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

Botty Dimanov\textsuperscript{1} and Umang Bhatt\textsuperscript{2} and Mateja Jamnik\textsuperscript{3} and Adrian Weller\textsuperscript{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 \cite{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 \cite{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 \cite{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.

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 \cite{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 \cite{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 \cite{11}. This was analysed in terms of the geometry of the learned function \cite{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. \cite{22} examined how attention-based methods could be fooled. \cite{15} showed that ‘attention is not explanation’, demonstrating that attention maps could be manipulated after training without altering predictions. \cite{14} considered modifying vision models so that explanations could be controlled. \cite{26} employed a ‘scaffolding’ construction specifically to fool Local Interpretable Model-Agnostic Explanations ‘LIME’ \cite{23} and Shapley Values ‘SHAP’ \cite{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 \cite{25}, 2. Gradients × input \cite{24}, 3. Integrated Gradients \cite{29}, 4. Guided-backpropagation \cite{28}, 5. SHAP \cite{20}, and 6. LIME \cite{23}.

Our approach introduces an explanation loss term during training. This is similar to \cite{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 \( f : \mathbb{X} \rightarrow \mathbb{Y} \), which map an input matrix in \( \mathbb{X} \subseteq \mathbb{R}^{n \times m} \) with \( n \) samples and \( m \) features (attributes), to an output matrix in \( \mathbb{Y} \subseteq \mathbb{R}^{n \times d} \), 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 \( x^{(i)} \) for the input vector row \( i \) with \( m \) feature columns, and \( X \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \...
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
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 $\alpha$ 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 x 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: $25 > x$ privileged, $25 < x$ 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 ($L$) 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 $k = \text{highest rank}$.

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, hold-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 $L^\infty$ norms. One possible explanation for this result is that the gradient of the $L^1$-norm explanation is consistent across the explanations for all data-points, whereas the $L^2$-norm explanation loss penalises datapoints with large magnitudes for the target feature importance $\left(\frac{\alpha c_i}{\sum c_i}\right)$ explanations. However, the $L^\infty$-norm does not affect as many data-points with relatively small values of the target feature importance. The effect is emphasised with the $L^\infty$ norm, where the gradient is non-zero only for the point with the highest value explanation. Hence, training with $L^\infty$ 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 $\alpha$. All three norms converge to very similar configurations with the appropriate $\alpha$. Since the $L^\infty$-norm over emphasises extremely high value explanations, it requires a lower $\alpha$. This is in contrast to $L^\infty$-norm, which reflects the loss of a single example and requires an $\alpha$ of orders of magnitude higher than the $L^1$-norm.

**Explanation Loss Weight** $\alpha$ Figure 4 demonstrates that the learning dynamics of the adversarial explanation attack vary with the explanation penalty term $\alpha$. Observe that at one extreme, the penalty term $\alpha$ 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 ($\alpha \in [10^{-1}, 10^3]$), 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
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 be 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

---

**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.

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

---

**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.

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

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.