You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods

Botty Dimanov[1] and **Umang Bhatt**[2] and **Mateja Jamnik**[3] and **Adrian Weller**[4]

Abstract. Transparency of algorithmic systems is an important area of research, which has been discussed as a way for end-users and regulators to develop appropriate trust in machine learning models. One popular approach, LIME [23], even suggests that model explanations can answer the question "Why should I trust you?". Here we show a straightforward method for modifying a pre-trained model to manipulate the output of many popular feature importance explanation methods with little change in accuracy, thus demonstrating the danger of trusting such explanation methods. We show how this explanation attack can mask a model's discriminatory use of a sensitive feature, raising strong concerns about using such explanation methods to check fairness of a model.

1 INTRODUCTION

There is great interest in transparency, or interpretability, as a way to aid our understanding of the inner workings of a machine learning model. One motivation is to ensure fairness as part of the 'Fair, Accountable, and Transparent' research agenda [8, 32]. Fairness is a key concern in many application areas including selecting candidates for hire or approving loans in banking. A popular family of approaches for transparency provide feature importance, or saliency, scores for a given input test case - the scores show how important each feature of the input was to the algorithm in coming to its conclusion. Indeed, a recent survey reports that these local saliency methods are the most popular approaches for transparency currently in practice [6].

It has been common to suggest that such saliency methods can be used to inspect a model for fairness as follows. We observe if a model's outputs depend significantly on a protected feature such as gender or race, which are termed *sensitive*. If there is high dependence on a sensitive attribute then the model appears to be unfair.

In this paper we show that the apparent importance of a sensitive feature does not reliably reveal anything about fairness of a model. We explain how this can happen and provide an instructive example demonstrating that a model could have arbitrarily high levels of unfairness across a range of popular measures, even while appearing to have zero dependence on the relevant sensitive feature. We introduce a practical approach to modify an existing model in order to downgrade the apparent importance of a sensitive feature to explanation methods. We demonstrate the success of our method empirically, with little change in model accuracy, while model unfairness can still remain high.

[1] University of Cambridge, United Kingdom, btd26@cam.ac.uk
[2] University of Cambridge, United Kingdom, usb20@cam.ac.uk
[3] University of Cambridge, United Kingdom, mateja.jamnik@cl.cam.ac.uk
[4] University of Cambridge and The Alan Turing Institute, United Kingdom, aw665@cam.ac.uk

Our observations raise serious concerns for organisations or regulators who hope to rely on feature importance interpretability methods to validate the fairness of models. We focus here on deep learning models, but our ideas extend naturally to other model classes.

2 RELATED WORK

There is a rapidly growing literature on *adversarial examples* [30], which considers how to fool *classification* accuracy by perturbing data points. Once a model has been well trained, it is possible to take a particular data point which was successfully classified, then change it just a tiny amount, in a particular way, such that the pretrained model now misclassifies the point with high confidence.

Later it was observed that many *explanation* methods are fragile with respect to small changes in a data point, even if the classification is unaffected [2, 16, 3]. It was shown that tiny adversarial perturbations to data inputs can be generated so that the classification remains unchanged, but the explanation returned is very different [11]. This was analysed in terms of the geometry of the learned function [9].

In this work we do not perturb the data. Instead, we modify the *model* in order to modify the explanations produced when common explanation methods are applied. In particular, our aim is to modify the model so that for any given data point, explanation methods will not show the sensitive feature as being important - even if in fact it is. Very recently, some works explored similar ideas. [22] examined how attention-based methods could be fooled. [15] showed that 'attention is not explanation', demonstrating that attention maps could be manipulated after training without altering predictions. [14] considered modifying vision models so that explanations could be controlled. [26] employed a 'scaffolding' construction specifically to fool Local Interpretable Model-Agnostic Explanations 'LIME' [23] and Shapley Values 'SHAP' [20] explanation methods.

We believe we are the first to focus on fairness of a model in relation to popular explanation methods. We describe our approach to modifying a model in order to hide unfairness in Section 3. We show in Section 4 how unfairness can be arbitrarily high, despite no dependence on a sensitive feature. In Section 5 we show empirically that our approach has little impact on a model's accuracy while being able to fool simultaneously many popular approaches to explanation: 1. Gradients [25], 2. Gradients $\times$ input [24], 3. Integrated Gradients [29], 4. Guided-backpropagation [28], 5. SHAP [20], and 6. LIME [23].

Our approach introduces an explanation loss term during training. This is similar to [17], who propose a loss function which enforces an L^1 penalty on the learned function gradient to ensure the final model has sparse explanations. However, we reduce the importance score for a particular specified target feature.

3 METHOD

Our approach consists of retraining an existing model to minimise a modified loss objective function: to the original loss we add an 'explanation loss' term, which is the gradient of the original loss with respect to a chosen target feature. Our attack method achieves three objectives: 1. We obtain a model with low local sensitivity to the chosen feature, yet with little loss in accuracy; 2. The low sensitivity generalises to unseen test points; and 3. Low feature sensitivity leads to low attribution for the target feature across all six feature importance explanation methods that we experimented with.

3.1 Notation

We consider differentiable functions $f : \boldsymbol{X} \mapsto \boldsymbol{Y}$, which map an input matrix in $\boldsymbol{X} \subseteq \mathbb{R}^{n\times m}$ with n samples and m features (attributes), to an output matrix in $\boldsymbol{Y} \subseteq \mathbb{R}^{n\times d}$, where each row is a 1-hot vector of softmax probabilities over d output classes. While our approach applies to arbitrary d, in this paper, we focus on $d = 2$ corresponding to a 'good' and 'bad' output classes (e.g., receive a loan or not). We write $\boldsymbol{x}^{(i)}$ for the input vector row i with m feature columns, and $\boldsymbol{X}_{:,j}$ for an entire feature j column vector. Aiming for readability, we allow for various number of points n to be processed, and may write $f(\boldsymbol{x})$ for the function evaluated on one input point $\boldsymbol{x}$. We write g for a local feature explanation function which take as input a model f and an input point of interest $\boldsymbol{x}$, and returns feature importance scores $g(f, \boldsymbol{x}) \in \mathbb{R}^m$, where $g(f, \boldsymbol{x})_j$ is the importance of (or attribution for) feature x_j for the model's prediction $f(\boldsymbol{x})$. We consider neural network functions f_θ parameterised by $\boldsymbol{\theta}$. Although some input features are categorical (e.g. male or female), as is standard, here we encode as numeric values and treat the variables as continuous.

3.2 Formulation

Suppose we have trained a model f_θ with acceptable performance but with undesirably high target feature explanations. We would like to find a **modified classifier** $f_{\theta+\delta}$, with the following properties:

1. *Model similarity:* the new model has similar performance
$$\forall i, \quad f_{\theta+\delta}(\boldsymbol{x}^{(i)}) \approx f_\theta(\boldsymbol{x}^{(i)}).$$
2. *Low target feature attribution:* the importance of the target feature j (e.g., gender or race), as given by a chosen explanation method g, decreases significantly
$$\forall i, \quad |g(f_{\theta+\delta}, \boldsymbol{x}^{(i)})_j| \ll |g(f_\theta, \boldsymbol{x}^{(i)})_j|.$$

3.3 Adversarial Model Explanation Attack

To manipulate the feature importance explanations, we initialise with a pre-trained model and then modify it by optimising with an extra penalty term, *explanation loss*, weighted by a hyperparameter α, which is normalised over all n training points (full batch):

$$\mathcal{L}' = \mathcal{L} + \frac{\alpha}{n} \left|\left|\nabla_{\boldsymbol{X}_{:,j}} \mathcal{L}\right|\right|_p, \tag{1}$$

where j is the index of the target feature that we want the model to appear to avoid using, and $\nabla_{\boldsymbol{X}_{:,j}}\mathcal{L}$ is the gradient vector of the original cross-entropy loss with respect to the entire feature column vector $\boldsymbol{X}_{:,j}$. We apply the L^p norm.[5] We define a new objective that regularises for low derivative with respect to the target feature across the training points, and results in the modified classifier, $f_{\theta+\delta}$. We outline the procedure in Algorithm 1, where we used $\tau = 100$ consistently since this was sufficient for convergence across runs. In all experiments we use $\alpha = 3$. We discuss varying α in Section 5.4.

[5] We use $p = 1$ since it led to rapid convergence and good results.

Algorithm 1 Learning a Modified Model with Concealed Unfairness

Input: Original classifier f_θ, target feature's index i, input matrix $\boldsymbol{X} \in R^{n\times m}$ with corresponding targets $\boldsymbol{y} \in \mathbb{R}^d$, and number of iterations τ.

Initialise $\boldsymbol{\delta} = \mathbf{0}$
for $t \in [0, \tau]$ iterations **do**
  Calculate the cross entropy loss $\mathcal{L}$ with respect to $f_{\theta+\delta}$
  Calculate the explanation loss
$$\zeta = \frac{\alpha}{n} \times L^p \left(\left[\left|\frac{\partial \mathcal{L}}{\boldsymbol{X}_{1,i}}\right|, \left|\frac{\partial \mathcal{L}}{\boldsymbol{X}_{2,i}}\right|, \ldots, \left|\frac{\partial \mathcal{L}}{\boldsymbol{X}_{n,i}}\right| \right] \right)$$
  Calculate the total loss $\mathcal{L}' = \mathcal{L} + \alpha \times \zeta$ (equation 1)
  Update model parameters with $\nabla_{\boldsymbol{\theta}}\mathcal{L}'$ using Adam
end for

Output: Modified classifier $f_{\theta+\delta}$

We clarify a difference between our approach for explanation loss and the recent method of [14]. While their approach takes the gradient of the one correct label element from the logits layer just before the softmax output, we take the gradient of the cross-entropy loss. Taking the gradient of the loss, rather than only the correct label element, contains extra information about the other classes, with the potential to improve generalisation across explanation methods and test points.

3.4 Fairness Metrics

In this paper we emphasise that an explanation method does not reliably reveal much about fairness of a model. A key question is then whether or not in fact the model is fair. We explore this using standard definitions from the literature [13, 5], evaluating the three fairness metrics below before and after learning the modified model. We consider model predictions for two primary sub-groups based on a sensitive feature, designating the sub-groups as privileged or unprivileged following [4] (e.g., gender males or females).

1. Demographic Parity (DP): the predicted *positive rates* for both groups should be the same.
2. Equal Opportunity (EQ): the *true positive rates (TPR)* for both groups should the same.
3. Equal Accuracy (EA): the classifier accuracy for both groups should be the same.

Note that it is typically not possible to satisfy these fairness notions simultaneously [18].

4 HOW EXTREME COULD UNFAIRNESS BE, YET STILL BE HIDDEN?

Here we consider the limits of how unfair a model might be, yet still appear to be fair according to explanation methods. Worryingly, and perhaps surprisingly, we show that in fact a model can be extremely unfair with respect to a feature, yet appear to have no sensitivity at all to the feature.

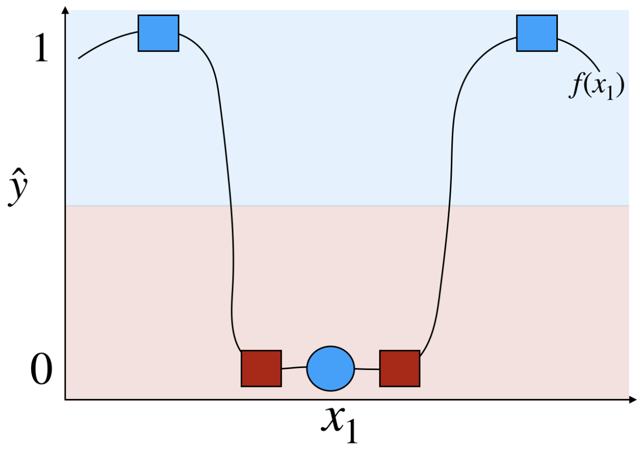

Figure 1: This example illustrates a function with no dependence on target feature yet extreme unfairness, showing the softmax predicted label $\hat{y}$ versus an input feature x_1, which is not the target feature. Each shape shown is a data point. The colour indicates the true label, i.e., blue means $y = 1$ and red means $y = 0$. The shape shows the value of the target feature: let square be male and circle be female. The black curve shows a function mapping from features to estimated output label $\hat{y}$. Assume the function is constant across gender. The blue circle is in the red zone, whereas it should be in the blue zone (see Section 4). Best viewed in colour.

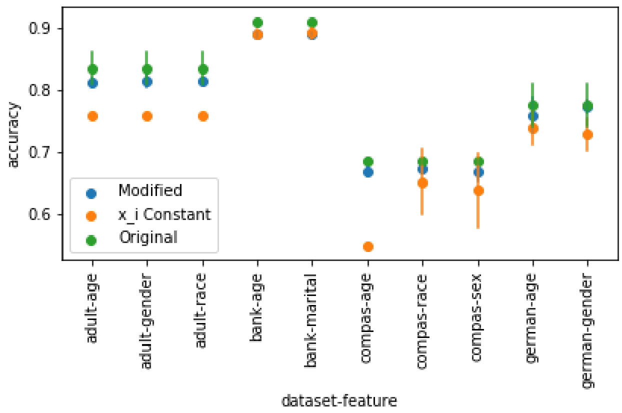

Figure 2: A comparison of accuracies of the modified model, a model trained with the target feature held at constant x_2, and the original model. Observe that across datasets and target features, our method achieves an accuracy comparable to the one of the original model and significantly higher than that of the constant model, demonstrating that the modified model is not merely ignoring the target feature. Results are averaged across 10 initialisations for a model with 5 hidden layers. Best viewed in colour.

Consider the situation shown in Figure 1. Each data point has two features: a continuous x_1 and a binary x_2. Let x_2 be a sensitive feature, such as gender, given by the shape of the point: assume circles are female, and squares are male. The true label y for each point is indicated by its colour: red for good and blue for bad. The black curve indicates the model's softmax predicted label value $\hat{y}$ as a function of the features (x_1, x_2). If above 0.5, then 1 is output, else 0 is output; this is shown by the pale blue/red boundary in background colour. Further, assume the model does not vary in the direction of x_2 (hence in particular has 0 gradient).

Five data points are shown. The model makes only one classification mistake (the blue circle receives $\hat{y} = 0$ yet has $y = 1$). However, this model is highly unfair with respect to the sensitive feature for all three metrics described in Section 3.4. Equal Opportunity is maximally violated: for female circles, $0/1 = 0\%$ deserving points get the good (blue) outcome; for male squares, $2/2 = 100\%$ deserving points get the good (blue) outcome. Equal Accuracy is also maximally violated: for female circles, $0/1 = 0\%$ points are accurate (blue circle should be placed in the blue zone); for male squares, $4/4 = 100\%$ points are accurate (correctly, blue squares are in the blue zone, red squares are in the red zone).

Finally, consider demographic parity (DP): for female circles, $0/1 = 0\%$ get the good outcome; for male squares, $2/4 = 50\%$ get the good outcome. Observe that if we keep adding more blue square data points near the ones already shown then the female ratio stays unchanged while the male ratio tends to 1, thus we can obtain any arbitrarily high level of DP unfairness.

Remark. Another way to view our example is that we have a model which by construction ignores the sensitive feature x_2. This is sometimes considered a form of process fairness via unawareness [7, 12]. It is known that even if a model cannot access a sensitive feature, it may still be unfair with respect to it – for example, the model might be able to reconstruct the sensitive feature with high accuracy from other features. This may lead one to wonder how our approach differs from simply removing the target feature.

The difference is that our approach attempts to learn a function which has very low derivative with respect to the sensitive feature at training points – hence, we might learn a function which varies significantly between the two possible sensitive feature settings yielding different outputs for male versus female. We explored this by comparing modified models learned with our approach against models where the sensitive feature was held constant (we did this, rather than simply remove the feature, in order to maintain model complexity). Accuracy results are shown in Figure 2, illustrating that our method attains higher accuracy. Further, see partial dependence plots in Figure 7.

5 RESULTS

Here we report and discuss empirical results of applying our adversarial model explanation attack.

5.1 Experimental Set-up

Datasets We conduct experiments on three datasets with sensitive features from the UCI machine learning repository [10] (adult (*Adult*) – gender, race; German credit (*German*) – age, gender; bank market (*Bank*) – age); and the dataset for Correctional Offender Management Profiling for Alternative Sanctions [19] (*COMPAS*) – gender, race, age.

Models For each dataset we train 0-5 hidden layer multilayer perceptrons (MLPs) with 100 units in each layer, regularised with a layer-wise L^2-norm penalty weighted by 0.03 for up to 1,000 epochs with early stopping and patience of 100 epochs with 10 random initialisations. We use L^2-norm regularisation because we want to avoid the regime of sparse weights. The penalty 0.03 was empirically validated to give the best validation accuracy. We use Tensorflow [1] to conduct the original optimisation with Adam [31], a global learning

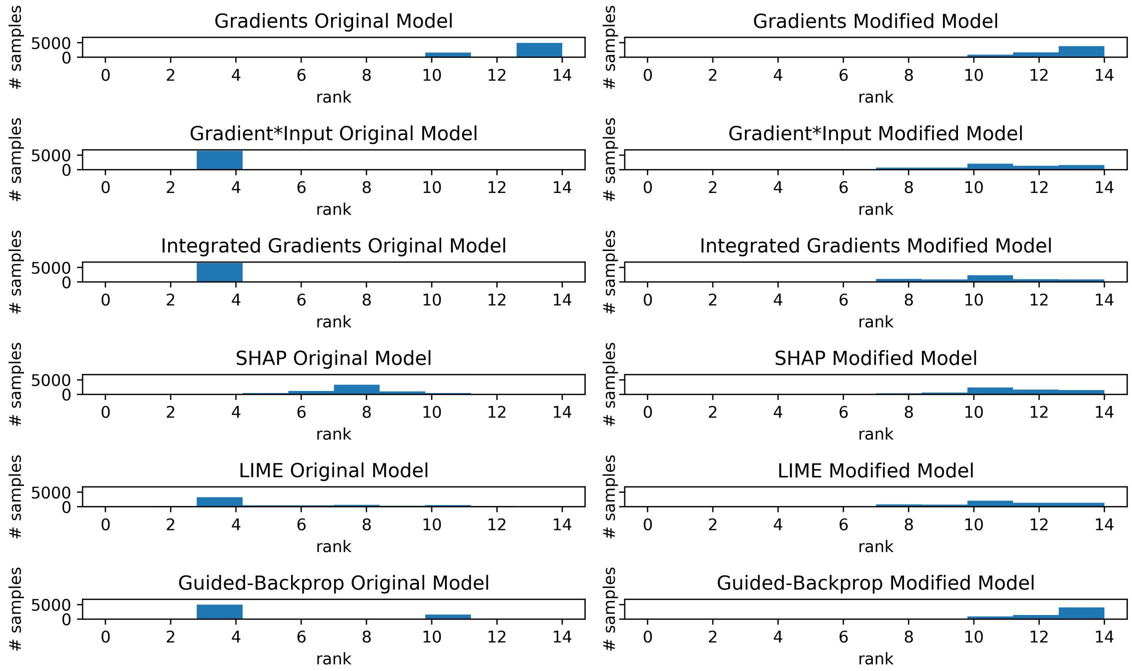

Figure 3: Importance ranking histograms for gender as the sensitive feature on adult test set of the original (left) and modified (right) models. Each histogram represents the ranking across the test set assigned by the designated feature importance method. A *higher ranking number* (further to the right) indicates *smaller feature importance*. Observe that the modified model has successfully shifted the ranking for all explanation methods.

rate of 0.01 and 0.005 learning rate decay over each update and with full batch gradient descent.

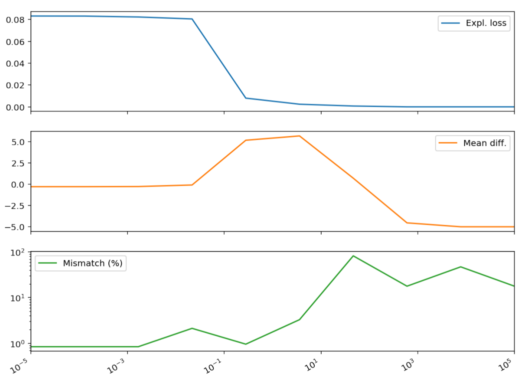

Figure 4: Effect of $\alpha \in [10^{-5}, 10^{-5}]$ in applying our explanation attack to the adult dataset and gender target feature on the model similarity and low target feature attribution metrics (y-axis): (top) average explanation loss per sample (Expl. loss); (middle) the mean of the sensitive property importance ranking distribution (Mean diff.); and (bottom) the percentage difference between the two models' predictions (Mismatch). Notice that optimal α values lie in the range $[10^{-1}, 10^{1}]$.

Feature Attribution Methods We evaluate six popular feature attribution methods: Sensitivity analysis gradients [25] (**Grads**), Gradients $\times$ input [24] (**GI**), Integrated Gradients [29] (**IG**), an approximation of Shapley values Expected Gradients [20] (**SHAP**), Local Interpretable Model-Agnostic Explanations [23] (**LIME**), and Guided-backpropagation [28] (**GB**). We conceal unfairness using the training data and report evaluations both on the training data, and on a test set that was used neither for training the original model, nor for the modified model.

Fairness For the fairness evaluation, we use the implementation of IBM AI360 Toolkit [4] and we binarise each sensitive features in the following fashion: Gender: Male - privileged, Female - unprivileged; Age: $25 > \mathrm{x}$ privileged, $25 < \mathrm{x}$ unprivileged; Race: White - privileged, Non-white - unprivileged; Martial status: Single - privileged, Not single - unprivileged.

5.2 Evaluation Criteria

5.2.1 Attack

We consider the concealing procedure successful when both properties from Section 3.2 are well satisfied. We measure **model similarity** between the modified model and the original model through three metrics:

- **Loss diff.**: Difference between the categorical cross entropy losses ($\mathcal{L}$) of both models averaged over all test points.
- **Acc. diff.**: Difference in the accuracy of both models.
- **Mismatch (%)**: Difference in the output of the two models, as measured by the percentage of datapoints, where the predictions of the two models differ.

Measuring the effect of the concealing procedure on feature importance is more complex. We want to avoid the pathological case of the attack shrinking the importance of all features and inducing a random classifier. Therefore, we introduce four metrics for interpretation dissimilarity based on the relative importance ranking of a

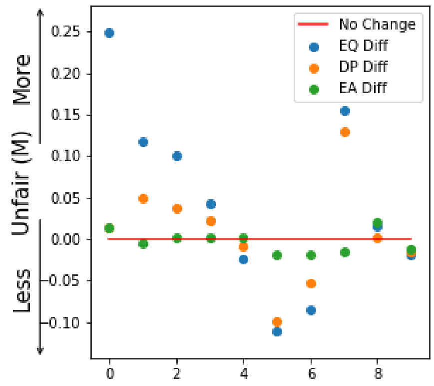

Figure 5: Evaluation of the impact our explanation attack has on unfairness (higher indicates more unfairness). We show three fairness metrics across 4 datasets and their sensitive features, averaged over 6 model complexities (number of hidden layers) and 10 random initialisations. We find no consistent pattern of impact, though Equal Opportunity (EQ) appears the most variable.

feature. Figure 3 illustrates the ranking histogram of the relative feature importance, which describes the ranking probability mass distribution of the explanation. We show a case where the initial model had low target feature gradient, demonstrating that even in this case, the attack was successful. An effective attack shifts the distribution from left to right. We use five metrics to measure attack susceptibility through this shift:

- **Top k:** the number of datapoints where the sensitive feature received rank k or above.
- **Mode shift:** the difference between the modes of the distribution (*Avg. #shifts*).
- **Mean shift:** the difference between the means.
- **Highest rank:** the highest rank that the sensitive feature received across all datapoints.
- **Highest ranking datapoints (HRD):** the number of datapoints where the sensitive feature received the highest rank. This is the same as Top k, where k = highest rank.

5.3 Low Target Feature Attribution

Figure 3 illustrates three important points. First, out method significantly decreases the relative importance of the target feature, effectively making it the least important of all features. Second, the attack transfers across six different explanation methods. Third, the attack generalises for unseen, held-out test datapoints.

Transferability Tables 1 and 2 illustrate that the explanation attack transfers across explanation methods.

The attack transfers to other gradient-based explanation methods and significantly decreases the importance for 6 other explanation methods (Gradients * Input, Integrated Gradients, Guided-backpropagation, LIME, SHAP).

Notice in Table 1 that in the case of the Adult dataset and gender target feature for GI, IG, GB and LIME, the explanation attack has moved down the target feature importance out of the Top ranking features for thousands of data points, demonstrating that the attack works even when the target feature has high relative importance.

Generalisation The generalisation of the attack to test points is noteworthy since we might expect that the decision boundary would be perturbed locally around the training points to affect only their explanations, without significant change for test points, especially if far away in feature space. We investigate this hypothesis in Section 5.6.

Further, Table 2 confirms that the attack generalises across datasets and features since it is capable of shifting the importance ranking distribution considerably for a total of 10 features over 4 datasets. The table indicates that the test values for both the model similarity and low target feature attribution are either similar or lower.

5.4 Hyper-parameter Investigation

Explanation Loss Norm We observe that the L^1-norm converged slightly faster and to slightly better configurations both in terms of model similarity and low target feature attribution metrics across different settings in comparison to both the L^2 and L^∞ norms. One possible explanation for this result is that the gradient of the L^1-norm explanation is consistent across the explanations for all datapoints, whereas the L^2-norm explanation loss penalises datapoints with large magnitudes for the target feature importance $\left(\left|\frac{\partial \mathcal{L}}{\partial \mathbf{X}_{i,j}}\right|\right)$ explanations. However, the L^2-norm does not affect as many datapoints with relatively small values of the target feature importance. The effect is emphasised with the L^∞ norm, where the gradient is non-zero only for the point with the highest value explanation. Hence, training with L^∞ norm resembles a single sample gradient descent and results in significantly slower convergence. Further, we observed that the choice of the explanation loss norm is strongly coupled with the value of the explanation penalty term α. All three norms converge to very similar configurations with the appropriate α. Since the L^2-norm over emphasises extremely high value explanations, it requires a lower α. This is in contrast to L^∞-norm, which reflects the loss of a single example and requires an α of orders of magnitude higher than the L^1-norm.

Explanation Loss Weight α Figure 4 demonstrates that the learning dynamics of the adversarial explanation attack vary with the explanation penalty term α. Observe that at one extreme, the penalty term α corresponds to unnoticeable changes in the explanation loss (first sub-figure), while at the other to a catastrophic change that leads to a constant model which ignores all features and drastically changes the model predictions (sub-figure three). Within the optimum range ($\alpha \in [10^{-1}, 10^1]$), we can minimise the explanation loss significantly while keeping the model prediction dissimilarity relatively low. We set $\alpha = 3$ for all experiments.

Learning algorithm We tried various parameter learning approaches and observed that it could make a significant difference. Similarly to regular training, adaptive learning rate algorithms achieve significantly better results. A vanilla-SGD optimisation is much more likely to converge to constant classifiers that predict the label distribution and requires bespoke learning rate scheduling routines similar to [27], where the learning rate is adopted dynamically based on the explanation loss. In all experiments, we used Adam [31].

5.5 Fairness Evaluation

Figure 6 demonstrates that overall the modified model can hide a sensitive feature in such a way that the modified model would appear

	Mode (O)	Mode (M)	# shifts	Mean (O)	Mean (M)	Mean Diff	Top Rank(O)	Top Rank(M)	#Top_O (O)	#Top_O (M)	Top-5 (O)	Top-5 (M)	Top-1 (O)	Top-1 (M)
Gradients	13	13	0	12.28	12.36	0.08	10	4	1571	322	0	1	0	0
Gradient*Input	4	13	9	4.00	10.81	6.81	4	2	6513	12	6513	12	0	0
Integrated Gradients	4	11	7	4.00	10.17	6.17	4	3	6513	36	6513	36	0	0
SHAP	7	12	5	7.47	11.16	3.69	3	4	1	0	51	1	0	0
LIME	4	13	9	5.71	10.69	4.98	0	2	13	0	3642	17	13	0
Guided-Backprop	3	13	10	4.93	12.37	7.44	3	6	4942	0	4942	0	0	0

Table 1: Evaluation of model similarity & low target feature attribution after an adversarial explanation attack for five explanation methods on Adult Gender Train ('O' is original model, 'M' is modified model). Notice that the mode and mean ranking of the sensitive feature increases after our attack. For nearly all datapoints, the sensitive feature moves out of the top five most important features.

Dataset	Feature	Train Exp Loss	Test Exp Loss	Train Acc Δ	Test Acc Δ	Train Mismatch (%)	Test Mismatch (%)
adult	age	9.79e-3 $\pm$ 3.61e-3	9.82e-3 $\pm$ 3.59e-3	-2.76 $\pm$ 1.03	-3.07 $\pm$ 1.16	10.88 $\pm$ 1.67	10.72 $\pm$ 1.66
	gender	1.10e-2 $\pm$ 3.36e-3	1.11e-2 $\pm$ 3.38e-3	-2.43 $\pm$ 0.86	-2.71 $\pm$ 0.94	10.37 $\pm$ 2.44	10.29 $\pm$ 2.49
	race	1.01e-2 $\pm$ 2.75e-3	1.02e-2 $\pm$ 2.76e-3	-2.47 $\pm$ 0.85	-2.78 $\pm$ 0.9	10.24 $\pm$ 1.31	10.37 $\pm$ 1.35
bank	age	1.28e-2 $\pm$ 4.12e-3	1.34e-2 $\pm$ 4.17e-3	-1.81 $\pm$ 0.35	-2.23 $\pm$ 0.4	7.35 $\pm$ 0.73	7.5 $\pm$ 0.75
	marital	1.25e-2 $\pm$ 5.26e-3	1.3e-2 $\pm$ 5.46e-3	-1.73 $\pm$ 0.34	-2.27 $\pm$ 0.4	7.25 $\pm$ 0.71	7.43 $\pm$ 0.7
compas	age	4.00e-3 $\pm$ 1.69e-3	4.34e-3 $\pm$ 1.82e-3	-2.23 $\pm$ 0.66	-3.2 $\pm$ 0.91	19.83 $\pm$ 1.68	18.96 $\pm$ 1.6
	race	3.40e-3 $\pm$ 1.9e-3	3.62e-3 $\pm$ 1.97e-3	-1.54 $\pm$ 0.75	-2.7 $\pm$ 0.87	18.85 $\pm$ 2.48	18.38 $\pm$ 2.82
	sex	3.01e-3 $\pm$ 1.53e-3	3.20e-3 $\pm$ 1.59e-3	-1.9 $\pm$ 0.83	-2.78 $\pm$ 0.99	19.46 $\pm$ 2.85	18.39 $\pm$ 3.02
german	age	1.77e-3 $\pm$ 1.34e-3	1.82e-3 $\pm$ 1.43e-3	-7.38 $\pm$ 6.38	-5.83 $\pm$ 6.6	18.59 $\pm$ 10.33	17.72 $\pm$ 10.25
	gender	2.21e-3 $\pm$ 1.31e-3	2.24e-3 $\pm$ 1.38e-3	-6.07 $\pm$ 3.27	-4.21 $\pm$ 4.01	17.14 $\pm$ 4.84	15.88 $\pm$ 4.87

Table 2: Summary of model similarity and low target feature attribution metrics over four **train** and **test** datasets and six features averaged over all complexities. We find that the explanation loss for **both** the train and test sets is low. Also the change in accuracy between the original and modified model over the train and test set is similar. These results suggest that our attack is successful in generalising across unseen test points.

fair using local-sensitivity explanation techniques, yet actually could become more or less unfair. The low local-sensitivity can result in a decision boundary that varies irrespective of the sensitive feature values, such as the one illustrated in Figure 1. We investigate the effects of the adversarial explanation attack on decision boundary in Section 5.6.

We run further experiments across model complexities and different initialisations. Figure 5 shows that the adversarial explanation attack does not have a consistent impact on the fairness metrics, despite the fact that the apparent importance of the feature is negligible. The attack causes the resulting model to have unpredictable unfairness behaviour, becoming more unfair for some features, less unfair for others, or maintains a relatively similar fairness levels to the original model. The unpredictability of the unfairness argues strongly against relying solely on transparency to verify model fairness.

Nevertheless, in most cases, the fairness metrics are affected similarly in the sense that if one of the models becomes more unfair according to one metric, most of the remaining metrics vary accordingly. One possible explanation for the inconsistent behaviour of the fairness metrics after the attack could the presence of confounding factors. Although the explanatory importance of a feature could be low, the model might have learned to rely on other features, which could be used to infer the target feature (e.g., someone's marital status of a husband or wife can be used to infer their gender). Another possibility is that the adversarial explanation attack results in a model that: a) effectively keeps the same model, but flattens the derivatives to make it locally insensitive to a feature; or b) ignores the feature altogether. We discussed evidence in favour of a) over b) in Section 4. Further, Figure 6 shows that the unfairness of our modified model does not match that of a model which simply ignores the target feature.

5.6 Decision Boundary: How much does the model really change?

Despite the significant changes in explanation, our results suggest that the model has not changed significantly. This is demonstrated by the small number of mismatches shown in Table 2, and the small change to the decision boundary, as illustrated in Figure 8. However, Figure 7 shows that the model can change significantly with respect to the target attribute.

6 CONCLUSION AND FUTURE WORK

We demonstrated that many popular explanation methods used in real-world settings are not able to indicate reliably whether or not a model is fair. We provided an intuitive explanation to show how this can happen. We introduced a method to modify an existing model and showed its empirical success in downgrading the feature importance of key sensitive features across six explanation methods and

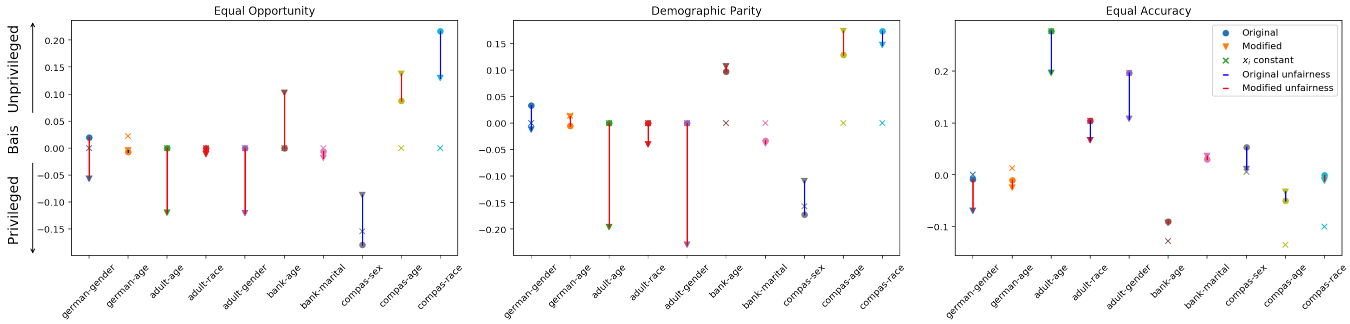

Figure 6: Unfairness across 3 metrics: Equal Opportunity, Demographic Parity and Equal accuracy. We find no consistent pattern. To some extent, we see that the unfairness with respect to Equal Opportunity is higher for the original model and behaves similarly to removing the feature. Similarly for demographic parity, we find that the modified model is less biased than the original model with respect to the sensitive feature. Equal accuracy (of subgroups between both models) was least affected by our attack.

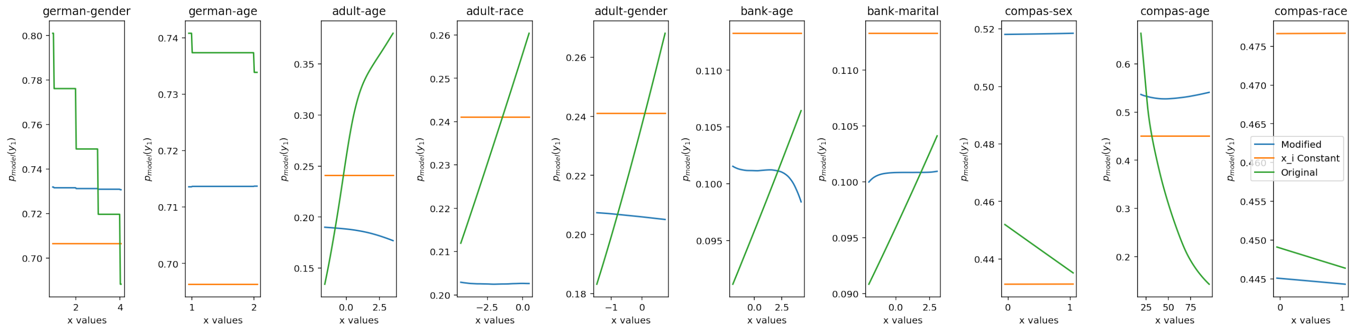

Figure 7: Partial dependence plots showing how the predicted output varies according to the sensitive feature shown. Results shown are for 5 hidden layers. Best viewed in colour.

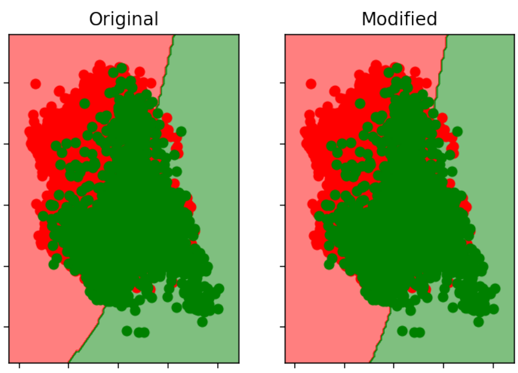

Figure 8: Comparison of the decision boundary between the original (left) and modified (right) classifier after an adversarial explanation attack on Adult capital gains (most important feature) in 2D reduced input space. We use scikit-learn [21] to implement PCA. Red and green backgrounds indicate negative and positive prediction, respectively. Notice that the boundary is slightly modified in the lower end within a region where there are few datapoints. The circles represent the 2D projections of each point in the training and the test set, while their colour indicates the true label.

unseen test points across four datasets, while having little effect on model accuracy.

Our work raises concerns for those hoping to rely on such explanation methods to measure or enforce standards of fairness. For example, a trained loan scoring system might be unfair with respect to a sensitive feature such as gender. However, the model's parameters might be modified in such a way that a feature importance explanation could falsely suggest that the output does not depend on this sensitive feature. If transparency methods are to be used, we argue for rigorous tests of robustness to understand and control the extent to which they can be manipulated.

There are many interesting questions to explore in future work. How might the explanation attack be refined (e.g., to explore its performance if extended in the natural way to be used against multiple target variables), and how might it be well defended against? One could further explore how the attack relates to the dataset, to the model class, to the explanation method, and the difference between the model's representational capacity and the dataset's complexity.

Acknowledgements

AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI). UB acknowledges support from the CFI. BD acknowledges support from EPSRC Award 1778323.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., 'Tensorflow: A system for large-scale machine learning', in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, (2016).

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim, 'Sanity checks for saliency maps', in *Advances in Neural Information Processing Systems*, pp. 9505–9515, (2018).

[3] David Alvarez-Melis and Tommi S Jaakkola, 'Towards robust interpretability with self-explaining neural networks', in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 7786–7795. Curran Associates Inc., (2018).

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

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi, 'Data decisions and theoretical implications when adversarially learning fair representations', *CoRR*, **abs/1707.00075**, (2017).

[6] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley, 'Explainable machine learning in deployment', *arXiv preprint arXiv:1909.06342*, (2019).

[7] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell, 'Fairness under unawareness: Assessing disparity when protected class is unobserved', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 339–348. ACM, (2019).

[8] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen, 'Principles for accountable algorithms', (2018).

[9] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Mller, and Pan Kessel. Explanations can be manipulated and geometry is to blame, 2019.

[10] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[11] Amirata Ghorbani, Abubakar Abid, and James Zou, 'Interpretation of neural networks is fragile', *AAAI*, (2019).

[12] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller, 'Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning', in *Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[13] Moritz Hardt, Eric Price, and Nati Srebro, 'Equality of opportunity in supervised learning', in *Advances in Neural Information Processing Systems (NeurIPS)*, (2016).

[14] Juyeon Heo, Sunghwan Joo, and Taesup Moon, 'Fooling neural network interpretations via adversarial model manipulation', *arXiv preprint arXiv:1902.02041*, (2019).

[15] Sarthak Jain and Byron C Wallace, 'Attention is not Explanation', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, (2019).

[16] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim, 'The (un) reliability of saliency methods', in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 267–280, Springer, (2019).

[17] Keisuke Kiritoshi, Ryosuke Tanno, and Tomonori Izumitani, 'L1-norm gradient penalty for noise reduction of attribution maps', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, (June 2019).

[18] Jon Kleinberg, 'Inherent trade-offs in algorithmic fairness', in *ACM SIGMETRICS Performance Evaluation Review*, volume 46, pp. 40–40. ACM, (2018).

[19] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the COMPAS recidivism algorithm, Mar 2019.

[20] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems*, pp. 4765–4774, (2017).

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, **12**, 2825–2830, (2011).

[22] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. Learning to deceive with attention-based explanations, 2019.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why should I trust you?: Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, (2016).

[24] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje, 'Not just a black box: Learning important features through propagating activation differences', *arXiv preprint arXiv:1605.01713*, (2016).

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, 'Deep inside convolutional networks: Visualising image classification models and saliency maps', *arXiv preprint arXiv:1312.6034*, (2013).

[26] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju, 'How can we fool LIME and SHAP? Adversarial attacks on post hoc explanation methods', *arXiv preprint arXiv:1911.02508*, (2019).

[27] Leslie N. Smith, 'A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay', *CoRR*, **abs/1803.09820**, (2018).

[28] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, 'Striving for simplicity: The all convolutional net', *arXiv preprint arXiv:1412.6806*, (2014).

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, 'Axiomatic attribution for deep networks', in *International Conference on Machine Learning (ICML)*, (2017).

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, 'Intriguing properties of neural networks', *arXiv preprint arXiv:1312.6199*, (2013).

[31] Tijmen Tieleman and Geoffrey Hinton, 'Lecture 6.5-rmsprop, coursera: Neural networks for machine learning', *University of Toronto, Technical Report*, (2012).

[32] Adrian Weller, 'Transparency: motivations and challenges', in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 23–40, Springer, (2019).