Characterizing the risk of fairwashing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fairwashing refers to the risk that an unfair black-box model can be explained by a fairer model through post-hoc explanations' manipulation. However, to realize this, the post-hoc explanation model must produce different predictions than the original black-box on some inputs, leading to a decrease in the fidelity imposed by the difference in unfairness. In this paper, our main objective is to characterize the risk of fairwashing attacks, in particular by investigating the fidelity-unfairness trade-off. First, we demonstrate through an in-depth empirical study on black-box models trained on several real-world datasets and for several statistical notions of fairness that it is possible to build high-fidelity explanation models with low unfairness. For instance, we find that fairwashed explanation models can exhibit up to 99.20% fidelity to the black-box models they explain while being 50% less unfair. These results suggest that fidelity alone should not be used as a proxy for the quality of black-box explanations. Second, we show that fairwashed explanation models can generalize beyond the suing group (*i.e.*, data points that are being explained), which will only worsen as more stable fairness methods get developed. Finally, we demonstrate that fairwashing attacks can transfer across black-box models, meaning that other black-box models can perform fairwashing without explicitly using their predictions.

1 Introduction

As machine learning models are increasingly integrated into the pipeline of high-stake decision processes, concerns about their transparency are becoming prominent and difficult to ignore for the actors deploying them. As a result, post-hoc explanation techniques have recently gained popularity as they may appear as a potentially viable solution to regain trust in machine learning models' predictions. More precisely, post-hoc explanation techniques refer to methods used to explain how black-box ML models produce their outcomes [26, 9]. Current existing techniques for post-hoc explanations include global and local explanations. In a nutshell, global explanations focus on explaining the whole logic of the black-box model by training a surrogate model that is interpretable by design (*e.g.*, linear models, rule-based models or decision trees) while maximizing its fidelity to the black-box model. In contrast, local explanations aim at explaining a single decision by approximating the black-box in the vicinity of the input point through an interpretable model.

However, several issues need to be addressed when deploying post-hoc explanation techniques. First, a growing body of works has recently shown that post-hoc explanation techniques not only can be misleading [37, 44] but are also vulnerable to adversarial manipulations, wherein an adversary misleads users' trust by devising deceiving explanations. This phenomenon has been demonstrated for a broad range of post-hoc explanation techniques, including global and local explanations [4, 47], example-based explanations [25], explanation map-based explanations [30, 20] and counterfactual explanations [35].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

For instance, in a fairwashing attack [4], the adversary manipulates the explanations provided to under-report the unfairness of the black-box models being explained. This attack can significantly impact individuals who have received a negative outcome following the model's prediction while depriving them of the possibility of contesting it. Second, post-hoc explanations can fool models' designers when used as debugging tools [1, 2]. Finally, they can also be leveraged to perform powerful model extraction attacks [41], which can ultimately lead to the leakage of private information. However, fairwashing in the context of machine learning being a relatively new research topic, a fundamental question is to understand the situations in which fairwashing is possible. In particular, it is not known yet how the success of fairwashing attacks is influenced by the choice of the suing group (*i.e.*, data points being explained) or the particular black-box model considered.

In this work, our aim is to address this gap by characterizing the risk of fairwashing through the lens of post-hoc explanations' manipulability. In particular, to quantify the fidelity-unfairness trade-off, we focus on fairwashing attacks targeting global explanations[1]. Our main contributions can be summarized as follows:

- **Characterization of fairwashing.** We formulate fairwashing as a multi-objective optimization problem, in which an adversary aims to design a high-fidelity interpretable model satisfying a given fairness constraint. Then, we use an ϵ-constraint method to compute the Pareto frontier of the problem and use it to investigate, given a black-box model and a fairness metric, the fidelity-unfairness trade-off that can be achieved through fairwashing. This Pareto frontier can be used as a proxy to characterize the manipulation power of an adversary performing fairwashing on a given black-box model.
- **Generalization of fairwashing beyond the suing group.** Previous works have suggested that fairwashing has to be tailored to the suing group. To investigate this issue, we have performed an in-depth analysis on the generalization of fairwashing to non-members of the suing group for a broad range of fairness metrics. In particular, our results suggest that explanations built to *fairwash* a suing group can generalize to another group that is not explicitly targeted by the fairwashing attack.
- **Transferability of fairwashing.** In addition to the generalization to other groups, our study also demonstrates that fairwashing can transfer across different black-box models. In the spirit of adversarial learning, transferability is defined here as the ability for an explanation model designed to fairwash predictions made by a specific unfair black-box model to successfully fairwash decisions taken by another unfair black-box model without even using its predictions. This property is particularly worrisome as it means that a model producer accused of being unfair could design a more powerful model and use the latter to produce high-fidelity explanations that appear fair (*e.g.*, to clear the doubts about the black-box model's decisions). This finding also suggests that auditing for fairwashing should go beyond the incriminated black-box model and also consider the data distribution and the learning task.

Overview of related work. Table 1 summarizes recent works that have investigated the risk that post-hoc explanation techniques can be manipulated. Most of these attacks are agnostic to the black-box model being explained except for techniques based on explanation maps.

For instance, in the context of example-based explanations' manipulation, Fukuchi et al. [25] have demonstrated the risk of stealthily biased sampling, which occurs when a model producer explains the behaviour of its black-box model by sampling a subset of its training dataset. Slack et al. [47] have shown that variants of LIME [43] and SHAP [38], two popular post-hoc local explanation techniques, can be manipulated to underestimate the unfairness of black-box models. Following the same line of work, Le Merrer and Trédan [36] have demonstrated that a malicious model producer can always craft a fake local explanation to hide the use of discriminatory features.

Another work by Laugel et al. [35] has specifically focused on the use of counterfactual explanations, which is a form of example-based explanation. They have demonstrated that generated instances can be unjustified (*i.e.*, not supported by ground-truth data) for most techniques existing for such type of explanation. Finally, visualization-based explanation techniques can also be manipulated, as shown

[1] For local explanations, the fidelity is not relevant as a local explainer may have a perfect fidelity with respect to the black-box model on the data instance considered.

Table 1: Summary of related work for the manipulation of post-hoc explanations.

Name of attack	Reference	Model agnostic attack	Explanation type	
			Global	Local
Fairwashing	[4]	✓	✓	✓
Stealthily biased sampling	[25]	✓	✓	×
Misleading explanation maps	[30]	×	×	✓
	[20]	×	×	✓
	[6]	×	×	✓
Unjustified counterfactual explanations	[35]	✓	×	✓
Public relations attack	[36]	✓	×	✓
Scaffolding	[47]	✓	×	✓
Misleading black-box explanations	[34]	✓	✓	×

by recent works on saliency map based explanations' manipulation. Heo et al. [30] have shown that these types of explanations are vulnerable to the so-called adversarial model manipulation while Dombrowski et al. [20] have demonstrated their vulnerability to adversarial input manipulation.

The contribution closest to our work is [4], in which the authors first introduced the notion of fairwashing as a rationalization exercise. They have devised `LaundryML`, an algorithm that can systematically rationalize black-box models' decision through global or local explanations. The details of this method, which form the basis of the fairwashing attacks that we have conducted, are given in Section 2.2. A subsequent work [34], similar to [4], has also investigated the possibility that black-box models can be explained with high fidelity by global interpretable models whose features are very different from that of the black-box and look innocuous.

Outline. First, in Section 2, we review the preliminary notions necessary to understand this work and we describe how to explore the fidelity-unfairness trade-offs of a fairwashing attack by using an ϵ-constraint method to solve its underlying multi-objective optimization problem. Afterwards, in Section 3, we present the results obtained from our study for a diverse set of datasets, fairness metrics and black-box models. Section 4 discuss how to quantify the risk of fairwashing. Finally, we discuss the main implications of our findings and conclude in Section 5.

2 Setting and problem formulation

2.1 Notations

Let $X \in \mathcal{X} \subset \mathbb{R}^n$ denote a feature vector, $Y \in \mathcal{Y} = \{0, 1\}$ its associated binary label (for simplicity we assume a binary classification setup without loss of generalization) and $G \in \mathcal{G} = \{0, 1\}$ a feature defining a group membership (*e.g.*, with respect to a sensitive attribute) for every data point sampled from $\mathcal{X}$. In addition, we assume that $b : \mathcal{X} \to \mathcal{Y}$ refers to a black-box classifier of a particular model class $\mathcal{B}$ (*e.g.*, neural network or ensemble model) mapping any input $X \in \mathcal{X}$ to its associated prediction $\hat{Y} \in \mathcal{Y}$. Finally, let $e : \mathcal{X} \to \mathcal{Y}$ be a global explanation model from a particular model class $\mathcal{E}$ (*e.g.*, linear model, rule list or decision tree) designed to explain b.

We train a classifier f (black-box model or explanation model) by minimizing its average loss $L_{\mathcal{D}}(f)$, for a given loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ (*e.g.*, cross entropy), over a dataset of interest $\mathcal{D}$ (*e.g.*, training set or suing group), which means $L_{\mathcal{D}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[l(f(x_i), y_i)]$. In addition, the performance of f is also measured in terms of its unfairness $\texttt{unf}_{\mathcal{D}}(f)$ on $\mathcal{D}$.

In this work, we focus on *statistical notions of fairness* [13, 15, 16, 29], which require a model to exhibit approximate parity according to a statistical measure across the different groups defined by the group membership G. In particular, we consider four different statistical notions of fairness, namely statistical parity [22, 13, 32, 23, 49], predictive equality [15, 16], equal opportunity [29] and equalized odds [15, 33, 29, 48]. The definitions of these fairness metrics are listed in Table 2.

Table 2: Summary of the different statistical notions of fairness considered.

Fairness notion	Definition
Statistical Parity (Δ_{SP})	$\lvert P(\hat{Y}=1\lvert G=0) - P(\hat{Y}=1\lvert G=1)\rvert$
Predictive Equality (Δ_{PE})	$\lvert P(\hat{Y}=1\lvert Y=0,G=0) - P(\hat{Y}=1\lvert Y=0,G=1)\rvert$
Equal Opportunity (Δ_{EOpp})	$\lvert P(\hat{Y}=1\lvert Y=1,G=0) - P(\hat{Y}=1\lvert Y=1,G=1)\rvert$
Equalized Odds (Δ_{EOdds})	$\lvert P(\hat{Y}=1\lvert Y=1,G=0) - P(\hat{Y}=1\lvert Y=1,G=1)\rvert$ and $\lvert P(\hat{Y}=1\lvert Y=0,G=0) - P(\hat{Y}=1\lvert Y=0,G=1)\rvert$

2.2 Problem formulation

Our investigation is motivated by the previous work from Aïvodji et al. [4], who have defined fairwashing in model and outcome explanations as a manipulation exercise in which high-fidelity and fairer explanations can be designed to explain unfair black-box models.

Definition 1 (Global explanation fidelity) Let b be a black-box model, e a global explanation model for b and X a set of data instances. Following the definition in [18], the fidelity of e with respect to b on X is expressed as:

$$\mathsf{fidelity}(e) = \frac{1}{|X|} \sum_{x \in X} \mathbb{I}(e(x) = b(x)).$$

Definition 2 (Global fairwashing attack) Let b be a black-box model and X_{sg} a set of data instances hereafter referred to as suing group. A global fairwashing attack consists in finding an interpretable global model $e = p(b, X_{sg})$ derived from the black-box b and the suing group X_{sg} using some attack process $p(\cdot, \cdot)$, such that e is fairer than b for a given fairness metric.

Aïvodji et al. [4] devised `LaundryML`, an algorithm that can systematically fairwash unfair black-box models' decisions through both global and local explanations. `LaundryML` is a constrained model enumeration technique [28] that searches for explanation models maximizing the fidelity while minimizing the unfairness for a given unfair black-box model.

While Aïvodji et al. [4] focused on the search of high-fidelity explanation models to perform fairwashing attacks, our work goes a step further by determining the fidelity-unfairness trade-offs of those attacks. This allows for a better characterization of the *manipulability* of the explanations. For this purpose, we will compute the set of Pareto optimal explanation models describing all the achievable fidelity-unfairness trade-offs, by solving the following problem:

$$\begin{aligned} \text{minimize} \quad & L_{\mathcal{D}_{sg}}(e) \\ \text{subject to} \quad & \texttt{unf}_{\mathcal{D}_{sg}}(e) \leq \epsilon, \end{aligned} \tag{1}$$

in which e is the explanation model, $\mathcal{D}_{sg} = \{X_{sg}, b(X_{sg})\}$ is formed by the suing group and the prediction of the black-box model b on the suing group, while ϵ is the value of the unfairness constraint.

3 Experimental evaluation

The main objective of this section is to demonstrate empirically that (1) the incurred error in the fidelity imposed by fairwashing can be very small (Section 3.1), (2) fairwashing can generalize beyond suing groups (Section 3.2) and (3) transfer across black-box models (Section 3.3).

Datasets. We have investigated four real-world data commonly used in the fairness literature, namely `Adult Income`, `COMPAS`, `Default Credit` and `Marketing`. Table 3 in the appendix summarizes the main characteristics of these datasets.

- **Adult Income.** The UCI Adult Income [21] dataset contains demographic information about $48,842$ individuals from the 1994 U.S. census. The associated classification task consists in predicting whether a particular individual earns more than $50,000\$$ per year. We used `gender` (`Female`, `Male`) as group membership.

- **COMPAS.** The COMPAS [8] dataset gathers $6,150$ records from criminal offenders in Florida during 2013 and 2014. Here, the classification task consists in inferring who will re-offend within two years. We used `race` (`African-American`, `Caucasian`) as group membership.
- **Default Credit.** The Default Credit [21] dataset is composed of information of $29,986$ Taiwanese credit card users. The classification task is to predict whether a user will default in its payments. We used `sex` (`Female`, `Male`) as group membership.
- **Bank Marketing.** The UCI Bank Marketing [42] dataset contains information about $41,175$ customers of a Portuguese banking institution contacted as part of a marketing campaign (*i.e.*, phone calls), whose goal was to convince them to subscribe to a term deposit. The classification task consists of predicting who will subscribe to a term deposit. We used `age` (`30-60`, `not30-60`) as group membership.

Preprocessing. Before running the experiments, each dataset is split into three subsets, namely the *training set* (67%), the *suing group* (16.5%) and the *test set* (16.5%). Overall, we created 10 different samplings of the three subsets using different random seeds, the results of all the experiments being averaged over these 10 samples. The training set is directly used to train the black-box models, while the suing group dataset is used to prepare the explanation models as well to evaluate their fidelity-unfairness trade-offs. Finally, the test set is used to assess the accuracy of the black-box models as well as the generalization of the explanation models beyond their suing groups. For all models (*i.e.,* black-boxes and explanation models), we used a one-hot encoding of the features of the datasets.

Black-box models. We have trained four different types of black-box models on each dataset, namely a Deep Neural Network (DNN), a Random Forest (RF) [12], a AdaBoost classifier [24] and a XgBoost classifier [14]. To tune the hyperparameters of these models, during their training, we performed a hyperparameter search with 25 iterations using `HyperOpt` [10]. The performances (*i.e.*, accuracy and unfairness) of the four black-box models on the training set, the suing group set as well as the test set of the four datasets are provided in Table 4 in the appendix.

Explanation models. We solved the optimisation problem defined in Equation 1 for three model classes, namely logistic regression, rule lists and decision trees. For rule lists, we used Fair-CORELS [5], a modified version of `CORELS` [7], which trains rule lists under fairness constraints. For both logistic regression and decision trees, we used the exponentiated gradient technique [3], which is a model agnostic technique to train any classifier under fairness constraints (we use its implementation in the `Fairlearn` library [11]).

3.1 Experiment 1: fidelity-unfairness trade-offs in fairwashing

The main objective of this experiment is to characterize the fidelity-unfairness trade-offs incurred by fairwashing when using different datasets, fairness metrics, black-box models and explanation models.

Setup. Given a suing group X_{sg}, for each black-box model b and each fairness metric m, the Pareto fronts are obtained by first sweeping over 300 values of fairness coefficients $\epsilon_m \in [0, 1]$. Afterwards, for each value of ϵ_m, an explanation model e_{ϵ_m} is trained to satisfy the unfairness constraint ϵ_m on X_{sg}, by solving the problem in Equation 1 for logistic regression, rule list and decision tree. Then, its effective unfairness and fidelity (with respect to b) on X_{sg} are returned. Finally, the set of non-dominated points is computed.

Results. Top rows in Figure 1 show the fidelity-unfairness trade-offs of fairwashed decision trees found for the four black-box models, respectively on Adult Income, for members of the suing group, using four different fairness metrics: equalized odds, equal opportunity, predictive equality and statistical parity. We present more detailed results on the fidelity-unfairness trade-offs of fairwashed decision tree, logistic regression, and rule list explainers, for Adult Income (Figure 5), COMPAS (Figure 6), Default Credit (Figure 7), and Marketing (Figure 8) in the appendix.

Finally, Figure 2 displays the fidelity of the fairwashed decision trees when they are designed to be at least 50% less unfair than the black-box models they are explaining. Results are shown for all

datasets, fairness metrics and three family of black-box models (AdaBoost, DNN and RF). A more complete diagram with results for the XgBoost black-box model is shown in the appendix (Figure 9), as well as results for fairwashed logistic regression models (Figure 10) and rule lists (Figure 11).

Consistently over all these results, we observe that the fairwashed explanation models found for the suing groups were significantly less unfair than the black-box models while maintaining high fidelity. More precisely, for any combination of fairness metric m and black-box model b, a fairwashed decision tree displays an unfairness less than 50% of the unfairness of b while maintaining a fidelity greater than 87%, 90%, 95% and 96% respectively for COMPAS, Adult Income, Default Credit and Marketing. This fidelity can go up to 99% for some particular combination of black-box model and fairness metrics (*e.g.,* AdaBoost's fairwashed explanation models for Marketing dataset). Furthermore, the small percentage change of the fidelity of the fairwashed explanation models indicate that fairwashing does not introduce an important loss in fidelity when compared to non-fairwashed explanations models.

3.2 Experiment 2: generalization of fairwashing beyond suing groups

The main objective of this experiment is to assess the generalization of a fairwashed explanation models beyond the suing group. This generalization indicates whether or not fairwashing is an attack that has to be tailored for a particular subset of data instances or whether it is more generic in its scope and thus also more problematic.

Setup. We used the same experimental setup as in `Experiment 1`. However, the unfairness and fidelity of the explanation model are computed on a test set X_{test} such that $X_{sg} \cap X_{test} = \emptyset$. The role of X_{test} is to mimic non-members of the suing group X_{sg}, which is targeted by the fairwashed explanation model.

Results. Bottom rows in Figure 1 show the fidelity-unfairness trade-offs of fairwashed decision trees found for the four black-box models on Adult Income, for non-members of the suing group, using four different fairness metrics: equalized odds, equal opportunity, predictive equality and statistical parity. As mentioned in Section 3.1, complete results for other explanation model classes and all datasets are presented in the appendix.

Overall, the results show that the explanation models designed for a particular suing group generalize well also to non-members of that suing group by achieving similar fidelity-unfairness trade-offs. The small gap between results on members of the suing group X_{sg} and those on non-members X_{test} can be explained by the fact that methods for learning fair classifiers are usually not robust to perturbations in the training distribution [31], resulting in fairness violations when evaluated on test sets. In fact, the fairwashing attack defined in Equation 1 is equivalent to a problem of training an interpretable model under a fairness constraint, in which the training pair (X, Y) is formed by the suing group X_{sg} and the predictions $b(X_{sg})$ of the black-box b. As a result, the issue of generalizing beyond the suing group can be reduced to the problem of generalizing fairness beyond the training set. In our scenario, the fairness violations are small enough to favor high-fidelity explanation models on the test set. Nonetheless, with the development of robust fairness-enhancing techniques (*e.g.*, [39]), it is reasonable to expect that this small gap in fidelity could be narrowed further in the future.

3.3 Experiment 3: transferability of fairwashing

The objective of this experiment is to verify whether an explanation model specifically designed for a particular black-box model (called the teacher model) can be used to fairwash decisions of other black-box models (here the student models). The motivation for this experiment is similar to the study of transferability in the context of adversarial learning, which shows that it is not so much the characteristics of the black-box model rather than that of the dataset and the classification task that makes the attack possible.

Setup. Given a suing group X_{sg}, a teacher black-box model $b_{teacher}$, a fairness metric m, its associated fairness constraint ϵ_m and a set of student black-box models $b^i_{student}$, with $i = 1, \ldots, n$, an explanation model e_{ϵ_m} is trained to satisfy the unfairness constraint ϵ_m on X_{sg}, by solving the problem in Eq 1 for logistic regression, rule lists and decision trees. Afterwards, the unfairness and fidelity of e_{ϵ_m} are evaluated with respect to each of the student black-box models $b^i_{student}$

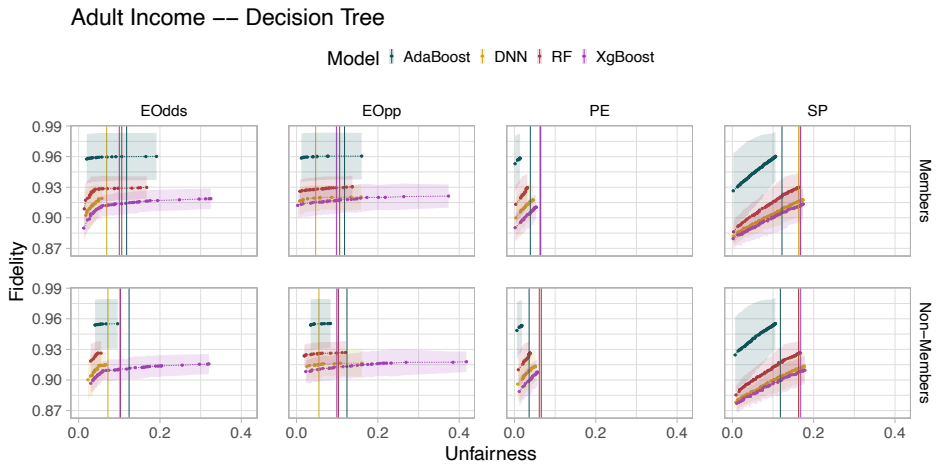

Figure 1: Fidelity-unfairness trade-off of fairwashing attacks for equalized odds, equal opportunity, predictive equality and statistical parity metrics on Adult Income, using decision trees as explanation models. Vertical lines denote the unfairness of the black-box models. Results are averaged over 10 fairwashing attacks.

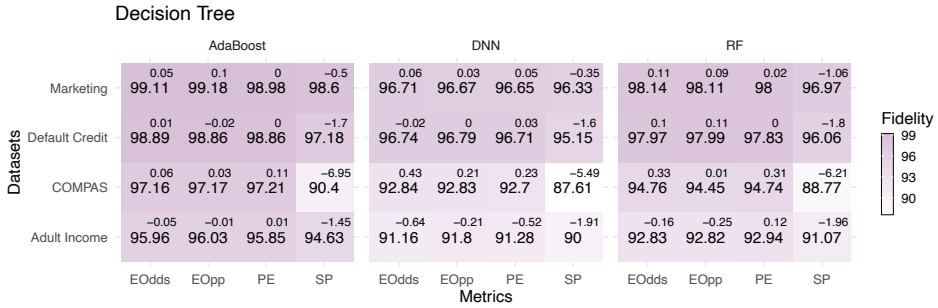

Figure 2: Fidelity of the fairwashed decision trees that are 50% less unfair than the black-box models they are explaining. Results (averaged over 10 fairwashing attacks) are shown for AdaBoost, DNN and RF black-box models, for all datasets and fairness metrics. The content of each cell is in the form of x^y, in which x represents the fidelity of the fairwashed explanation model, and y its percentage change with respect to the fidelity of the unconstrained explainer, used here as a baseline.

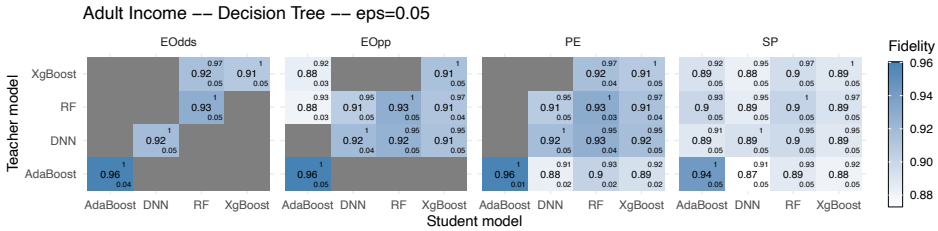

Figure 3: Analysis of the transferability of fairwashing attacks for equalized odds, equal opportunity, predictive parity and statistical parity on Adult Income, for $\epsilon = 0.05$, and for decision tree explanation models. The result in each cell is in the form of x^y_z, in which y denotes the label agreement between the teacher black-box model and the student black-box model, x is the fidelity of the fairwashed explanation model and z is its unfairness. Blank cells denotes the absence of transferability for the unfairness constraint imposed. Results are averaged over 10 fairwashing attacks.

on X_{sg}. For this experiment, we considered four black-box models (AdaBoost, DNN, RF and XgBoost). First, we fixed one model as the teacher black-box model and used the remaining ones as the student black-box models. We conducted the experiments for all four possible combinations of the teacher-student models. We evaluated the transferability for two value of the unfairness constraint ($\epsilon \in \{0.01, 0.05, 0.1\}$) to simulate both strong and loose fairness constraints.

Results.

Figure 3 displays the fidelity and unfairness of fairwashed decision trees with respect to both the teacher and the set of student black-box models on the suing group X_{sg}. Results are shown for Adult Income, for all fairness metrics and different values of the unfairness constraint ($\epsilon = 0.05$)[2]. Additional results for Adult Income using linear regression and rule list explainers are provided in the appendix (Figures 13 and 14), as well as results obtained for COMPAS (Figures 15, 16, and 17), Default Credit (Figures 18, 19 and 20) and Marketing (Figures 21, 22 and 23).

Overall, our results demonstrate that fairwashing can transfer with high fidelity for most models. For instance, on Adult Income with an equal opportunity constraint set to 0.05, a fairwashed decision tree that had a fidelity of 93% for a RF teacher model successfully transferred to AdaBoost, DNN and XgBoost student models with a fidelity of 88%, 91% and 91% respectively (*c.f.*, Figure 3). For other datasets, an even higher fidelity can be achieved. In particular, on Default Credit with an equal opportunity constraint set to 0.05, a fairwashed decision tree that had a fidelity of 98% for a Random Forest teacher model successfully transferred to AdaBoost, DNN and XgBoost student models with a fidelity of respectively 95%, 96% and 97% (*c.f.*, Figure 18).

4 Quantifying the risk of fairwashing

At a high level, one can say that fairwashing is possible because the set of all the explainers that can explain a particular black-box model (at a particular level of fidelity) is diverse in terms of unfairness (*i.e.,* contains both fair and unfair explainers). This phenomenon has been observed for a broad range of models and domains, and received different names including predictive multiplicity [40], underspecification [19], and multiplicity in the set of almost-equally-accurate models (also referred to as *Rashomon set*) [46, 27, 45].

Given a model class $\mathcal{F}$, a loss function $L_{\mathcal{D}}(\cdot)$ over a dataset $\mathcal{D}$ of interest, a reference model f^* (*e.g.,* optimal model, explainer with at least 95% fidelity), and a performance threshold $\tau \in [0, 1]$, the Rashomon set $R_s(\mathcal{F}, f^*, \tau)$ can be defined as :

$$R_s(\mathcal{F}, f^*, \tau) = \{f \in \mathcal{F} \mid L_{\mathcal{D}}(f) \leq (1 + \tau) L_{\mathcal{D}}(f^*)\} \tag{2}$$

Quantifying the unfairness of the explainers that are in the Rashomon set can help in evaluating the risk of fairwashing. More precisely, if the range of the unfairness of explainers that lie within the Rashomon set of a black-box model's explainers covers both high and low regions of unfairness, then the risk of fairwashing is high: one can find high-fidelity explainers with low unfairness. Conversely, suppose the range is centred around the black-box model's unfairness. In that case, it will be difficult to fairwash: one can not find high-fidelity explainers with an unfairness very different than that of the black-box model. To explore the effectiveness of this technique, we use the recent approach proposed by Coston et al. [17] to compute the range of the unfairness of high-fidelity explainers by solving the following problem:

$$\begin{aligned} \text{minimize} \quad & \texttt{unf}_{\mathcal{D}_{sg}}(e) \\ \text{subject to} \quad & L_{\mathcal{D}_{sg}}(e) \leq v, \end{aligned} \tag{3}$$

in which e is the explanation model, $\mathcal{D}_{sg} = \{X_{sg}, b(X_{sg})\}$ is formed by the suing group and the prediction of the black-box model b on the suing group, while v is the value of the constraint on the loss to explore different levels of fidelity.

In this experiment, we explore different values for the fidelity ranged from 90% to 99%. In particular, we solved the problem of minimizing the disparity (and not the absolute disparity) under constraints

[2] Results for $\epsilon \in \{0.03, 0.1\}$ are displayed in the appendix (Figure 12).

of the loss as described in [17, Problem 2]. The authors proposed a reduction approach inspired by the work of Agarwal et al. [3] to explore the range of unfairness for a fixed value of the loss. We used this approach and swept over different values of the loss function to compute the range of unfairness of high-fidelity explainers.[3]

The results obtained for Adult Income and COMPAS (*c.f.*, Figure 4) effectively corroborate with the performances we observed for both datasets for fairwashed logistic regression explainers on statistical parity (*c.f.*, Figures 5, 6, and 10): fairwashing is possible for both datasets, and it is easier (*i.e.,* possibility to have high-fidelity with low unfairness) to do it on Adult Income than COMPAS.

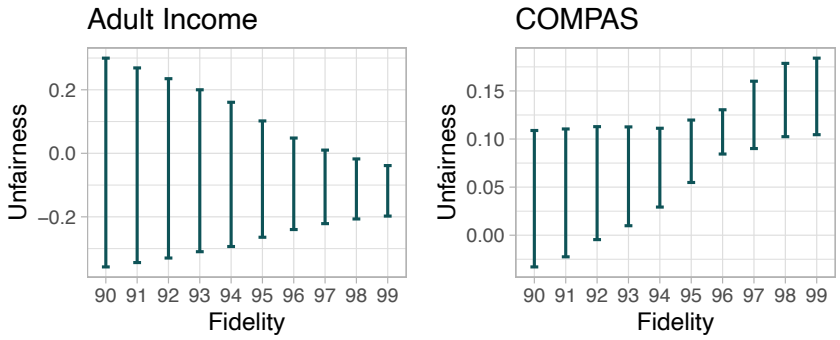

Figure 4: Range of the statistical parity of logistic regression explainers for different values of the fidelity for a DNN black-box trained on Adult Income and COMPAS.

5 Conclusion

In this paper, we have characterized the manipulability power of fairwashing attacks by analyzing their fidelity-unfairness trade-offs in diverse situations. In particular, we have demonstrated for different fairness metrics and black-box models that (1) fairwashed explanation models can exhibit significantly low unfairness while having a high fidelity to the black-box, (2) fairwashed explanation models can generalize beyond the suing group and (3) fairwashing attacks can transfer across black-box models.

The first lesson to draw from our investigation is that *relying on the fidelity as a proxy for the quality of a post-hoc explanation can be misleading* as a fairwashed explanation model can exhibit high fidelity while being significantly less unfair than the black-box model being explained. In addition, the results obtained for the generalization of fairwashed models beyond suing groups demonstrate that a *fairwashed explanation model can also rationalize subsequent unfair decisions made by the original black-box model for free*. This fact preclude the possibility of designing fairwashing detection techniques that leverage on the instability of the unfairness with respect to variations in the suing group. Indeed, such technique will most likely fail against fairwashed explanation models designed using stable fair classification algorithms [31, 39].

Furthermore, while generalizing beyond the suing group enables a dishonest black-box model producer to reuse its fairwashed explanation models for subsequent unfair decisions, the transferability property shown in the experiments could help the latter *to use a fairwashed explanation models to rationalize unfair decisions for other black-box models than the one it was designed for*. As a direct consequence, a model producer accused of deploying a black-box model providing unfair decisions could develop after *ex post facto* another black-box model to rationalize the unfair decisions of the first one. These results also suggest that auditing a black-box model for fairwashing should go beyond the model itself and consider the data distribution as well as the learning task, and consider the Rashomon set of the explainer. Our investigation of the range of the unfairness of high-fidelity explainers effectively show that such an approach can help in quantifying the risk of fairwashing.

[3] We used the implementation provided by the authors in the public Github repository available at `https://github.com/asheshrambachan/Fairness_In_The_Rashomon_Set`

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[2] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems*, 2020.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69, 2018.

[4] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 161–170, 2019.

[5] U. Aïvodji, J. Ferry, S. Gambs, M.-J. Huguet, and M. Siala. Learning fair rule lists. *arXiv preprint arXiv:1909.03977*, 2019.

[6] C. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning*, pages 314–323, 2020.

[7] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 35–44, 2017.

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[9] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 50:82–115, 2020.

[10] J. Bergstra, D. Yamins, and D. Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123, 2013.

[11] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft Technical Report MSR-TR-2020-32*, 2020.

[12] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[14] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[15] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[16] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[17] A. Coston, A. Rambachan, and A. Chouldechova. Characterizing fairness over the set of good models under selective labels, 2021.

[18] M. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, pages 24–30, 1996.

[19] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[20] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13589–13600, 2019.

[21] D. Dua and C. Graff. UCI machine learning repository. 2017. URL http://archive.ics.uci.edu/ml.

[22] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

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

[24] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[25] K. Fukuchi, S. Hara, and T. Maehara. Faking fairness via stealthily biased sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 412–419, 2020.

[26] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93, 2019.

[27] L. Hancox-Li. Robustness in machine learning explanations: does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 640–647, 2020.

[28] S. Hara and T. Maehara. Enumerate Lasso solutions for feature selection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1985–1991, 2017.

[29] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[30] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pages 2925–2936, 2019.

[31] L. Huang and N. Vishnoi. Stable and fair classification. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2879–2890, 2019.

[32] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[33] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.

[34] H. Lakkaraju and O. Bastani. "How do I fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

[35] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2801–2807, 2019.

[36] E. Le Merrer and G. Trédan. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2(9):529–539, 2020.

[37] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[38] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[39] D. Mandal, S. Deng, S. Jana, J. Wing, and D. J. Hsu. Ensuring fairness beyond the training data. In *Advances in Neural Information Processing Systems*, 2020.

[40] C. Marx, F. Calmon, and B. Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6765–6774, 2020.

[41] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[42] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[43] M. T. Ribeiro, S. Singh, and C. 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*, pages 1135–1144, 2016.

[44] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[45] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[46] L. Semenova, C. Rudin, and R. Parr. A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*, 2019.

[47] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. pages 180–186, 2020.

[48] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.

[49] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3.1 for results related to the characterization of the fidelity-unfairness trade-offs in fairwashing, Section 3.2 for results related to the generalization of fairwashing beyond suing groups, and Section 3.3 for results related to the transferability of fairwashing across black-box models.
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] All the code to reproduce all the experimental results is included in the supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code of our experiments.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]