You shouldn’t trust me: Models can be Learned to Conceal Unfairness from Explanation Methods

Botty Dimanov, Umang Bhatt, Prof. Mateja Jamnik and Dr. Adrian Weller

AAAI-20 SafeAI Workshop
Feature importance reveals nothing reliable about model fairness.

- Instructive example
- Practical approach to modify a pre-trained model:
  - Low feature importance
  - High feature unfairness
  - Little change in accuracy
Why do we care?

Model A

Feature Importance

Feature

Model B

Feature Importance

Feature
Why do we care?

Model A

Model B

Feature Importance

Feature

Feature Importance

Feature
What is fair?

Demographic Parity
positive rate

50%

Equal Opportunity
ture positive rate

100%

Equal Accuracy
true positive +
true negative

100%
Illustrative Example

\[ \frac{\partial f(x_1, x_2)}{\partial x_2} = 0 \]

\[ x_1 \in R \quad (\circ \circ \circ) \]

\[ x_2 \in \{0, 1\} \quad (\bullet \), \quad (\bullet) \]

\[ y \in \{0, 1\} \quad \text{[Orange, Blue]} \]

\[ f(x_1, x_2) \]

\[ 0 \quad 1 \]

\[ \hat{y} \]

\[ x_1 \quad (\circ \circ \circ) \]
Illustrative Example

\[ \frac{\partial f(x_1, x_2)}{\partial x_2} = 0 \]

\( x_1 \in \mathbb{R} \)
\( x_2 \in \{0,1\} \)
\( y \in \{0,1\} \)

Demographic Parity

\[ 50\% \]

\[ 0\% \]
Illustrative Example

\[ \frac{\partial f(x_1, x_2)}{\partial x_2} = 0 \]

\( x_1 \in \mathbb{R} \)

\( x_2 \in \{0,1\} \)

\( y \in \{0,1\} \)

\( y \in \{0,1\} \)

\[ f(x_1, x_2) \]

\[ \hat{y} \]

Equal Opportunity

100%

0%
Illustrative Example

\[ \frac{\partial f(x_1, x_2)}{\partial x_2} = 0 \]

\( x_1 \in R \)

\( x_2 \in \{0,1\}, \quad y \in \{0,1\}, \quad \hat{y} \in \{0,1\} \)

Equal Accuracy

\[ 100\% \quad + \quad 0\% \]
Can we manipulate explanations?

**Modified explanations** via adversarial perturbations of **inputs**
- (Ghorbani, Abid, and Zou 2019)
- (Dombrowski et al. 2019)
- (Slack et al. 2019)

**Control visual explanations** via adversarial perturbations of **parameters**
- (Heo, Joo, and Moon 2019)

**Downgrade explanations via adversarial perturbations of parameters to hide unfairness**
Method

Classifier: \( f : X \mapsto Y \)

Explanation: \( g(f, x)_j \)

\[
\begin{align*}
\hat{f}_\theta & \rightarrow f_{\theta+\delta} \\
\text{Model Similarity:} & \forall i, \; f_{\theta+\delta}(x^{(i)}) \approx f_\theta(x^{(i)}) \\
\text{Low target feature} & \forall i, \; |g(f_{\theta+\delta}, x^{(i)})_j| \ll |g(f_\theta, x^{(i)})_j|
\end{align*}
\]
Adversarial Explanation Attack

\[ \argmin_{\delta} L' = L(f_{\theta+\delta}, x, y) + \frac{\alpha}{n} \left\| \nabla_{x \neq j} L(f_{\theta+\delta}, x, y) \right\|_p \]
Experimental Set-up

Datasets

1. Adult — age, gender, race;  
   (Dua and Graff 2017)
2. German credit — age, gender;  
   (Dua and Graff 2017)
3. Bank market — age, marital;  
   (Dua and Graff 2017)
4. COMPAS — gender, race, age  
   (Larson et al. 2019).

Feature Importance Explanation Methods

1. Sensitivity analysis (Simonyan, Vedaldi, and Zisserman 2013),  
2. Gradients × input (Shrikumar et al. 2016),  
3. Integrated Gradients (Sundararajan, Taly, and Yan 2017),  
4. Guided-backpropagation (Springenberg et al. 2014)  
5. Shapley values (Expected Gradients) (Lundberg and Lee 2017),  
6. LIME (Ribeiro, Singh, and Guestrin 2016)
Results (Importance Ranking)

\[ f_{\theta} \]

Gradients Original Model

Gradient*Input Original Model

Integrated Gradients Original Model

SHAP Original Model

LIME Original Model

Guided-Backprop Original Model

\[ f_{\theta+\delta} \]

Gradients Modified Model

Gradient*Input Modified Model

Integrated Gradients Modified Model

SHAP Modified Model

LIME Modified Model

Guided-Backprop Modified Model
The adversarial explanation attack:
1. Significantly decreases relative importance.
2. Generalises to test points.
3. Transfers across explanation methods.
### Results (Model Similarity)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Feature</th>
<th>Train $\zeta$ (10$^{-2}$)</th>
<th>Test $\zeta$ (10$^{-2}$)</th>
<th>Train Acc $\Delta$</th>
<th>Test Acc $\Delta$</th>
<th>Train Mismatch (%)</th>
<th>Test Mismatch (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>age</td>
<td>9.79±3.61</td>
<td>9.82±3.59</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>11.03±3.36</td>
<td>11.11±3.38</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>10.1±2.75</td>
<td>10.18±2.76</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>bank</td>
<td>age</td>
<td>12.79±4.12</td>
<td>13.39±4.17</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>12.5±5.26</td>
<td>12.96±5.46</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>compas</td>
<td>age</td>
<td>4.0±1.69</td>
<td>4.34±1.82</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>3.4±1.9</td>
<td>3.62±1.97</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>3.01±1.53</td>
<td>3.2±1.59</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>german</td>
<td>age</td>
<td>1.77±1.34</td>
<td>1.82±1.43</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>2.21±1.31</td>
<td>2.24±1.38</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>
How much does the model change?

Original Model: \( f_\theta \)

Modified Model: \( f_{\theta+\delta} \)

Constant Model: \( f_\theta : X_{., j} = 0 \)
How much does the model change?

\[ f_\theta \quad f_\theta : x_{-j} = 0 \quad f_\theta + \delta \]

Partial Dependence Plots
Does the model ignore the target feature?

Original Model: $f_{\theta}$

Modified Model: $f_{\theta} + \delta$

Constant Model: $f_{\theta}: x_{:,j} = 0$
Fairness
Fairness via Unawareness?

\[ f_\theta \quad \text{more unfair} \quad f_{\theta+\delta} \quad \text{more unfair} \]
Fairness via Unawareness?

\[ f_\theta \quad \text{more unfair} \quad f_{\theta+\delta} \quad \text{more unfair} \]

\[ f_{\theta+\delta} \]

\[ f_\theta : \mathbf{x}_{-j} = 0 \]
Conclusions

Feature importance reveals nothing reliable about model fairness.

Main Contributions:
- Practical approach modify a pre-trained model:
  - Low feature importance across 6 explanation methods, 4 datasets, 10 features
  - High feature unfairness across 3 fairness metrics
  - Little change in accuracy, but the difference in outputs is detectable
Conclusions and Future Work

Feature importance reveals nothing reliable about model fairness.

Practical approach modify a pre-trained model:
- Low feature importance across 6 explanation methods, 4 datasets, 10 features
- High feature unfairness across 3 fairness metrics
- Little change in accuracy, but the difference in outputs is detectable

Main Contributions:

Future Work:

1. Increase model similarity

\[ L' = L(f_{θ+δ}, x, y) + \frac{α}{n} \left\| \nabla_{X_{-j}} L(f_{θ+δ}, x, y) \right\|_p \]

2. Investigate relation to model capacity and dataset complexity