, or explainable artificial intelligence (XAI), become widely used, it is crucial for practitioners to be alert that in situations with label bias, these methods will not be reliable – indeed, they might even do harm [24]. *In this paper*, we seek to illustrate this by using a case study to illustrate how, in the presence of label bias:

- Algorithmic fairness detection and mitigation techniques fail to appropriately diagnose and mitigate the actual bias.
- Bias mitigation can result in overcompensation, potentially even creating a stronger bias. As such, we risk blue-stamping an algorithm as being "fair" while in reality – without our knowledge – they are actually still highly unfair.
- Label bias can also affect XAI explanations: If the labels belonging to a certain group are systematically incorrect, then features that are predictive of group membership could be used by the model to re-calibrate the predictions accordingly. In such a situation, the XAI explanations would likely include features that are predictive of the group, rather than the (real) outcome. If XAI explanations are used to learn about the underlying phenomena, this could lead to highly skewed conclusions that incorrectly attribute outcomes to the group.
- All of these cases come as a consequence of applying – with good intentions – Trustworthy AI techniques in a scenario where labels are systematically biased. While algorithmic fairness and XAI are highly relevant in addressing the problems, not taking the label bias into account leaves standard methods unable to address the root cause of the bias.

Medicine After Death. In this paper, we therefore liken the application of Trustworthy AI methods in the presence of label bias, to the process of trying to rescue a dead person by administering them a drug: The medicine is not only late but could be unethical. Even though there is nothing wrong with Trustworthy AI methods applied, they are applied in a situation where they are unable to work as intended – the deeds have already been done and we have a medicine-after-death (MAD) scenario.

2 Related work

Algorithmic fairness and XAI are, by now, established research fields with methods used across both academic and industrial applications. Through increasing AI legislation such as the EU AI act, they will be increasingly used for validating and certifying AI tools, especially in critical applications such as healthcare, welfare or law.

2.1 Algorithmic fairness

The literature offers a number of different fairness definitions, metrics and mitigation techniques. Fairness definitions and metrics include concepts such as Equalized Odds and Equal Opportunity [19], Demographic parity and Conditional Statistical Uniformity [34]; Fairness by Awareness [23]; Fairness by Unawareness [18] and [23]; Treatment Equality [8]; Counterfactual Fairness [26]; and Domain Relation Fairness [14], we refer to [6] for a systematic overview of the most common fairness metrics with their various criteria. In this paper, we will focus on three representative fairness criteria from [6], namely independence, equalized odds/separation, and sufficiency.

Despite the myriads of fairness mitigation techniques developed in the past, it remains unclear (under what circumstances) which methods work [15]. Consequently, early work considered fairness measures as a legal constraints [16]. The advantage of this technique is that it maintains accuracy at a high level while constraining models to those that satisfy the fairness constraint. However, this approach is inconsistent due to the impossibility of having a strict mathematical interpretation of the legal definition of discrimination. For example, [15], mathematically demonstrates the key relationships between fairness conditions and bias injections. It reveals that while the model's behavior under label bias can be explained—particularly since the algorithm is designed to address selection bias—its performance in mitigating the specific bias it targets remains inadequate. However, the study acknowledges that the conditions outlined in the paper are primarily theoretical, and identifying when a distribution meets these conditions still poses significant challenges.

Similarly, independence, also known as demographic parity, defines "fairness" as equal probability of favorable (desirable) or undesirable outcomes for all groups, regardless of individual attributes. This definition does come with undeniable advantages: It does not rely on labels and therefore does not suffer from problems with label quality. Moreover, it is highly intuitive, as it corresponds directly to the acceptance rates frequently used to study discrimination in practice. However, the approach also comes with obvious downsides: Whenever prevalence actually differs between groups – for instance, breast cancer is more prevalent among women than men – enforcing independence will enhance errors, and may reduce the overall predictive accuracy of the model, resulting in demographic tokenism, where diversity is pursued superficially without addressing deeper, systemic issues. Moreover, independence can conflict with other fairness metrics, such as Equalized Odds or Sufficiency.

Equalized odds requires equality of both the true positive rate (TPR) (and, as a consequence, the false negative rate) and the false positive rate (FPR), ensuring that the model recognizes false negatives and positives equally across groups. This approach helps to prevent discrimination in multiple dimensions. However, it still faces some of the same challenges as demographic parity: ensuring Equalized odds might come at the expense of overall model accuracy, and it might not align well with simultaneously satisfying other fairness definitions [6, 26, 29].

2.2 Explainable AI (XAI)

Explainable AI (XAI) refers to a collection of techniques and methods aimed at making the decision-making processes and predictions of machine learning models more transparent, interpretable, and comprehensible to humans. As artificial intelligence systems are increasingly implemented in critical sectors such as healthcare, welfare, and criminal justice, the importance of explainability has grown considerably. Consequently, there are many existing XAI technique - such as model-specific explainability techniques tailored to specific types of models, such as linear models, decision trees, or neural networks, [20] and [5]; post-hoc explainability techniques, such as local interpretable model-agnostic explanations (LIME) [28], SHapley Additive exPlanations (SHAP) [25] and counterfactual explanations [35]. These are all applicable to black-box models after they have been trained, without requiring access to the model's internal structure. Other techniques are visualization-based techniques (saliency maps [31], t-SNE and PCA [33]), rule extraction techniques, e.g., RuleFit, combines decision rules with linear models to provide interpretable explanations [17] and fairness-aware XAI, such as fairness constraints in models [36] and fairness metrics and explainability [7] which focuses on explaining and mitigating biases in AI models such as adversarial debiasing ensure fairness while maintaining interpretability.

In this paper, as we used tree-based models for tabular data, we used Gain-based Importance typically useful to interpret and explain the behavior of decision trees and random forests [11], and gradient boosting machines [10]. It provides insights into which features are most important for the model's predictions, offering a global explanation of the model by summarizing the importance of features across the entire dataset.

2.3 Label bias

The effect of label bias on machine learning algorithms has been previously documented, and several techniques exist for mitigating the effect of label bias on learned algorithms [21, 30]. However, most of these methods rely on assumptions that are hard to verify when correct labels are unavailable. For instance, [21] relies on setting parameters using true labels y_{truth}, which are typically unavailable. On the other hand, [30] proposes a method to measure and correct label bias, based on using majority group models to correct the labels on minority groups. This rests on an underlying assumption that the domain shift between groups is *only* in the labels, and not in the input features x. Finally, [37] hypothesizes that instances with label errors will result in low confidence for the given label, and use this to prune the training set. However, such approaches are bound to suffer when applied to underrepresented and potentially out-of-distribution groups.

Another challenge for studying label bias is that the most relevant problems are those where true labels are not available. Most existing datasets do not come with two sets of biased and unbiased labels, and many studies therefore resort to randomly flipping labels in order to obtain a dataset with biased labels [13, 37]. We note, however, that this is most likely not how label bias occurs in reality, meaning that the observed effects might not be realistic and that the assumptions underlying the associated methods might not hold for real life label bias.

In an attempt to systematize different techniques for mitigating label bias, [15] theoretically demonstrates the key relationships between fairness conditions and bias injections. It reveals that while the model's behavior under label bias can be explained—particularly since the algorithm is designed to address selection bias—its performance in mitigating the specific bias it targets remains inadequate. However, the conditions outlined in the paper are primarily theoretical, and identifying when a distribution meets these conditions still poses significant challenges.

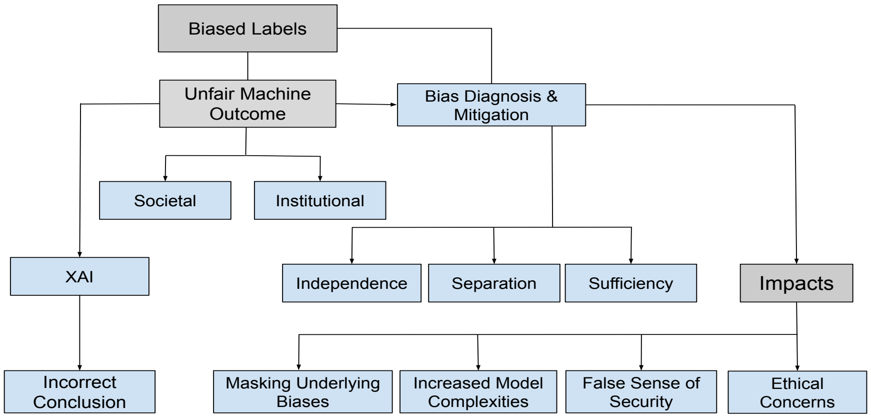

Fig. 1. Medicine After Death: While it is well known that biased labels are likely to lead to biased models, their effect on trustworthy AI methods such as XAI and bias diagnosis and mitigation have not been studied. Training on biased labels is likely to change the explanations given by XAI – which could be positive if it helps highlight that there is a problem. Here, however, we concern ourselves with what happens if the issue goes undiagnosed – in this case, using XAI to understand the underlying predictive mechanisms could lead to incorrect conclusions about the causes of different outcomes. This, again, could manifest as increased discrimination if, e.g., it becomes an established fact that gender influences disease prevalence. Similarly, we showcase how label bias affects bias diagnosis and mitigation across three different fairness definitions, and discuss the associated negative impacts.

Moreover, the paper points out that there are situations where minimizing fairness measures does not result in the fairest possible distribution.

As we see above, existing work addresses the situation where we as developers are aware that we have a certain label bias problem, and want to mitigate its effect on our predictive performance. The viewpoint that we take in this paper is different in two crucial ways:

- We do not just study the effect of label bias on predictive performance, but also on the next step of the Trustworthy AI pipeline: What is the effect of label bias on bias diagnostics and mitigation? On XAI?
- We take a practitioner's point of view and ask ourselves: What would happen if our data suffers from a label bias that we do not know and/or do not understand?

We do not aim to invalidate existing fairness metrics or XAI methods, but rather to raise questions about the reliability of these metrics under biased labels. Our long-term goal would be to automatially assess when trustworthy AI algorithms can be expected to be robust – and when they cannot.

To this end, we present our results in two parts. First, we consider the scenarios of training and testing the model with an observed biased label y_{bias}, and compare the observed performance to the real one obtained using the true label y_{truth}. In this way, we observe the true effect of diagnosing and mitigating algorithmic bias using systematically biased labels. Next, to observe how label bias affects model explainability, we invert the process by training with the true labels y_{truth} and compare the importance of features across the two experiments. Our study is illustrated schematically in Fig. 1.

3 Case Study: Imitating a Trustworthy AI pipeline

Let us start by imitating how a "Trustworthy AI" pipeline would look like when a developer is – without realizing it – training, testing and interpreting a machine learning model on labels that are systematically biased, and next using those biased labels to mitigate any found algorithmic biases.

3.1 Details on data

We base our experiments on the well-known COMPAS dataset, and argue that it is a great real-life dataset for studying label bias. COMPAS is an algorithm developed for the US judicial system, which assesses the likelihood of a defendant becoming a recidivist [4, 12], to assist decisions on whether to allow defendants bail, or whether they would have to await trial in jail. As shown by ProPublica [3], a major known problem with COMPAS is that the labels suffer from disparate measurement bias. Comparing the risk predictions made by COMPAS to actual 2-year recidivism rates, it was shown that Caucasians were overrepresented in false negatives (allowed bail and then committing new crime), whereas African Americans were overrepresented in false positives (denied bail but not committing new crime).

Since the available COMPAS dataset contains both sets of labels, we in this paper consider the thresholded COMPAS *decile score* as our observed, biased label y_{bias}, whereas we consider the *2-year recidivism* as the underlying true label y_{truth}. Of course, in reality, the 2-year recidivism is also not unbiased – but we find it safe to use it as such in our illustrative example, as it is generally accepted as a good proxy for y_{truth}.

The COMPAS dataset consists of 4743 instances of US criminal cases, each represented with input features along with two different outputs: the decile score $D \in \{1, \ldots, 10\}$ produced by the (to us unknown) COMPAS algorithm, as well as the 2-year recidivism score documented by ProPublica [2]. For the purpose of our analyses, the decile score is thresholded into > 5 or ≤ 5. Details on the distribution of scores across the dataset are found in Table 1. To observe the bias before and after mitigation effect across races, we dichotomized the race variable into binary one-versus-all features, namely - Race: African-American; Race: Caucasian; Race: Hispanic and Race: Others, respectively.

Table 1. This data summary provides descriptive statistics across different racial groups and key features, including Juvenile Misdemeanor Charges (JMC), Juvenile Felony Count (JFC), Prior Count (PC), and Criminal Charge Degree (CCD).

Race	Total	y_{bias} (D > 5)	y_{truth}	JMC > 2	JFC > 2	PC > 2	CCD	age (μ)
African-American	2266	1166 (51.46%)	471 (20.78%)	45 (1.99%)	32 (1.41%)	1159 (51.15%)	789 (34.82%)	32.6
Caucasian	1701	455 (26.75%)	209 (12.29%)	14 (0.82%)	3 (0.176%)	656 (38.57%)	740 (43.50%)	37.6
Hispanic	459	113 (24.62%)	53 (11.55%)	1 (0.22%)	0 (-)	152 (33.12%)	208 (45.32%)	35.2
Others	318	65 (20.44%)	42 (13.21%)	2 (0.63%)	2 (0.63%)	95(29.87%)	126 (39.62%)	35.8

3.2 Model training and testing

We train the Sklearn Gradient Boosting model on randomly sampled training / test splits of 70 / 30%. At this point, we train using the observed biased labels y_{bias}, and report both the observed performance with respect to the test labels y_{bias}, but also the real underlying performance with respect to the test labels y_{truth}. As we see in Table 2, and as we would naturally expect, the observed performances that we obtain when using the observed test labels y_{bias} do not truthfully represent the actual performance of the algorithm but tend to be over-estimated.

Table 2. Observed and real performance of the trained model: We trained the model on the COMPAS dataset using the training labels y_{bias} consisting of the thresholded *decile score* from the COMPAS algorithm. These labels were also used to generate "observed" performance, whereas ProPublica's *2-year recidivism* labels are used to generate the "real" performance with respect to the less biased label y_{truth}.

Race	**Accuracy**		**AUC**		**Sensitivity**		**Specificity**	
Train with y_{bias}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.7088	0.6397	0.7028	0.6595	0.7715	0.7738	0.6340	0.5453
Caucasian	0.7707	0.7098	0.6950	0.6049	0.4651	0.3548	0.9249	0.8549
Hispanic	0.8282	0.7423	0.6865	0.5855	0.4211	0.2619	0.9520	0.9091
Others	0.7970	0.7820	0.6422	0.6589	0.3838	0.3500	0.9406	0.9677

3.3 Bias diagnostics

Next, we illustrate how systematic bias in test labels affects metrics typically used for bias diagnostics. In Table 3, we report group-wise selection rates (quantifying the independence fairness criterion), true and false positive rates (used to quantify the separation criterion, also called equalized odds), and positive and negative predictive values (used to define the sufficiency criterion).

Moreover, in Table 4 we report the following targeted metrics for the three fairness definitions, computed using each group g as an "anchor" for a more complete picture:

For independence, we consider the maximum group-wise difference between selection rates S_i (for the anchor group g and remaining groups i):

$$M_{ind}(g) = max\left[abs\Big(S_g - S_i\Big)\right]_{i=1,\ldots G}. \quad (1)$$

Note that $M_{ind}(g) \in [0,1]$, and that independence is satisfied if $S_i = S_g$ for all $i = 1, \ldots G$, which happens exactly if $M_{ind}(g) = 0$.

For separation, we again consider maximum group-wise differences between error rates. Since we need both error rates to be equal, we add their differences to get a single metric:

$$M_{sep}(g) = max\left[abs\Big(TPR_g - TPR_i\Big) + abs\Big(FPR_g - FPR_i\Big)\right]_{i=1,\ldots G} \quad (2)$$

Note that $M_{sep}(g) \in [0,2]$. Since separation is satisfied if: $TPR_i = TPR_g$ and $FPR_i = FPR_g$ for all i, we see that separation is satisfied if $M_{sep}(g) = 0$.

Table 3. Overview of observed and true metrics related to the three fairness criteria. Note, in particular, that the observed true positive rates and negative values are overestimated, whereas the observed false positive rates are underestimated.

Race	**Selection rate**		**TPR**		**FPR**		**PPV**		**NPV**	
Train with y_{bias}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.5866	0.5866	0.7715	0.7738	0.3660	0.4547	0.7156	0.5453	0.6992	0.7738
Caucasian	0.2059	0.2059	0.4651	0.3548	0.0751	0.1451	0.7576	0.5000	0.7741	0.7642
Hispanic	0.1350	0.1350	0.4211	0.2619	0.0480	0.0909	0.7273	0.5000	0.8440	0.7801
Others	0.1278	0.1278	0.3438	0.3500	0.0594	0.0323	0.6471	0.8235	0.8190	0.7759

Table 4. Overview of observed and real fairness metrics. Note that while separation is independent of labels y, the separation and sufficiency metrics are consistently underestimated by the observed criteria.

Race	**Independence**		**Separation**		**Sufficiency**	
Train with y_{bias}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.4588	0.4588	0.7343	0.8757	0.1883	0.2803
Caucasian	0.3807	0.3807	0.5973	0.7286	0.1554	0.3352
Hispanic	0.4516	0.4516	0.6684	0.8757	0.1564	0.3277
Others	0.4588	0.4588	0.7343	0.8462	0.1883	0.3352

Finally, for sufficiency, we consider maximum group-wise differences between predictive values. Imitating our metric for sufficiency, we add the differences for both predictive values to get a single metric:

$$M_{suff}(g) = max\left[abs\Big(PPV_g - PPV_i\Big) + abs\Big(NPV_g - NPV_i\Big)\right]_{i=1,\ldots G}. \quad (3)$$

Again, $M_{suff}(g) \in [0,2]$ and it takes the value 0 if sufficiency is achieved.

Tables 3 and 4 show the observed (with respect to y_{bias}) and true (with respect to y_{truth}) bias diagnostics. Note that, as expected, there is no differences between the observed and real bias metrics for the independence criterion, as the independence criterion does not rely on labels y. We note, however, that for the separation and sufficiency metrics, the observed bias metrics are clearly underdiagnosing bias compared to the true situation. To further illustrate the difference between the observed and real performances, we also show the ROC curves, which relate closely to the separation criterion, in Figure 2. Note how, when computing ROC curves with respect to the true labels, the performance is lower and the curves show a higher variability.

3.4 Bias mitigation

Next, we illustrate what would happen – i.e. both what we would observe and what would *really* happen – when we use the observed but biased labels to perform bias mitigation. Here, we utilize FairLearn's Threshold Optimizer to mitigate with respect to the Equalized Odds criterion using randomization [9].

Similar to above (Table 3), Table 5 shows the observed and real base performance metrics used in fairness criteria are after mitigation. The derived observed and real fairness metrics are shown in Table 6.

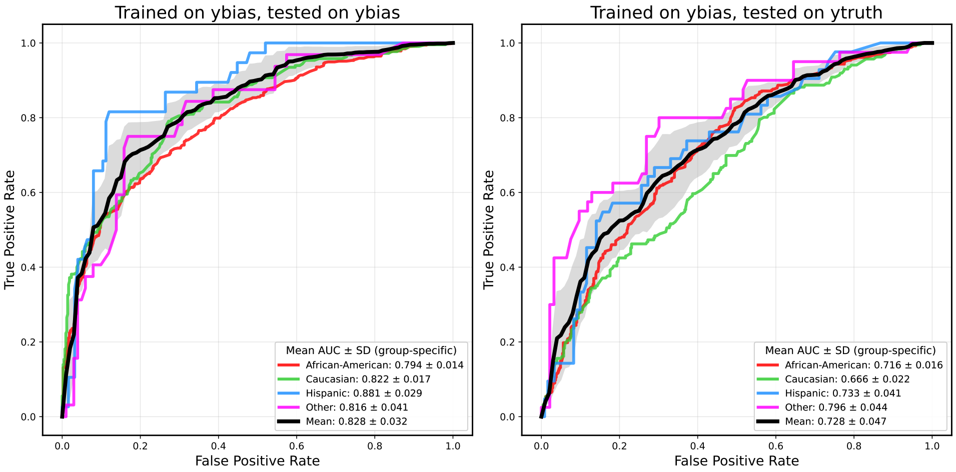

Fig. 2. Per-group ROC curves computed with respect to the observed label y_{bias} (left) and the real variables y_{truth} (right).

Table 5. Overview of observed and true metrics related to the three fairness criteria, following mitigation.

Race	**Selection rate**		**TPR**		**FPR**		**PPV**		**NPV**	
Mititgate with y_{bias}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.4123	0.4123	0.5840	0.5758	0.2075	0.2971	0.7706	0.5773	0.6148	0.7016
Caucasian	0.3058	0.3058	0.5953	0.4624	0.1596	0.2418	0.6531	0.4388	0.8045	0.7753
Hispanic	0.2270	0.2270	0.6579	0.4524	0.0960	0.1488	0.6757	0.5135	0.8968	0.8175
Others	0.2481	0.2481	0.5313	0.6000	0.1584	0.0968	0.5152	0.7273	0.8500	0.8400

Table 6. Overview of observed and real fairness metrics- following mitigation.

Race	**Independence**		**Separation**		**Sufficiency**	
Train with y_{bias}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.1853	0.1853	0.1854	0.2717	0.4906	0.2884
Caucasian	0.1065	0.1065	0.1262	0.2826	0.3072	0.3535
Hispanic	0.1853	0.1853	0.1890	0.2717	0.3769	0.2363
Others	0.1642	0.1642	0.1890	0.2826	0.4906	0.3535

In terms of mitigation effect, we make the following observations: While before, we were generally just underestimating bias, we are now still underestimating bias with respect to the separation criterion, but we are now overestimating the bias with respect to the sufficiency criterion.

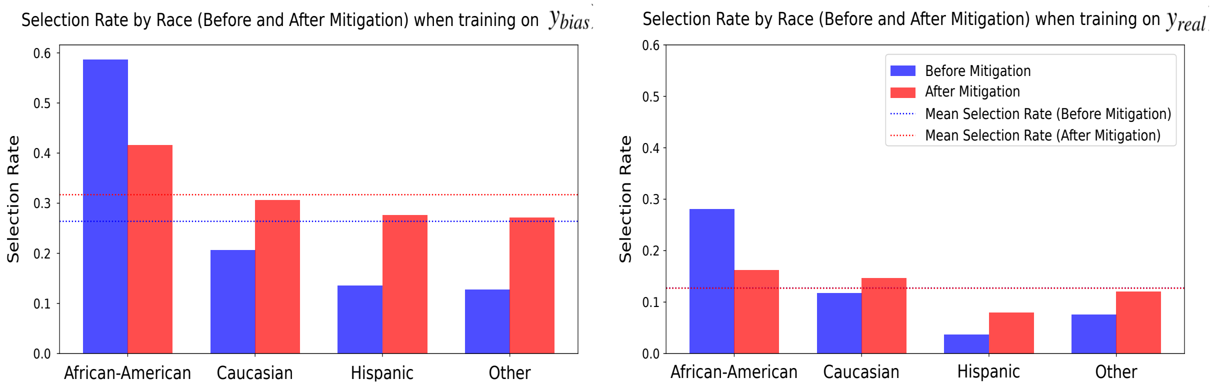

Fig. 3. Independence criterion via selection rate, before & after mitigation, when training on biased labels (left) and real labels (right).

Table 7. Observed and Real performance of model trained on less biased labels y_{truth}

Race	**Accuracy**		**AUC**		**Sensitivity**		**Specificity**	
Train with y_{truth}	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*	*Observed*	*Real*
African-American	0.6557	0.6759	0.6196	0.6421	0.4414	0.4473	0.9697	0.8370
Caucasian	0.7535	0.7207	0.6708	0.5729	0.3070	0.2204	0.9695	0.9253
Hispanic	0.7791	0.7423	0.6196	0.5233	0.1053	0.0714	0.9760	0.9752
Others	0.7744	0.7444	0.6166	0.5892	0.1875	0.2000	0.9208	0.9785

3.5 Model XAI – how training on biased labels affects explanations

Often, model explanations are used to get insights into the world – in this section we show how those insights can be corrupted if the model is trained on biased labels: We expect that a model that is easily able to recognize group from data, will learn to recalibrate predictions to match the label bias. As such, features that are important for predicting the group, could show up as predictive for the target. If the label bias is not known to the user, this could be incorrectly interpreted as group features being more predictive of the target than they actually are.

To show this effect, we re-train the model with the less biased labels y_{truth}. Before looking at differences in model explanations, let's first understand how the models trained on y_{bias} and y_{truth} differ. We first note that the selection rates of the models are rather different, see Fig. 3. As we already know from Table 1, the 2-year recidivism rates y_{truth} are much lower across groups than the corresponding decile scores y_{bias}, and as a consequence the selection rates decrease noticeably when training on y_{truth}. This also affects the model's predictive performance, see Table 7. In spite of under-predicting, however, the model still has predictive signal, in particular with respect to the observed labels – and we conclude that interpreting the model explanations is reasonable.

To generate explanations, we used the Gain-based Importance (Mean Decrease in Impurity) variable importance technique - $\left\{I = \frac{1}{M}\sum_T \sum_{g\in T, features=N} G\right\}$ where: M is the number of trees, T is the tree, g is the split in the tree, N is the features and G is the improvement in the loss function from the split. This was used because it gives

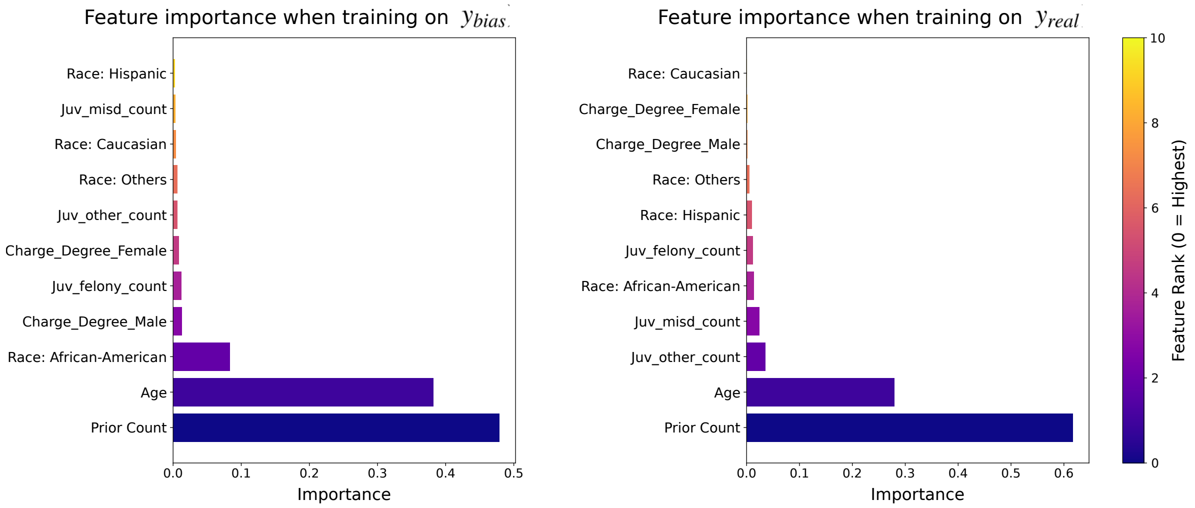

Fig. 4. Model XAI: Feature importance changes depending on whether the model is trained on the observed (biased) labels y_{bias} (left) or trained on the real (less biased) labels y_{truth}. While Prior Count and Age are important features for both scenarios, we see that when training on y_{bias}, being African American is a far more important feature than when training on y_{truth} – in the latter case, variables related to juvenile crime pay a far greater role. This highlights that model explanations are influenced by the status of the label used. In particular, we risk drawing incorrect conclusions about which features drive particular outcomes if we are not attentive to the effect of biased labels.

the contribution of the features to the model performance (e.g., Gini impurity, entropy or mse) attributed to each features.

In Figure 4, we see the clear effect of the training labels on explanations. Here, we show the GradientBoosting-Classifier feature importance generated by sklearn.ensemble software, across the two different scenarios: Training on the observed, biased labels y_{bias} compared to training on the less biased, underlying labels y_{truth}. In particular, we observe that the features **Race: African-American** is far less prominent when predicting the less biased 2-year recidivism y_{truth} than when predicting the biased COMPAS decile score y_{bias}. **Charge_degree_Female** but also **Charge_degree_Male** are less important for the unbiased model, whereas the juvenile counts **Juv_other_count**, **Juv_misd_count** and **Juv_felony_count** are all promoted to more important features.

4 Discussion

We have carried out a case study with a real example of observed (biased) and real (less biased) labels from the COMPAS dataset – but what did we learn?

4.1 Effect on bias diagnostics

As shown in Table 4, the observed fairness metrics that one would get by using the observed labels y_{bias} for bias diagnostics, consistently underestimate the bias with respect to both separation and sufficiency.

4.2 Effect on bias mitigation

Next, we observe what happens when using those same, observed but biased labels y_{bias} for mitigation. Tables 3 and 5 present the fairness metrics before and after mitigation. We found that fairness scores were significantly reduced after mitigation, regardless of whether they were computed with respect to the observed y_{bias} or the real y_{truth} . The selection rate reveals that mitigation reduces discrepancies observed before mitigation and artificially lowers the scores after mitigation compared to before mitigation, which might give a sense of security to the deprived or discriminated groups as by masking the observed bias before the mitigation. However, we still observe biases after mitigation. Indeed, Tables 5 and 6 show that mitigation makes the picture more complex: While we are underestimating bias (before and after mitigation) with respect to the separation criterion, we are now overestimating the bias with respect to the sufficiency criterion.

If, as developers, we navigate with respect to the observed labels y_{bias}, we might conclude that the mitigated model in Table 6 favors Caucasians for the separation and sufficiency criteria. However, when looking at the same metrics with respect to the real labels y_{true}, the conclusions are opposite. Moreover, when navigating with respect to the observed labels y_{bias}, we would conclude that after mitigation, the model has more than doubled its sufficiency bias scores per group – while in reality they have stayed quite fixed.

In short, it is clear that bias diagnostics with respect to the observed labels y_{bias} do not leave a truthful picture neither before nor after mitigation, and mitigation with respect to the observed – and only available – variable y_{bias} does not ensure fairness. Indeed, the observed metric for separation indicates that the model is much closer to achieving separation than it actually is with respect to the real labels.

4.3 Why does bias mitigation leave a more muddy picture?

The trend before mitigation was clear: The observed labels led to underestimation of bias. However, the trend was more complex after mitigation. This could be a result of the mitigation process enforcing conditions on the label to meet the desired outcome. This could be problematic: in the quest of enforcing the conditions to achieve the desired outcomes (i.e., balancing the model performance across groups using the protected attributes), there may be significant information loss leading to distinct after-mitigation model complexities.

To investigate the effect of mitigation further, we create a violin plot of the predicted probability scores of the model before and after mitigation, see Figure 5. We note that the African-Americans had an evenly distributed probability, including substantially higher probability scores before mitigation than the other groups, who have their main mode in the low probability range. Recall that the African-American group had a higher False Positive Rate than other groups before mitigation. After mitigation however, as shown, the mitigation technique forces the probability score for all groups to follow similar distributions, likely to achieve a more equal True Positive and False Positive Rates. While this might look great for the mitigation outcome, this suggests that mitigation may have corrected previous observed bias, but it is probably introducing a new bias for another group.

We can also view the mitigated model as overconfident: While the original model's probability scores followed a rather wide distribution, the probabilities post mitigation are distinctly bimodal for all groups. This overconfidence lends further insight into the reduced performance of the mitigated model, and raises the question whether the mitigated models actually are more trustworthy?

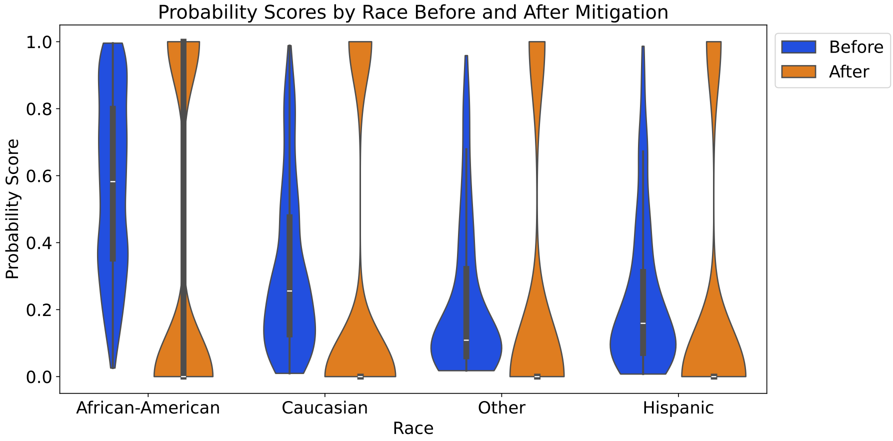

Fig. 5. The Probabilities predicted by the original and mitigated models show that mitigation distinctly changes the distribution of probabilities, across groups. This could be caused by the randomization for separation seeking to equalize false positives and -negatives.

4.4 Effect on XAI

We have also observed the effect of label bias on model explanations in Figure 4. The protected attribute (Race: African-American) was among the top three prominent features when the model was trained on the biased label y_{bias}, while considerably less prominent when training on the less biased label y_{truth}. Similarly, the gender of the offender Charge_Degree_Male was notable when training on y_{bias} and less prominent when training on y_{truth}. In short, we see that the label bias significantly impacts the explanations of the model, potentially leading to incorrect conclusions about which features lead to the outcome of interest.

5 Summary and conclusion

We have shown how label bias leads to incorrect bias diagnostics, incorrect bias mitigation, and changes in model explanations, which could lead to misinterpretation if not detected. Moreover, we have shown that when mitigating bias using randomization, we obtain overconfident models that are artificially tuned to make a certain amount of mistakes.

While all of this is problematic, it is not necessarily easy to avoid: In real life scenarios such as healthcare, welfare or criminal justice, the *true* label y_{truth} might not be available in any form.

Unless we can actually measure and correct the bias, we are left in a medicine-after-death (MAD) scenario because we are only able to (incorrectly) address the symptoms but not the root cause of the bias. Our submission does not imply or in any way downplay existing mitigation efforts, but we consider our case study an argument that addressing label bias should go beyond the after-effect mitigation. Consequently, we make the following recommendations:

First, transparency in model development and mitigation efforts is crucial to building trust and accountability. With this in mind, we encourage collaboration between researchers, practitioners (domain experts), and policy-makers to share knowledge and best practices in highlighting potential label bias to practitioners before model development and testing. This must happen across different domains of research areas to promote social and institutional well-being for all, regardless of race, gender or any unique attributes. This involves discarding data sets inflicted with historical biases and setting a benchmark for what an acceptable training set should be and should not be. However, this also involves being alert to those situations where obtaining a better dataset is not possible – e.g. in healthcare scenarios where correct diagnostic labels do not exist.

Second, efforts should go beyond after-effect mitigation and consider the broader context of fairness. This may involve developing new metrics and approaches to balance group disparity without introducing synthetic bias. It might mean a pre-training (or pilot) approach, continuous performance and fairness assessment in real-life applications, such as real-time updating of training data and adjusting models based on new insights, controversies and findings.

Finally, not all problems can be solved with existing tools – here, we have pointed to a weakness in the application of standard Trustworthy AI algorithms. This, however, does not mean that Trustworthy AI methods should not be used – it only means that we need to pay increased attention to *when* these methods can be trusted. In other words – even trustworthy AI algorithms can be applied incorrectly, leading to less trustworthy AI. To avoid such problems, future research should include automatic assessment of the robustness of Trustworthy AI tools – as well as further case studies on how these tools are affected under different real examples of label bias.

References

[1] Adewole S. Adamson. 2021. Racial Disparities in Dermatology: How Structural Racism Contributes to Underdiagnosis of Skin Diseases in Darker Skin Tones. *JAMA Dermatology* 157, 6 (2021). https://doi.org/10.1001/jamadermatol.2021.0484

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. ProPublica, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[4] Julio Arroyo, Pietro Perona, and Elijah Cole. 2023. Understanding Label Bias in Single Positive Multi-Label Learning. *ICLR 2023, Tiny Papers Track Computer Vision and Pattern Recognition*, 1–35. https://doi.org/10.48550/arXiv.2305.15584

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. *International Conference on Learning Representations (ICLR)* (2015). arXiv:1409.0473

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning Limitations and Opportunities* (2nd. ed.). Number 340 pp. The MIT Press, MIT, Cambridge, England.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint* (2018). arXiv:1810.01943

[8] Richard Berk, Hoda Heidari, Michael Kearns Shahin Jabbari, and Aaron Roth. 2023. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2023), 3–44. https://doi.org/10.1177/0049124118782533

[9] Sarah Bird, Roman Lutz, Miro Dudík, Harini Suresh, John Miller, Alexandra Antoine, Peter Friedlander, Seth Neel, Katrina Ligett, Richard Zemel, Jens Lundgaard, Michael Backes, and Giulia Fanti. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. arXiv:2009.14396 [cs.LG]

[10] Leo Breiman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. https://doi.org/0.1023/A:1010933404324

[11] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–21. https://doi.org/0.1023/A:1010933404324

[12] Sam CorbeŠ-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *In Proceedings of KDD '17,Halifax, NS, Canada* 10, 13–17. https://doi.org/10.1145/3097983.3098095

[13] Jessica Dai and Sarah M. Brown. 2020. Label Bias, Label Shift: Fair Machine Learning with Unreliable Labels. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada* 29 (2020), 1066–1077. chrome-extension: //efaidnbmnnnibpcajpcglclefindmkaj/https://dynamicdecisions.github.io/assets/pdfs/29.pdf

[14] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. 2018. User Profiling through Deep Multimodal Fusion. *WSDM '18: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 171 – 179. https://doi.org/10.1145/3159652.3159691

[15] Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? A study of label and selection bias. *Machine Learning* 112, 24 (2023), 5081–5104. https://doi.org/10.1007/s10994-023-06401-1

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259 – 268. https://doi.org/10.1145/2783258.278331

[17] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics* (2008). https://doi.org/10.1214/07-AOAS148

[18] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Proceedings of the Web Conference (WWW 2018)*. https://doi.org/10.48550/arXiv.1802.09548

[19] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference*. Springer, Springer.

[21] Heinrich Jiang and Ofir Nachum. 2020. Identifying and Correcting Label Bias in Machine Learning. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 108, 702–712. https://proceedings.mlr.press/v108/jiang20a.html

[22] Ronald C. Kessler and T. Bedirhan Üstün. 2004. The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research* 13, 2 (2004). https://doi.org/10.1002/mpr.168

[23] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. *31st Conference on Neural Information Processing Systems (NIPS 2017).* https://doi.org/10.48550/arXiv.1703.06856

[24] Yunyi Li, Maria De-Arteaga, and Maytal Saar-Tsechansky. 2024. Label Bias: A Pervasive and Invisibilized Problem. *Notice of the American Mathematical Society* 71, 8 (2024), 1066–1077. https://doi.org/10.1090/noti2941

[25] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS).* arXiv:1705.07874

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35. https://doi.org/10.1145/3457607

[27] Sarah E. Reese and Margarita Alegría. 2016. Disparities in Mental Health Treatment in U.S. Racial and Ethnic Minority Groups: Implications for Policy and Practice. *Psychiatric Services* 67, 5 (2016). https://doi.org/10.1176/appi.ps.201500175

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* https://doi.org/10.1145/2939672.2939778

[29] Boris Ruf and Marcin Detyniecki. 2022. A tool bundle for AI fairness in practice. *CHI Conference on human factors in computing systems extended abstracts*, 1–3. https://doi.org/10.1145/3491101.3519878

[30] Srinivasan H Sengamedu and Hien Pham. 2023. FAIRLABEL: Correcting Bias in Labels. *arXiv:2311.00638v1* 112, 24 (2023), 2311.00638v1.

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2017. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *International Conference on Learning Representations (ICLR).* arXiv:1312.6034

[32] Susan C. Taylor. 2020. Dermatologic Health Disparities: Challenges and Opportunities for Addressing Skin of Color. *Dermatologic Clinics* 38, 4 (2020). https://doi.org/10.1016/j.det.2020.05.001

[33] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research (JMLR)* (2008), 579–2605. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

[34] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. *In FairWare'18: IEEE/ACM International Workshop on Software Fairness*. https://doi.org/10.1145/3194770.3194776

[35] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law Technology* (2017). arXiv:1711.00399

[36] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES).* https://doi.org/10.1145/3278721.3278779

[37] Yixuan Zhang, Boyu Li, Zenan Ling, and Feng Zhou. 2024. Mitigating label bias in machine learning: Fairness through confident learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16917–16925.