The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations

APARNA BALAGOPALAN, Massachusetts Institute of Technology, USA
HAORAN ZHANG, Massachusetts Institute of Technology, USA
KIMIA HAMIDIEH, University of Toronto, Vector Institute, Canada
THOMAS HARTVIGSEN, Massachusetts Institute of Technology, USA
FRANK RUDZICZ, University of Toronto, Vector Institute, Unity Health Toronto, Canada
MARZYEH GHASSEMI, Massachusetts Institute of Technology, Vector Institute, USA

Machine learning models in safety-critical settings like healthcare are often "blackboxes": they contain a large number of parameters which are not transparent to users. Post-hoc explainability methods where a simple, human-interpretable model imitates the behavior of these blackbox models are often proposed to help users trust model predictions. In this work, we audit the quality of such explanations for different protected subgroups using real data from four settings in finance, healthcare, college admissions, and the US justice system. Across two different blackbox model architectures and four popular explainability methods, we find that the approximation quality of explanation models, also known as the *fidelity*, differs significantly between subgroups. We also demonstrate that pairing explainability methods with recent advances in robust machine learning can improve explanation fairness in some settings. However, we highlight the importance of communicating details of non-zero fidelity gaps to users, since a single solution might not exist across all settings. Finally, we discuss the implications of unfair explanation models as a challenging and understudied problem facing the machine learning community.

CCS Concepts: • **machine learning** → **explanations**; **fairness**.

Additional Key Words and Phrases: explainability, machine learning, fairness

1 INTRODUCTION

Machine learning (ML) models are increasingly used in safety-critical settings like healthcare [23, 107], college admissions [53], and law [6]. Several studies have shown that human decisions can become more accurate when assisted by such ML models [16, 106, 111]. However, many ML models are "blackboxes"—they might have too many parameters or be proprietary—and cannot explain their predictions in ways humans understand [99]. In such scenarios, users may struggle to understand a model's outputs enough to trust and use its predictions [33, 35, 64].

Post-hoc explainability methods have recently begun helping users better understand why blackbox models make certain predictions [61, 93, 94]. A popular post-hoc approach is to train simple, human-interpretable models to *imitate* a blackbox model's behaviour [92] by maximizing the congruity between simple approximations and blackbox model predictions. Such approximation quality is known as *fidelity* [30]. Then, the simpler model can be used either as a new stand-alone model or to explain one prediction at a time [92, 105]. By highlighting important inputs, these explainability methods provide a path towards helping users trust machine learning models in high-impact settings [93].

However, it remains unknown when and if these models approximate behavior *fairly*. If fidelity differs between different pre-defined groups (e.g., between demographics) in a dataset, explainability methods may perpetuate machine

Authors' addresses: Aparna Balagopalan, Massachusetts Institute of Technology, USA, aparnab@mit.edu; Haoran Zhang, Massachusetts Institute of Technology, USA, haoranz@mit.edu; Kimia Hamidieh, University of Toronto, and Vector Institute, Canada, kimia@cs.toronto.edu; Thomas Hartvigsen, Massachusetts Institute of Technology, USA, tomh@mit.edu; Frank Rudzicz, University of Toronto, and Vector Institute, and Unity Health Toronto, Canada, frank@cs.toronto.edu; Marzyeh Ghassemi, Massachusetts Institute of Technology, and Vector Institute, USA, mghassem@mit.edu.

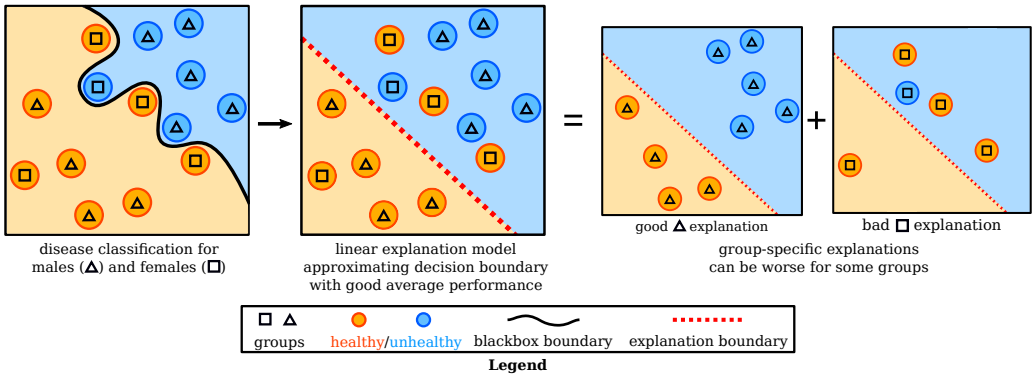


Fig. 1. An example of an unfair global explanation model. Orange and Blue circles indicate predicted classes, *healthy* or *unhealthy*, respectively. □ and △ denote group membership. The red dashed line is a linear explainability model fit to approximate the black blackbox decision boundary. The two figures on the right show that the linear approximate is worse for the □ group.

bias by encouraging users to trust model predictions for some people but not others. In this work, we study *to what degree do gaps in fidelity exist between subgroups?*

To answer this question, we measure group-wise fidelity for different post-hoc explanation methods on real tabular datasets that include group membership. Intuitively, an explanation model is *fair* if it has equally high fidelity for all protected groups. This definition is similar to common group fairness definitions which seek to eliminate gaps in predictive performance across groups [46, 83, 119]. We introduce two definitions of *fidelity gaps*, or disparities in fidelity across different subgroups. Using these measures, we benchmark two popular families of post-hoc explanation models: local methods, which imitate the boundary of a blackbox around one instance [70, 92, 94], and global methods, which imitate the blackbox across all instances [61, 98]. We also motivate fidelity gap measurements by showing mathematically that measuring fidelity gaps across subgroups directly connects with prior work on fairness preservation for explainability [31]. With a comprehensive audit of explanation fairness, we find that significant fidelity gaps exist between subgroups.

A popular way to train fairer models is through robust optimization [65, 89, 100]. To see how robust training impacts large fidelity gaps, we also study a simple technique for retraining explainability methods to improve their fairness. We also study potential causes for these fidelity gaps, and highlight mechanisms by which group information can indirectly be used in post-hoc explanations as an important contributing factor. Lastly, we assess the impact of the observed fidelity gaps on real-world decision-making accuracy with a carefully designed simulation study. The major findings of our evaluation are as follows:

Explanation fidelity varies significantly between subgroups: We find that fidelity gaps grow up to 7% between subgroups in our experiments using four popular datasets. In comparison to average fidelity across all data points, the fidelity of explanations for disadvantaged groups is often significantly lower (up to 21%). These findings indicate that judging the quality of explanations by their *average* fidelity alone—a common approach—overestimates explanation quality for some subgroups, potentially leading to worse downstream decision making. This effect is illustrated in Fig. 1.

Balanced and robust training can reduce but not eliminate fidelity gaps: We use robust training by adaptively reweighting or balancing groups in training data while training explanation models. This turns out to be a promising direction: fidelity gaps improve across subgroups, though this depends on both the dataset and exact method utilized.

Fidelity gaps have an impact on decision-making in the real-world: Using a simulation study, we find that larger fidelity gaps may lead to disparities in decision making accuracy for different subgroups. This implies that ignoring fidelity gaps between subgroups can have detrimental effects to members of protected groups.

Finally, we categorize and discuss promising directions for evaluating and improving post-hoc explainability methods. In summary, our work is a step towards training fair and reliable explanation models.

2 RELATED WORK

2.1 Explainable Machine Learning

While ML models achieve outstanding performance, users often find them too complex to trust in practice. To make such *blackbox* models more useful, users require that they be understandable, often due to laws [13] or preference [12, 49, 95]. To fill this gap, recent approaches "explain" a blackbox model's behavior after it is trained [18, 36]. These *post-hoc* explainability methods are now used in safety-critical applications like healthcare [3] and finance [19].

Several explainability methods are increasingly-popular because they make no assumptions about a blackbox model's architecture [93], also known as model-agnostic. In contrast, some methods are designed exclusively for deep learning, requiring their internal structure and gradients [8, 55, 62, 103, 103]. We consider model-agnostic methods, which can be used for a wider family of blackbox models, including deep learning.

Model-agnostic explainability methods are primarily either *local* or *global* [38]. *Local* methods justify one model prediction at a time, typically by approximating the decision boundary around one data point [14, 70, 84, 91, 92, 94]. Then, the weights learned by the local models are used to rationalize the blackbox model's prediction. Some of the best-known local methods are LIME [92], which learns a sparse linear classifier on a dataset of perturbed samples, and SHAP [70], which uses feature-wise Shapley values [97]. *Global* methods, on the other hand, train interpretable surrogate models of the blackbox model's behavior on an entire dataset, which is then used in lieu of the blackbox. These methods primarily use tree-based models [69], rule lists [61, 88], sparse linear models [108, 120], and generalized additive models [68] as surrogates.

2.1.1 Explainable ML in Safety-critical Settings. The need for explainability in safety-critical applications is a nebulous and contested topic for several reasons:

Interpretability vs Explainability. Some prior works advocate for interpretability over explainability [41, 76, 99]. An explanation model without perfect fidelity is by definition incorrect for some data points [99]. Our work extends this point; these errors can occur for some groups more than others. Since explanations influence trust [16], it is important to conduct user studies on the impacts on aversion algorithmic advice [67] and over-reliance on algorithmic advice [34].

Anchoring Effects. Explanations can fool people into trusting incorrect models [9, 85]. For example, Poursabzi-Sangdeh et al. [85] find that when people are shown explanations from a bad model, they become more likely to trust the model, even when it is clearly wrong. In cases like this, people use the explanations while judging the quality of the blackbox models, even though the explanations themselves can be misleading [10].

Mismatched end-user and model-designer goals. Many explainability methods aim to assist model debugging, while non-engineer users only choose when to accept a blackbox's decisions [57, 86, 93]. This mismatch can have downfalls. For instance, Buçinca et al. [17] find that the explanations people find most useful are also the ones they trust *incorrectly*. Resolving this mismatch requires goal-aware explainability methods along with education to ensure end-users are properly trained in using these methods.

2.1.2 Desiderata for Post-hoc Explanations. Most post-hoc explainability methods have three goals:

Reliability. Explanations must be accurate for the right reasons. People often trust explained models [85, 96], so ensuring that explanations are faithful to the original model and not simply easy-to-rationalize is essential [41].

Robustness. Explanation models should not overfit to spurious patterns in the data [42, 59] and must be robust in the presence of small distribution shifts at test time [59].

Simplicity. Models should be sparse, and leave little room for effects for human cognitive biases such as the anchoring effect [85]. Ideal explanations will highlight only the key information needed to understand a model's behavior, encouraging users to engage with explanations in predictable ways [17]. However, there is often a trade-off between an explanation's faithfulness and its simplicity [60]. Recent work on cognitive forcing—where users explicitly interact and understand explanations—appears to be a promising direction to address this trade-off [17].

Along with other recent efforts [10, 42, 59, 85], we promote a fourth goal: **Fairness**. Explanation quality should not depend on group membership. We find that this requirement is not yet satisfied by popular explainability methods.

2.2 Algorithmic Fairness

Formalizing fairness is a flourishing research area [11, 22, 26, 27, 46, 66, 74, 116, 117]. Recent works define fairness at either the *individual-* or *group*-level. Individual fairness requires similar predictions for similar individuals; group fairness requires similar predictions for different groups (sex or race, for example). We consider group-level fairness for binary classification, which we quantify using demographic parity (DP) [46, 83], a standard group-fairness metric. We describe this metric probabilistically, allowing calculation of gaps across groups: $\text{DP} = E[\hat{Y}|A = a] - E[\hat{Y}|A = b] \quad \forall a, b \in A$, where $\hat{Y}$ is a predictor and its DP is measured with respect to attribute A.

There are three main strategies for encouraging group fairness [20]: pre-processing data to find less-biased representations [80]; enforcing fairness while training a model, typically through regularization [71, 118]; and altering a model's predictions to satisfy fairness constraints after it is trained [2, 24, 46, 83]. In this paper, we utilize the inprocessing method proposed by Zhang et al. [118] for training fair blackbox models. Further, recent work has demonstrated that group-robust training can increase fairness by improving the worst-group accuracy [101].

2.3 Bias in Model Compression and Risks of Fairwashing

Several recent works study the effects of model and data compression on fairness [50, 102]. For example, Samadi et al. [102] observe that reconstruction error associated with data dimensionality reduction via principal component analysis is higher for some populations. Hooker et al. [50] show that average accuracy after ML model compression hides disproportionately high errors on a small subset of examples. In a similar vein, we study post-hoc explanation models, which are often compressed blackbox models, and assess how they transmit bias. Another related topic is "fairwashing": the act of overlooking a model's unfair behavior by rationalizing its predictions via explanations [4]. Our paper instead considers fairness in how well explanation models imitate blackbox models (rather than the ground-truth), regardless of blackbox model fairness.

3 MEASURING THE FAIRNESS OF EXPLANATIONS

Here, we introduce metrics for measuring *fairness of explanation models* or fidelity gaps across subgroups.

3.1 Notation

Consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ that contains n training data points. $\boldsymbol{x}_i \in \mathbb{R}^d$ is the d-dimensional feature vector of the i-th data point in $\mathcal{D}$ and $y_i \in \{0, 1\}$ is its associated binary label. We assume binary classification for simplicity. Let $g \in \{1, \ldots, G\}$ be a variable defining group membership with respect to the protected attribute for every data point for a total of G subgroups. In all cases, g serves as auxiliary information and is not used during any model training, unless specified otherwise. A blackbox classifier $B : \mathbb{R}^d \to \{0, 1\}$ predicts one binary label per input x. Given classifier B, we wish to explain its prediction given some query point $\boldsymbol{x}^*$. To achieve this, an explanation model E is chosen from a set of interpretable models (*e.g.*, linear models or decision trees). Then, E is trained to *imitate* B either locally (for the feature space near $\boldsymbol{x}^*$) or globally (for all data points in $\mathcal{D}$).

3.2 Fidelity of Explanations

Given a blackbox B and explainability model E, we seek to describe how well E approximates B's behavior. Fidelity, as detailed in Definition 3.1 below, is a powerful measure for this approximation error [5, 60, 61], though it disregards group information.

Definition 3.1 (Explanation Fidelity [30]). Given blackbox model B and explanation model E, the *explanation fidelity* on data points $(x_i, y_i)_{i=1}^N$ is $\frac{1}{N}\sum_{i=1}^N L(B(x_i), E(x_i))$, where L is a performance metric.

For L, we use accuracy, AUROC[1], and mean error, denoted as $Fidelity^{Acc}$, $Fidelity^{AUROC}$, $Fidelity^{Err}$, respectively.

In the following sections, we build up to a definition of explanation fidelity that considers group information. First, we motivate the need for a metric that measures fidelity across groups (Section 3.3), then define two new notions for measuring the fairness of explanations (Section 3.4).

3.3 Fidelity Gaps are Critical to Fairness Preservation

Dai et al. [31] recently introduced *fairness preservation* in surrogate explanation models. Fairness is *preserved* when the fairness properties of the blackbox model and explanation model are identical. For example, consider Figure 1. A linear explanation E is a high-fidelity approximation of the blackbox B's decisions for one group (△), but not the other (□). Here, B seems unfair in predicting the "unhealthy" class for the two groups. Meanwhile, E appears fairer. In this example, B's degree of (un)fairness—the demographic parity gap—is not *preserved* by the explanation model. For demographic parity, fairness preservation in explanations implies that B and E should have similar DP Gaps (Section 2.2).

To reliably judge a blackbox's fairness using only its post-hoc explanations, preserving fairness is essential. If fairness is preserved, then when an explanation seems unfair, we can be confident that the blackbox model is likely similarly unfair as well. Next, we prove that fairness preservation is directly linked to fidelity gaps across subgroups.

3.3.1 *Fidelity Gaps are related to Fairness Preservation.*

Theorem 3.2. *Let E be a post-hoc explanation model trained to imitate predictions of blackbox model B, and mean residual error for a set of N data points $x \in \mathbb{R}^d$ is $\frac{1}{N}\sum_{x\in X}(E(x) - B(x))$. Then, the difference between the Demographic Parity Gaps of E and B, both with respect to binary valued-protected attribute g, is equal to the difference in mean residual error of data points with $g = 1$ and $g = 0$.*

[1] AUROC cannot be written directly as a sum but we slightly abuse notation for readability.

The full proof of this theorem is in Appendix A.1; the key idea is to expand $E(x_i) = B(x_i) + \epsilon_i$ where ϵ_i is a residual for each point x_i. This is valid when E is trained to imitate B with high fidelity (*e.g.*, minimizing mean squared error or cross-entropy loss). We also empirically validate this theorem on explanation models in Appendix A.2.

From Theorem 3.2, a *sufficient* condition for DP, as computed over instances x_i and their local linear classifiers $E_i(x_i)$ or a global model E (where $E_i = E \forall i$) is ensuring that the mean residual errors for each group is comparable. This is the same as low fidelity gaps across subgroups where L is the mean error. Note that this does not correspond to mean absolute difference between predictions of E and B, but instead their mean difference. Theorems of similar form could be derived for other group fairness definitions, but the ϵ values and data points considered would depend on the ground truth as well (e.g., for equal-opportunity, the ϵ difference would only contain terms for data points with positive-class ground truth). With this motivation in mind, we next introduce two new metrics that measure fidelity gaps across subgroups.

3.4 Measuring the Fairness of Explainability Methods

Building on the definition of average fidelity across groups (Defn. 3.1), we introduce two new measurements for the fairness of explanation models by evaluating their fidelity gaps between subgroups. The first metric (Definition 3.3) addresses the question: by what degree would relying on the average fidelity alone be detrimental to subgroups of data? The second metric estimates the mean difference in fidelity of explanations between subgroups of data (Definition 3.4).

Inspired by past work [30, 65], the *maximum fidelity gap from average* (Definition 3.3) computes the different between the overall, average fidelity and the worst-case subgroup fidelity. This way, we quantify the maximum degree to which an explanation model's fidelity is lower for disadvantaged groups compared to the average across all subgroups.

Definition 3.3 (Maximum Fidelity Gap from Average: Δ_L). Let the maximum fidelity gap from average be

$$\Delta_L = \max_j \left[\frac{1}{N} \sum_{i=1}^{N} L(B(x_i), E(x_i)) - \frac{1}{N_j} \sum_{i:g_i^j=1} L(B(x_i), E(x_i)) \right],$$

where $g_i^j = 1$ denotes that point x_i belongs to the jth subgroup defined by a specific protected attribute g (*e.g.* data points from females), and N_j is the number of data points with $g^j = 1$.

Next, the mean fidelity gap amongst subgroups (Definition 3.4) computes how much an explanation model's fidelity differs over subgroups. Here, we only consider groups defined by the same sensitive attribute (*e.g.* g^k is male, g^j is female).

Definition 3.4 (Mean Fidelity Gap Amongst Subgroups: Δ_L^{group}). Let the mean fidelity gap amongst subgroups be

$$\Delta_L^{group} = \frac{2}{G(G-1)} \sum_{k=1}^{G} \sum_{j=k+1}^{G} \left| \frac{1}{N_k} \sum_{i:g_i^k=1} L(B(x_i), E(x_i)) - \frac{1}{N_j} \sum_{i:g_i^j=1} L(B(x_i), E(x_i)) \right|,$$

where g_j denotes the j^{th} subgroup defined by a specific sensitive attribute (*e.g.* datapoints from females), and N_j is the number of datapoints in g_j.

Similar to average fidelity, we choose L to be Accuracy, AUROC, and Mean Error for both fidelity gap measurements (e.g., Δ_{AUROC} and Δ_{AUROC}^{group}). In all cases, we do not consider intersectional groups due to sample size concerns.

Dataset	Outcome Variable	n	d	d'	Protected Attribute (g)
`adult` [39]	Income > 50K	48,842	9	33	Sex (2 groups)
`lsac` [114]	Student passes the bar	20,427	8	14	Race (5 groups)
`mimic` [47]	Patient dies in ICU	21,139	49	49	Sex (2 groups)
`recidivism` [87]	Defendant re-offends	6,150	6	7	Race (2 groups)

Table 1. Binary classification datasets used in our experiments. n is the number of samples, d is the number of variables in the original dataset, and d' is the number of features after one-hot encoding categorical variables.

3.5 Experiments Overview

Since fidelity gaps across subgroups are closely linked to fairness preservation and risks of fairwashing, we design experiments to audit this quantity. We conduct the following experiments in the sections below[2]:

Measuring Fidelity Gaps Between Subgroups: We measure fidelity gaps using metrics defined in 3.3 and 3.4 for four different post-hoc explanation models, and two different blackbox model classes. The aim of this experiment is to study the presence and degree of fidelity gaps in standard explainability methods (Section 4).

Assessing the Impact of Robust Training: We use robust training strategies to train explanation models, and repeat the fidelity gap audits to study if robust training can provide reduced fidelity gaps (Section 5).

Studying Possible Causes for Fidelity Gaps: We analyze the impact of blackbox fairness and presence of protected attribute information in feature representations on the fidelity gap (Section 6).

Simulation Showing Impact of Fidelity Gaps: We conduct a simulation and study the quality of decisions made for groups to examine the impacts of unfair explanation models on real-world decision making (Section 7).

4 EXPLANATION FIDELITY VARIES SIGNIFICANTLY BETWEEN SUBGROUPS

Experimentally, we find that fidelity gaps indeed vary by group in many settings. To show this, we train four post-hoc explainability methods (two local, two global) to explain two different blackbox models trained on the four standard fairness benchmark tabular datasets described in Table 1. Following Aïvodji et al. [5], we randomly split each dataset into four subsets: a training set for blackbox models (50%), a training set for explanation models (30%), a validation set for explanation models (10%), and a held-out test set for evaluating both blackbox and explanation models (10%). For each dataset, we train both a Neural Network (NN) and a Logistic Regression (LR) model to serve as blackboxes. See Section B.3 in the Appendix for details on the training regimes, hyperparameter settings, and evaluation metrics for each. In the following sections, we describe the explainability models and fidelity gaps observed.

4.1 Local Explanation Models

Local explanation models explain individual predictions from classifiers by learning an interpretable model locally around each prediction. In our experiments, we consider LIME [93] and SHAP [70], which are popular methods that use linear models to elicit each feature's contribution to the blackbox model's prediction. More details are in Appendix B.1.

Experiment Setup. We measure fidelity gaps between subgroups using the two key metrics introduced in Section 3.4 (see Definitions 3.3 and 3.4). For each, we select three performance measures: Accuracy ($\Delta^{\text{group}}_{\text{Acc.}}$) following prior work [4], mean residual error ($\Delta^{\text{group}}_{\text{Err.}}$), and also include AUROC ($\Delta^{\text{group}}_{\text{AUROC}}$) as a threshold-independent metric. A full table with all metrics can also be found in the Appendix (Section D). For accuracy, we use a threshold of 0.5. Since the

[2]Code: https://github.com/MLforHealth/ExplanationsSubpopulations

Dataset	Blackbox Classifier	$\Delta_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{AUROC}}$	$\Delta^{\text{group}}_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{Err.}}$
adult	Logistic Regression	**0.8% ± 0.0%**	0.1% ± 0.0%	**2.4% ± 0.1%**	**1.9% ± 0.0%**
	Neural Network	**6.9% ± 0.7%**	**3.0% ± 1.2%**	**20.6% ± 2.0%**	**0.8% ± 0.5%**
lsac	Logistic Regression	**2.0% ± 1.0%**	0.0% ± 0.0%	**1.5% ± 0.5%**	**1.5% ± 0.1%**
	Neural Network	**21.4% ± 4.4%**	**6.6% ± 1.2%**	**12.2% ± 2.2%**	**3.8% ± 1.2%**
mimic	Logistic Regression	**0.4% ± 0.6%**	**3.0% ± 1.8%**	**1.1% ± 0.3%**	**2.0% ± 0.1%**
	Neural Network	**0.8% ± 0.4%**	**1.7% ± 1.5%**	**1.4% ± 0.7%**	**1.7% ± 0.5%**
recidivism	Logistic Regression	0.1% ± 0.1%	0.0% ± 0.0%	0.3% ± 0.2%	0.3% ± 0.0%
	Neural Network	**0.9% ± 0.3%**	**0.7% ± 0.3%**	**2.4% ± 0.7%**	**1.1% ± 0.1%**

Table 2. Performance fidelity gaps across subgroups for *LIME local explanations* using all available features. ± denotes standard deviation computed over 5 replications. Fidelity gaps are significant (one-sided Wilcoxon signed-rank tests at $p < 0.05$; marked in **bold**) between all five groups in the lsac dataset, and between two sensitive groups in other three datasets. $\Delta_{Acc.}$ denotes the maximum fidelity gap of subgroups from average (in terms of accuracy at 0.5 threshold), and Δ_m^{group} is the mean fidelity gap between subgroups using metric m.

four datasets are imbalanced, we use AUROC for model selection while tuning all hyperparameters. Non-zero fidelity gaps indicate disparities across groups in the explanation models.

Results. First, we find that LIME disproportionately favors different groups, as shown in Table 2, where the maximum accuracy gap ranges from 0-21.4%. This confirms that explanation quality can dramatically differ by subgroup, even without access to group-membership data. Furthermore, the AUROC/Accuracy between protected groups also ranges significantly (0-6.6%/0.3-12.2%), indicating that members of protected groups are disadvantaged in terms of explanations. Hence, when explanations are judged to be "high quality" based on average fidelity, it might be misleading and lead to errors in decision-making. Bolded non-zero fidelity gaps from are also significantly greater than 0 with a one-sided Wilcoxon signed-rank test at $p < 0.05$.

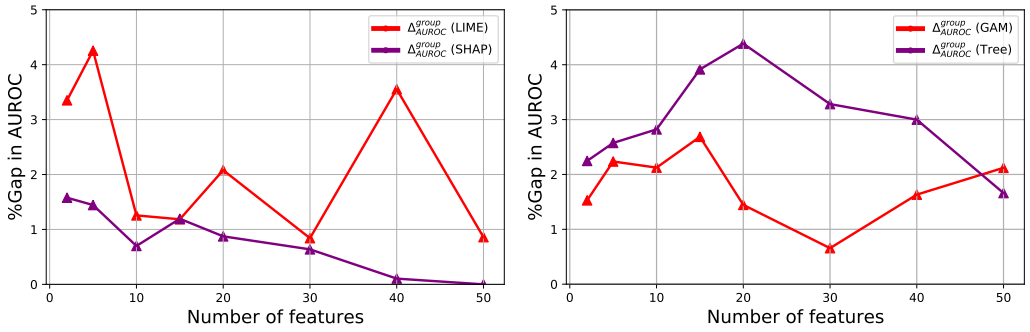
Second, as expected, SHAP's gaps are consistently zero. This is because the blackbox and explanation models are trained using identical features, in which case consistency is guaranteed [70]. However, using a subset of features to train the explanation model often leads to more useful explanations [115]. This increases the gaps significantly, as shown in Figure 2, indicating that SHAP can also suffer from significant gaps in fidelity when used in practice. Since LIME considers sparsity as well, we also run this same experiment for LIME and find that fewer features are indeed associated with larger fidelity gaps (Fig.2). Increasing sparsity is a common approach in training explanation models and these experiments indicate that this technique alone may contribute to substantially worse fidelity gaps.

Third, we observe that the fidelity gaps in AUROC tend to be lower for the logistic regression blackbox, possibly because the linearity of the local surrogate models matches logistic regression better than the neural network. Note that the overall fidelity of all models are greater than 85% (see Table 6 in the Appendix).

4.2 Global Explanation Models

Global explanation methods train one new surrogate model that approximates the behavior of a blackbox model. This surrogate model should itself be easily understood, and can then be used instead of the blackbox at test time (more background in Appendix B.1).

Experiment Setup. In this experiment, we generate global explanations using two popular choices of interpretable surrogate models: Generalized Additive Model (GAM) [48] and a sparse decision tree (Tree) [82]. GAM combines linear models of different variables during explanation [105], while Tree uses a low-depth, sparse, decision tree. We evaluate



(a) Local Explanations (b) Global Explanations

Fig. 2. The effect of varying the number of features on fidelity gaps using the `mimic` dataset with a neural network blackbox model. For (a) local explanation models, using fewer features leads to worse fidelity gaps. We observe larger fidelity gaps across subgroups with sparser models, i.e., fewer features in (a) local explanation models. For (b) global explanation models, the gap varies with number of features. We also observe similar trends on other datasets (see Appendix F).

the fidelity of each global method with the original blackbox and compare across subgroups. As with the local methods, we use both Accuracy and AUROC to evaluate fidelity gaps.

Results. First, we find that the fidelity gap between subgroups differs substantially from the average for the global explanation models, as shown in Table 3 where the accuracy gap ranges from 0-13.5%. We again observe that AUROC and accuracy vary substantially between protected subgroups (0-8.1% and 0.1-7.4% for protected groups such as sex and race groups in each dataset). This is especially true for more imbalanced subgroup proportions: having more subgroup categories leads to more disadvantage in protected groups, particularly when the classes are imbalanced themselves (e.g. `lsac`).

Second, we find that using fewer features (e.g., 15 in Fig. 2) may lead to larger gaps in performance between subgroups in sparse decision trees (Trees), bolstering prior findings on training trustworthy models [21]. Hence, the gaps shown in Tables 2 and 3 are likely underestimates when using fewer dimensions in explanation models, which is common. Interestingly, the subgroup with the lowest-quality explanations is not always the minority subgroup—which may be the most disadvantaged—in the datasets for fair ML. We expand this finding in Table 12 in the Appendix. Additionally, we see that subgroup gaps occur even after training blackbox models with a balanced number of data points from each subgroup for both global and local explanation models (see Table 10).

5 BALANCED AND ROBUST TRAINING REDUCES FIDELITY GAPS

Balanced and robust training methods could provide a path towards improving fidelity gaps, thereby learning fairer explanations [1, 45]. We showcase two such robust training methods, one for local methods and one for global methods. For both cases, we choose hyperparameters that maximize the worst-case fidelity across all groups.[3] Ultimately, our experiments indicate that while robust training improves fidelity gaps sometimes, they remain largely pervasive.

5.1 Robust Local Explanation Models

Experiment Setup. We train a more-robust version of LIME, using Just Train Twice (JTT) [65], a two-stage training paradigm for training robust ML Models. First, we train an identification model via empirical risk minimization. Then,

[3]The overall fidelity is not significantly affected by either training approach.

Dataset	Blackbox Classifier	Expl. Model	$\Delta_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{AUROC}}$	$\Delta^{\text{group}}_{\text{Acc.}}$	$\Delta^{\text{group}}_{\text{Err.}}$
adult	Logistic Regression	GAM	0.1% ± 0.0%	0.0% ± 0.0%	0.3% ± 0.0%	0.1% ± 0.0%
	Logistic Regression	Tree	**1.5% ± 0.1%**	**2.9% ± 0.4%**	**4.5% ± 0.2%**	**1.1% ± 0.1%**
	Neural Network	GAM	**0.8% ± 0.2%**	**0.5% ± 0.3%**	**2.4% ± 0.5%**	0.3% ± 0.2%
	Neural Network	Tree	**1.1% ± 0.1%**	**0.6% ± 0.4%**	**3.4% ± 0.2%**	**0.5% ± 0.4%**
lsac	Logistic Regression	GAM	**0.9% ± 0.9%**	0.0% ± 0.0%	**0.6% ± 0.4%**	**0.7% ± 0.3%**
	Logistic Regression	Tree	**3.7% ± 3.1%**	**1.1% ± 0.4%**	**2.8% ± 0.7%**	**1.8% ± 0.5%**
	Neural Network	GAM	**13.5% ± 0.9%**	**5.2% ± 1.2%**	**7.3% ± 1.0%**	**3.9% ± 2.6%**
	Neural Network	Tree	**11.5% ± 2.7%**	**5.8% ± 2.1%**	**7.4% ± 1.2%**	**4.9% ± 2.0%**
mimic	Logistic Regression	GAM	0.5% ± 0.1%	0.4% ± 0.1%	**0.9% ± 0.1%**	0.4% ± 0.2%
	Logistic Regression	Tree	**0.6% ± 0.0%**	**8.1% ± 0.8%**	**1.2% ± 0.1%**	**1.9% ± 0.0%**
	Neural Network	GAM	**1.2% ± 0.3%**	**1.8% ± 1.2%**	**2.2% ± 0.6%**	**0.9% ± 0.3%**
	Neural Network	Tree	**1.1% ± 0.5%**	**3.0% ± 1.5%**	**2.0% ± 0.9%**	**1.9% ± 0.9%**
recidivism	Logistic Regression	GAM	0.1% ± 0.0%	0.1% ± 0.0%	0.3% ± 0.0%	0.5% ± 0.0%
	Logistic Regression	Tree	0.0% ± 0.0%	0.4% ± 0.0%	0.1% ± 0.0%	**1.2% ± 0.0%**
	Neural Network	GAM	0.2% ± 0.2%	0.4% ± 0.2%	0.6% ± 0.6%	**1.1% ± 0.4%**
	Neural Network	Tree	**0.9% ± 0.3%**	**1.0% ± 0.9%**	**2.3% ± 0.7%**	**1.4% ± 0.3%**

Table 3. Fidelity gaps across subgroups for *global* explanation models GAM and Tree. ± denotes standard deviation computed over 5 replications. Fidelity gaps are significant (one-sided Wilcoxon signed-rank tests at $p < 0.05$; marked in **bold**) for all five groups in the lsac dataset, and between two sensitive groups in the other three datasets with both global explanation models. $\Delta_{Acc.}$ denotes the maximum fidelity gap of subgroups from average (in terms of accuracy at 0.5 threshold), and Δ_m^{group} is the mean fidelity gap between subgroups using metric m.

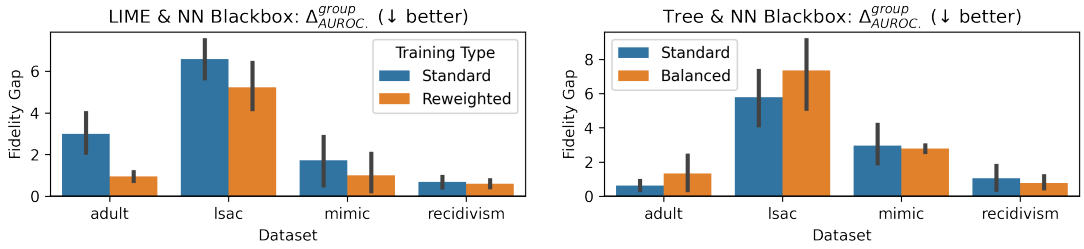
we extract its set of misclassified training examples. A final model is then trained by upsampling these misclassified examples, scaled by a hyperparameter λ. This reweighted loss is designed to make the second model more robust. We use JTT to train LIME's local linear approximations, using linear models for both the identification and final models.

Results. JTT successfully reduces gaps on three datasets with a NN blackbox model, as shown in Figure 3. Interestingly, this is not the case for the recidivism dataset, where JTT does not reduce the gaps and performs the same as standard training. With LR blackboxes (shown in Appendix 10), the fidelity gaps are already small, so JTT is less impactful. However, non-zero gaps between 1-2% still persist (e.g., NN blackbox on the adult dataset), indicating that the error-prone regions did not generalize to the test setting. Measuring fidelity gaps is therefore still critical, even if an explanation model is trained to be robust.

5.2 Robust Global Explanation Models

Experiment Setup. We next study balanced training for the global explanation method Tree. We rebalance the explainability training sets for each dataset by randomly oversampling minority groups, a common approach for improving test error on minority subpopulations [45, 112]. This way, the training set in each case consists of an equal number of examples from each protected group. Then, we train a Tree model to explain each blackbox model using these balanced datasets.

Results. As shown in Figure 3, we find that this common rebalancing approach to does not reduce gaps significantly across the board. Still, some cases look more promising than others. For example, mimic with NN which indicates this may be a fruitful direction for learning fairer explanations. This is especially true for cases like mimic with LR, where rebalancing the training set increases the fidelity gap substantially (see Figure 10 in Appendix).



(a) Local Explanations (LIME)

(b) Global Explanations (Tree)

Fig. 3. AUROC Fidelity gaps across subgroups with and without robust training for (a) LIME and (b) Tree-based Models. Improvements are significant with robust training for `adult` dataset in (a) with a Wilcoxon-signed rank test at $p < 0.1$ level. However, balanced training does not help for most datasets in (b). Error bars indicate 95% confidence intervals.

6 ON POSSIBLE CAUSES FOR FIDELITY GAPS

In our fidelity gap audits in prior sections, we noticed that the fidelity gaps are largest for the least-fair blackboxes (`adult` and `lsac` datasets; Tables 3 and 2). This indicates a potential relationship between the fairness of the blackbox and explainability models. To explore this further, we study the associations between blackbox fairness and fidelity gaps across subgroups in this section. First, we train fair models, and observe that significant non-zero fidelity gaps still persist. Second, we study if protected group information – e.g. if a data point belongs to a Male or Female individual – can be predicted from the *feature representations* alone, following prior work in fair representation learning [71, 118]. We find that complex mechanisms by which protected group information can be indirectly predicted could be contributing factors to the fidelity gaps observed.

6.1 Training Fair Blackbox Models

Our experiments in previous sections (see Tables 2 and 3) indicate that fidelity gaps across subgroups occur regardless of the blackbox model's fairness with respect to groundtruth label predictions. For example, a logistic regression model trained on the `mimic` dataset is fair with respect to the sex (Table 4 in Appendix; DP gap of 1%). However, fidelity gaps are non-zero across sex subgroups with a sparse decision tree global explanation (8.1%). Strikingly, the gaps in fidelity AUROC often exceed the gaps in AUROC of the blackbox models themselves (e.g., `mimic` dataset with the Tree model, the difference in classification performance of the blackbox between Male and Female individuals is 3.6%, while the fidelity gap between subgroups is 8.1%). However, we do observe larger fidelity gaps in datasets where blackbox models are more accurate but less fair (e.g. `adult` with an absolute DP gap of 16-17%).

Experiment Setup. To investigate this further, we train debiased neural network blackbox models for `lsac` and `mimic`: both highly-imbalanced by class label, and characterized by the largest and smallest gaps in blackbox model AUROC respectively. Adversarial debiasing following methodology proposed by Zhang et al. [118][4] is utilized, wherein an adversary tries to predict the protected group information from classification predictions and labels, while the main classification model (our blackbox model) jointly predicts the primary classification outcome. We use demographic parity as the desired fairness definition. Our results are shown in Table 13, which reports the performance of the fair(er) blackbox classifiers.

[4]We use an open-sourced implementation: https://github.com/Trusted-AI/AIF360

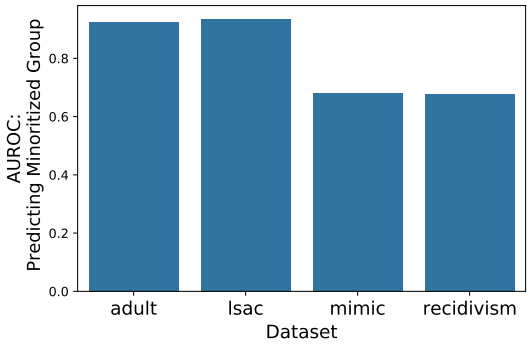
Results. We debias neural network blackbox models to be fairer, where a model is deemed to be fair if it has an absolute DP gap close to 1% (9% and 0.8% after debiasing for `lsac` and `mimic` respectively; improved from 14% and 2%). We find that despite fair training, fidelity gaps remain (Table 4, though they are significantly reduced in most cases: the fidelity gap in accuracy decreases from 2% to 1% for `mimic` (Tree) and 7.4% to 0.8% for `lsac`). Note that these results are dependent on our choice of fairness criterion: particularly, for the `lsac` dataset we find that overall performance is reduced to achieve parity (Table 13 in Appendix; for absolute DP gap close to 1%, over 99% of the blackbox predictions are that the student passes the bar). This indicates that while fair blackboxes could potentially reduce fidelity gaps across subgroups, there might be trade-offs and more potential causes for such gaps. We explore this in Section 6.2.

6.2 Performance of Predicting Protected Attributes from Feature Representations Alone

Experiment Setup. One way to achieve group fairness is by removing group information from or debiasing the representations (e.g. with the use of an adversary [71, 109]). Here, we quantify the amount of group information that is present in the data. This is relevant as the absence of group information is a sufficient condition to achieving fairness parity according to standard metrics in fair machine learning. For example, consider equality of opportunity for the positive class, which can be written as $\hat{Y} \perp G|Y = 1$. If we have $X \perp G|Y = 1$, then equality of opportunity is achieved for any form of the classifier $\hat{Y} = f(X)$, including E the explanation model and B the blackbox model. Therefore, if no protected group information is present in the positive examples, then the explanation fidelity would not differ between protected groups for positive examples [71]. In this experiment, we first compute the accuracy of detecting the protected group information from all datasets (with a cross-validated model). Then, we select features that have zero mutual information with respect to the protected attribute, and only use these in training the blackbox and explanation model. We expect that the performance of predicting group information from these features will be low. Then, we compute the fidelity gaps – this allows us to answer the question: do fidelity gaps exist when there is low group information in the data?

Results: First, we observe that in all cases the prediction AUROC is significantly greater than 0.5 (see Table 4) in identifying the minority group. This indicates that the protected group information – e.g. if the datapoint belongs to a Male/Female individual – can be predicted with good performance from the feature representations alone[5]. Since past work has shown that group information might be indirectly constructed and used by explanations [59], this is important to consider. Second, we only use features that have zero mutual information with respect to protected attribute labels in the `lsac` dataset for training blackbox and explanation models. We see that all models output a single-class prediction which limits our ability to make meaningful conclusions about the impact of group information on this dataset – note the fidelity gaps are technically zero, though the explanations are trivial. We repeat the same procedure with the `mimic` dataset by selecting 10 features. The AUROC of predicting protected attribute (sex) from these features is low (0.54; 0.56 and 0.54 for features from positive and negative class respectively). With this representation, we train both NN and LR blackbox models, and GAM/Tree global explanation models. We observe that accuracy-based fidelity gaps ($\Delta_{Acc.}$, $\Delta_{Acc.}^{group}$) decrease to low values not much higher than zero (to 0-0.6% with GAM and Tree; full table in Appendix 15 while blackbox model's AUROC is greater than 0.7). This indicates that fidelity gaps decrease when there is less group information in data representations. However, non-zero fidelity gaps in AUROC still persist for Tree-based models (up to 6.6%). This is due to low prevalence of positive class predictions with the blackbox model on using the reduced data representation (≈3% positive class at 0.5 threshold), which has a large impact on AUROC (since it is a ranking-based

[5]This was also observed when conditioned on only positive or negative label classes



(a) AUROC in detecting minoritized protected group

Dataset	Expl	$Fidelity^{Acc.}$	$\Delta^{\text{group}}_{AUROC}$	$\Delta^{\text{group}}_{Acc.}$
lsac	GAM	96.6	1.4	1.5
	Tree	96.9	7.0	0.8
mimic	GAM	96.2	0.6	1.0
	Tree	94.9	4.1	1.6

(b) Fidelity gaps with fair blackbox models

Fig. 4. We find minoritized protected groups can be detected with high AUROC from feature representations alone in (a). As a result, non-zero fidelity gaps persist even when underlying blackboxes are fair as seen in (b).

metric, and sensitive to degree of imbalance). We highlight that more experiments using interpretable, completely group-independent representations (i.e., an AUROC of 0.5 in predicting protected attribute labels) that still have high groundtruth predictive capability are required to accurately quantify the impact of group information on AUROC-based fidelity gaps. We also note that class imbalance – and varying degrees of class imbalance for data subgroups – may be an important factor. Our findings indicate that fidelity gaps persist across a range of class-imbalance ratios, but we leave the estimation of the effect of varying degrees of class imbalance (or positive-class prevalence) across subgroups on explanation fairness for future work.

7 SIMULATING THE REAL-WORLD IMPACT OF BIASED EXPLANATIONS

Unfair explanation models can have negative effects on real-world decision making. To demonstrate this, we conduct a simulation study of ML-assisted law school admissions using the `lsac` dataset [114]. Such systems are already being used in many cases [73, 78]. Our results clearly show that worse decisions are made for members of disadvantaged groups when explanations are less fair.

Experiment Setup. To set up our simulation study, we consider an admissions officer that uses a blackbox model that predicts whether a student will pass the bar, though this prediction may be incorrect. The admissions officer also has an explanation of the model's prediction, which may have low fidelity. The blackbox model's performance and the explanation fidelity can vary between protected groups—we vary these parameters in this experiment. We assume that the admissions officer then admits students solely based on their perceived likelihood of the applicant passing the bar, without any knowledge of the applicant's demographics. We assume parameters for the probability that the officer ultimately makes the correct decision. We obtain these parameters from prior user studies assessing the impact of explanations on human+AI decision making accuracy for a different task [10], but believe they serve as reasonable estimates to display the anchoring effect of decisions with explanations observed across a variety of decision-making settings [10, 85]. For further details, please see Appendix K.

To simulate the effect of fidelity gaps on decision-making accuracy, we vary the maximum fidelity gaps between the two groups (males and females) and the average from 0% to 15%, assuming an average fidelity of 85% across groups. We then compute the admissions officer's resulting decision-making accuracy for males and females. We use sex as the

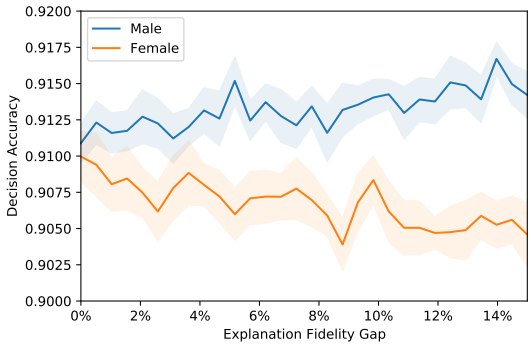


Fig. 5. Effect of fidelity gap size on a simulated admissions officer's decision accuracy between males and females using a neural network blackbox. Note that larger fidelity gaps lead to larger decision accuracy gaps between males and females; fidelity gaps could disadvantage different groups in practice. Each line is paired with 95% confidence intervals across 20 simulations.

protected attribute of interest for the simulation as both groups pass the bar equally in the dataset, so decision-making accuracy is a fair performance metric.

Results. As shown in Figure 5, we find that larger fidelity gaps lead to larger decision accuracy gap between groups. So when explanations are less fair, disadvantaged groups may be targeted by worse decisions. With over 60,000 law school applicants in the U.S. in a typical year [29], over 200 applications would be wrongly admitted/rejected based solely on the fairness of the explanation model, according to this simulation. Fidelity gaps should therefore be used as fairness metrics for explanation methods: minimizing these gaps leads to fairer decisions. However, we emphasize that the findings of this simulation are based on some strong assumptions (e.g., reliance on parameters extracted, anchoring effect existence in this admission setting, etc.). Real-world user studies are required to validate these expected findings rigorously across a variety of decision-making setups.

8 DISCUSSION

8.1 Takeaways for ML Practitioners

Analyze subgroup fidelities. Our results suggest that ML practitioners using post-hoc explainable models to interpret blackbox models should carefully analyse the fidelity of commonly-used explanations for different groups separately. Especially if there is a target subgroup of interest. If a fidelity gap exists, practitioners should carefully consider its source [113], and, where possible, take measures to minimize the impacts on downstream decision-making [90]. We also highlight the importance of carefully choosing the metric for measuring fidelity (e.g., accuracy, AUROC, precision, etc.): different metrics may be affected by properties of the dataset and hence predictions from a blackbox/explanation model (e.g., class imbalance) differently.

Consider the explanation model. Overall, our findings indicate the existence of fidelity gaps between subgroups is both a *model* and a *data* issue. From Section 4, we find that fidelity gaps can vary greatly for the same dataset depending on the explanation model used, and our results in Section 5 show that algorithms that seek to improve worst-case group performance may be a promising direction in reducing fidelity gaps. As such, we recommend careful selection and testing of various explanation models in order to select an equitable model with high overall fidelity.

In addition, model hyperparameters should also be carefully tuned. For example, in models like LIME, there are several hyperparameters that can affect fidelity gaps, such as the sampling variance (Figure 9 in Appendix), the number of perturbations, or number of features in the explanation. Exploring the effect of these hyperparameters on explanation quality and fairness is a promising direction of future work. Lastly, extending prior work on methods for fair supervised ML models [20, 65, 71, 80, 118], we call for similar approaches to training fair and explainable local and global explanation models which have reduced fidelity gap in addition to high overall fidelity.

Consider the data. Our results in Section 6 indicate that fidelity gaps also depend on data representations. Because some feature representations cause smaller fidelity gaps, practitioners should carefully consider the features used to learn both the blackbox and explanation models [44]. As machine learning models can encode historical biases present in the training corpora [37], it is crucial to consider the source of such potential biases, and, if possible, take actions to correct them by collecting additional data in a fairness-aware way [7, 54].

8.2 Implications of Fidelity Gaps

Algorithmic Modifications to Train Fair Explanation Models with Low Fidelity Gaps. While we benchmark robust training as an attempt to mediate the fidelity gaps across subgroups, we posit that data distribution-aware methods could lead to lower, less significant fidelity gaps. For example, recent work in causal bootstrapping [52] show potential to reduce confounding biases in ML model training given causal knowledge [43]. Similar strategies relying on partial or complete knowledge of the data generation graph [40] could prove effective in selecting features and training examples to train fair explanation models.

Post-processing solutions to standard explanation model training could also prove effective, similar to recent work in the space of improving worst-case generalization [75]. However, such solutions need to be appraised carefully to ensure that the resulting models are both *fair* and remain *interpretable* to users [110].

An interesting follow-up question is whether it is possible to have zero fidelity gaps—perfect worst-case generalization—while retaining good average fidelity under standard training settings. Zero fidelity gaps are possible, of course, when the blackbox and explanation models are identical. However in more-realistic scenarios, fidelity gaps may simply depend on the data distributions [75]. For example, rare subgroups may be more diffidult to approximate, and will naturally have lower fidelity than others [50, 102].

Fidelity Gaps as an Evaluation Metric. We focus mainly on evaluating the fairness of explanation models using the *fidelity gap* as a metric, assuming that models with smaller fidelity gaps are more desirable. However, recent work in group fairness has found that trying to achieve equal performance for all subgroups tends to worsen welfare for all [28, 51, 119]. Such a fairness/accuracy trade-off is well-documented in the algorithmic fairness literature [56, 121]. We posit that there is likely a similar trade-off between the fidelity gap and the overall fidelity of an explanation model. In such cases, motivated by definitions such as minimax Pareto fairness [32, 72], it may be more appropriate to select explanation models that maximize the fidelity of the worst-case group.

Human Implications of Fidelity Gaps. Explainable ML models form an integral part of sociotechnical systems, given their user-facing nature [16]. Several works have studied the utility of explanations in human–AI joint decision-making [10]. However, the potential failure modes we identify—fidelity gaps leading to worse explanations for some groups—need to be studied further in the context of real human decision-making (in addition to the simulations we conduct). For accurate decision-making in practice, like learning to defer decisions to an expert [77], it is important to communicate clearly and provide end-users with details of performance caveats [63]. This requires collaboration between

computer scientists and scholars working in the space of computer-mediated communication. A more design-centric approach is required to bridge the gap between researchers and consumers of these models [25].

9 CONCLUSION

In this work, we investigate fairness properties of post-hoc explainability methods. We ultimately find that significant gaps in performance exist between groups, indicating that some groups receive better explanations than others. First, we demonstrate experimentally that significant gaps occur in the two main branches of explanation methods using four explainability methods on four common datasets and two blackbox models. Second, we present a study of robust training methods for improving these gaps. We find that these methods can improve the fairness of explanation models in some cases. Third, using a simulation study, we demonstrate that improving explanation fairness could substantially improve decision making accuracy for underserved groups. Finally, we pose promising directions enhancing post-hoc explainability methods; future work should focus on ensuring explanation quality does not suffer according to group membership while remaining reliable and accurate.

ACKNOWLEDGMENTS

Aparna Balagopalan is supported by a grant from IBM-Watson. Haoran Zhang is supported by a grant from the Quanta Research Institute. Dr. Frank Rudzicz is supported by a CIFAR AI Chair. Dr. Marzyeh Ghassemi is funded in part by Microsoft Research, and a Canadian CIFAR AI Chair held at the Vector Institute. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. We would like to thank Hammaad Adam, Bret Nestor, Natalie Dullerud, Vinith Suriyakumar, Nathan Ng, and three anonymous reviewers for their valuable feedback.

REFERENCES

[1] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485* (2020).

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning (ICML)*. 60–69.

[3] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 559–560.

[4] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*. PMLR, 161–170.

[5] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. 2021. Characterizing the risk of fairwashing. *arXiv preprint arXiv:2106.07504* (2021).

[6] Benjamin Alarie, Anthony Niblett, and Albert H Yoon. 2016. Using machine learning to predict outcomes in tax law. *Can. Bus. LJ* 58 (2016), 231.

[7] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. 2020. Fair active learning. *arXiv preprint arXiv:2001.01796* (2020).

[8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. (2021).

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[11] Richard Berk. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13, 2 (2017), 193–216.

[12] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[13] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29, 2 (2021), 149–169.

[14] Tiago Botari, Frederik Hvilshøj, Rafael Izbicki, and Andre CPLF de Carvalho. 2020. MeLIME: meaningful local explanation for machine learning models. *arXiv preprint arXiv:2009.05818* (2020).

[15] Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[16] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[17] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[18] Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[19] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57, 1 (2021), 203–216.

[20] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[21] Chun-Hao Chang, Sarah Tan, Ben Lengerich, Anna Goldenberg, and Rich Caruana. 2021. How interpretable and trustworthy are gams?. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 95–105.

[22] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002* (2018).

[23] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2020. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4 (2020).

[24] John Chen, Ian Berlot-Attwell, Safwan Hossain, Xindi Wang, and Frank Rudzicz. 2020. Exploring Text Specific and Blackbox Fairness Algorithms in Multimodal Clinical NLP.

[25] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2022. Interpretable Machine Learning: Moving from mythos to diagnostics. *Queue* 19, 6 (2022), 28–56.

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

[27] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[28] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[29] The Law School Admission Council. 2018. Legal Education Data Library. https://www.lsac.org/data-research/data/current-volume-summaries-region-raceethnicity-gender-identity-lsat-score

[30] Mark Craven and Jude Shavlik. 1995. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems* 8 (1995), 24–30.

[31] Jessica Dai, Sohini Upadhyay, Stephen H Bach, and Himabindu Lakkaraju. 2021. What will it take to generate fairness-preserving explanations? *arXiv preprint arXiv:2106.13346* (2021).

[32] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 66–76.

[33] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[34] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. 1998. Persuasiveness of expert systems. *Behaviour & Information Technology* 17, 3 (1998), 155–163.

[35] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[36] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[37] Jad Doughman, Wael Khreich, Maya El Gharib, Maha Wiss, and Zahraa Berjawi. 2021. Gender Bias in Text: Origin, Taxonomy, and Implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. 34–44.

[38] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. *Commun. ACM* 63, 1 (2019), 68–77.

[39] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[40] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

[41] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[42] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.

[43] Sindhu CM Gowda, Shalmali Joshi, Haoran Zhang, and Marzyeh Ghassemi. 2021. Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 606–616.

[44] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. 2.

[45] Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Balancing out Bias: Achieving Fairness Through Training Reweighting. *arXiv preprint arXiv:2109.08253* (2021).

[46] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[47] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data* 6, 1 (2019), 1–18.

[48] Trevor J Hastie and Robert J Tibshirani. 2017. *Generalized additive models*. Routledge.

[49] Andreas Holzinger. 2018. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 55–66.

[50] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058* (2020).

[51] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.

[52] Guido Imbens and Konrad Menzel. 2018. *A Causal Bootstrap*. Technical Report. National Bureau of Economic Research, Inc.

[53] Joseph Jamison. 2017. Applying Machine Learning to Predict Davidson College's Admissions Yield. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. 765–766.

[54] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 306–316.

[55] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* (2015).

[56] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 100–109.

[57] Sanjay Krishnan and Eugene Wu. 2017. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. 1–6.

[58] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. 2020. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114* (2020).

[59] Himabindu Lakkaraju and Osbert Bastani. 2020. " 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*. 79–85.

[60] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).

[61] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[62] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066* (2015).

[63] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[64] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[65] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*. PMLR, 6781–6792.

[66] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*. PMLR, 4114–4124.

[67] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[68] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 150–158.

[69] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.

[70] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems* 30 (2017), 4765–4774.

[71] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*. PMLR, 3384–3393.

[72] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[73] Barbara Martinez Neda, Yue Zeng, and Sergio Gago-Masague. 2021. Using Machine Learning in Admissions: Reducing Human and Algorithmic Bias in the Selection Process. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 1323–1323.

[74] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[75] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. 2020. Overparameterisation and worst-case generalisation: friend or foe?. In *International Conference on Learning Representations*.

[76] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[77] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7076–7087. https://proceedings.mlr.press/v119/mozannar20b.html

[78] DJ Pangburn. 2019. Schools are using software to help pick who gets in. what could go wrong? https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-help-pick-who-gets-in-what-could-go-wrong

[79] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*.

[80] William Paul and Philippe Burlina. 2021. Generalizing Fairness: Discovery and Mitigation of Unknown Sensitive Attributes. *arXiv preprint arXiv:2107.13625* (2021).

[81] Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London* 58, 347-352 (1895), 240–242.

[82] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[83] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5684–5693.

[84] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. 2018. Model agnostic supervised local explanations. *arXiv preprint arXiv:1807.02910* (2018).

[85] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. *CHI 2021* (2021).

[86] Romila Pradhan, Jiongli Zhu, Boris Glavic, and Babak Salimi. 2021. Interpretable Data-Based Explanations for Fairness Debugging. *arXiv preprint arXiv:2112.09745* (2021).

[87] ProPublica. 2019. Compas recidivism risk score data and analysis.

[88] Nikaash Puri, Piyush Gupta, Pratiksha Agarwal, Sukriti Verma, and Balaji Krishnamurthy. 2017. Magix: Model agnostic globally interpretable explanations. *arXiv preprint arXiv:1706.07160* (2017).

[89] Hamed Rahimian and Sanjay Mehrotra. 2019. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659* (2019).

[90] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169, 12 (2018), 866–872.

[91] Shubham Rathi. 2019. Generating counterfactual and contrastive explanations using SHAP. *arXiv preprint arXiv:1906.09293* (2019).

[92] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " 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*. 1135–1144.

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[94] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[95] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. 2020. Explainable machine learning for scientific insights and discoveries. *Ieee Access* 8 (2020), 42200–42216.

[96] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 2662–2670.

[97] Alvin E Roth. 1988. *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

[98] Cynthia Rudin. 2014. Algorithms for interpretable machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1519–1519.

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

[100] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731* (2019).

[101] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. arXiv:1911.08731 [cs.LG]

[102] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair PCA: one extra dimension. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 10999–11010.

[103] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International Conference on Machine Learning*. PMLR, 3145–3153.

[104] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. 2021. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 723–734.

[105] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 303–310.

[106] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.

[107] Lukas Tuggener, Mohammadreza Amirian, Katharina Rombach, Stefan Lörwald, Anastasia Varlet, Christian Westermann, and Thilo Stadelmann. 2019. Automated machine learning in practice: state of the art and recent results. In *2019 6th Swiss Conference on Data Science (SDS)*. IEEE, 31–36.

[108] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (2016), 349–391.

[109] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[110] Caroline Wang, Bin Han, Bhrij Patel, Feroze Mohideen, and Cynthia Rudin. 2020. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *arXiv preprint arXiv:2005.04176* (2020).

[111] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–6.

[112] Qiong Wei and Roland L Dunbrack Jr. 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS one* 8, 7 (2013), e67863.

[113] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: a trade-off revisited. (2019).

[114] Linda F Wightman. 1998. *LSAC national longitudinal bar passage study*. Law School Admission Council.

[115] Eric Wong, Shibani Santurkar, and Aleksander Mądry. 2021. Leveraging Sparse Linear Layers for Debuggable Deep Networks. *arXiv preprint arXiv:2105.04857* (2021).

[116] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

[117] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[118] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[119] Haoran Zhang, Natalie Dullerud, Karsten Roth, Lauren Oakden-Rayner, Stephen Pfohl, and Marzyeh Ghassemi. 2022. Improving the Fairness of Chest X-ray Classifiers. In *Conference on Health, Inference, and Learning*. PMLR, 204–233.

[120] Haoran Zhang, Quaid Morris, Berk Ustun, and Marzyeh Ghassemi. 2021. Learning Optimal Predictive Checklists. *Advances in Neural Information Processing Systems* 34 (2021).

[121] Indre Zliobaite. 2015. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723* (2015).

A FAIRNESS PRESERVATION

A.1 Proof

(1) Samples x_i, each associated with a binary valued sensitive attribute $A_i \in [0, 1]$, and $i \in [1, N]$
(2) Pre-trained blackbox classifier, B, which generates predictions $B(x_i)$ for each point
(3) Local explanation model for each test instance, $E_i(x_i)$ or global explanation model $E(x_i)$, trained on tuples $(x_i^*, B((x_i^*))$. For local explanation models, $(x_i^*, B((x_i^*))$ are usually neighboring points around instance of interest, generated synthetically or sampling nearest neighbors in the training set.

In all cases, B and E_i output probabilities $\in [0, 1]$ or binary values.

By definition of explanation models being trained to mimic blackbox models (or be reasonable estimators of B with error minimization of errors such as mean squared error):

$$\forall i : E_i(x_i) = B(x_i) + \epsilon_i,$$

where ϵ_i is the residual error between outputs of blackbox and explanation surrogate model. In the case of local explanation models, E_i could be distinct for each point i. However, for global models, E_i is the same function (E) for all test instances. In the section below, we show how gaps between average and subgroup performance for producing explanations is directly tied to demographic parity metrics of the blackbox model and explanation model(s) in each case.

We want to attain sufficient conditions for DP of the explanation model to reflect that of the blackbox classifier, i.e.,

$$DP_{E(x)}$$

can be used as a proxy for

$$DP_{B(x)}$$

for the same x, i.e. same test instances.

$$DP_{B(x)} = \mathbb{E}_{x_i:A_i=1}[B(x_i)] - \mathbb{E}_{x_i:A_i=0}[B(x_i)] \tag{1}$$

$$DP_{E(x)} = \mathbb{E}_{x_i:A_i=1}[E_i(x_i)] - \mathbb{E}_{x_i:A_i=0}[E_i(x_i)] \tag{2}$$

$$DP_{E(x)} = \mathbb{E}_{x_i:A_i=1}[B(x) + \epsilon_i] - \mathbb{E}_{x_i:A_i=0}[B(x) + \epsilon_i] \tag{3}$$

$$DP_E = \mathbb{E}_{x_i:A_i=1}[B(x_i)] - \mathbb{E}_{x_i:A_i=0}[B(x_i)] + \mathbb{E}_{x \in X:A_i=1}[\epsilon_i] - \mathbb{E}_{x \in X:A_i=0}[\epsilon_i] \tag{4}$$

$$\implies DP_{E(x)} = DP_{B(x)} + Error$$

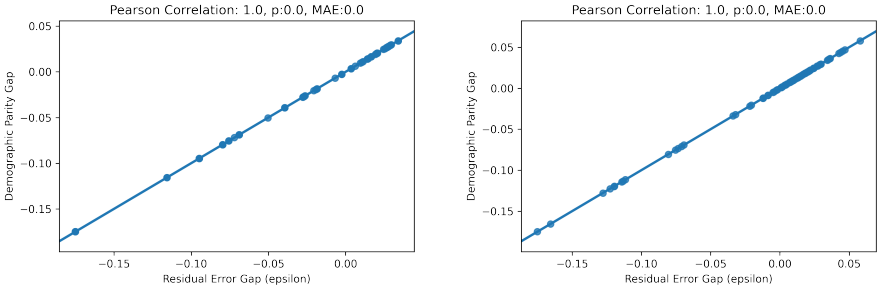
where $Error$=$\mathbb{E}_{x_i \in X:A_i=1}[\epsilon_i] - \mathbb{E}_{x_i \in X:A_i=0}[\epsilon_i]$, and corresponds to difference in average errors for instances i belonging to each group. ϵ_i directly corresponds to the residual error when minimizing the mean square error or cross-entropy error between continous predictions of B and E_i for a test instance i.

Additionally, the expansion above holds for probabilistic fairness definitions [83] when model outputs are probabilistic or constrained to be binary.

A.2 Empirical Validation

We empirically validate the theorem on global explanations model (Tree) that utilize standard estimators during explanation model training/optimization. We observe that observations hold for both probabilistic and generalized versions of fairness metrics [83]. This is pertinent because local explanation models are trained to imitate the blackbox

prediction probabilities. Note that an underlying assumption (which is validated in practice) is that E and B are both defined for the datapoints considered. We observe that all Pearson correlations [81] are significant and consistently 1.



(a) Tree + LR blackbox (b) Tree + NN blackbox

Fig. 6. Comparing gaps in Demographic Parity (DP) of blackbox and explanation model to gaps in residual errors. As per our theorem, these should be the same with 1 and mean absolute error of zero, which is empirically validated. Similar results are observed for all explanation models.

B METHODS AND EXPERIMENTAL SETUP

In the sections below, we describe types of post-hoc explanation models used as in our analyses. All experiments are performed for four tabular binary classification datasets (see Section B.2) commonly used in fair machine learning.

B.1 Post-hoc explanation Models

We experiment with two local and two global explanation models.

B.1.1 Local Explanation Models. Local explanation models explain individual predictions from classifiers by learning an interpretable model locally around each prediction. In our experiments, we consider LIME [93] and SHAP [70], which are popular methods that use linear models to elicit each feature's contribution to the blackbox model's prediction.

Both LIME and SHAP sample points in the neighborhood of query point of interest $\boldsymbol{x}$. Then, an auxiliary linear model is trained that approximates the behavior of the blackbox model on the perturbed examples of $\boldsymbol{x'}$, while achieving low complexity for interpretability: Both local methods consider a proximity measure $\pi_{\boldsymbol{x'}}(\boldsymbol{x})$ between inputs $\boldsymbol{x'}$ and $\boldsymbol{x}$ to define the neighborhood around a specific datapoint $\boldsymbol{x'}$, and $\mathcal{E}$ denote the class of linear models from which we want to choose an explanation model e. In both LIME and SHAP, an auxiliary linear model is trained that approximates the behavior of the blackbox model on the perturbed examples of $\boldsymbol{x'}$, while achieving low complexity for interpretability:

$$\text{argmin}_{E\in\mathcal{E}} L(B, E, \pi_{\boldsymbol{x}}) + \Omega(E)$$

where $\Omega(E)$ penalizes the complexity of explanation E, and L denotes a loss-function weighted by proximity to $\boldsymbol{x'}$.

LIME. Local Interpretable Model-agnostic Explanations (LIME) [93] enforces the faithfulness of explanations to the blackbox model over a set of examples in the simplified input space. More precisely, LIME does this by fitting a sparse linear model to perturbed samples around query point of interest $\boldsymbol{x'}$. Specifically, we first generate $\mathcal{Z} = (\boldsymbol{z}_j, w_j)_{j=1}^{n'}$, where $\boldsymbol{z}_j = \boldsymbol{x'} + \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d)$, and $w_j = \mathcal{D}(\boldsymbol{z}_j, \boldsymbol{x'})$. Here, $\mathcal{D}$ is some distance metric, typically the Euclidean distance.

Taking in k, the number of features we wish to explain, LIME then selects k features from d, and returns an explanation corresponding to the parameters of a linear model $H : \mathbb{R}^k \rightarrow [0, 1]$. Typically, H is a weighted (by w) ridge regression model fit on the sampled datapoints. In this work, we explore the fairness of LIME explanations, varying σ^2, k, and the training procedure of H. Note that since LIME implements a regression, the predicted outputs can lie outside $0 - 1$ range. While the metrics we report do not clip these values, we found that clipping led to the same classification performance. The mean residual error varied with clipping, but is always within 0-0.8% of the reported value.

SHAP. SHapley Additive exPlanations [70] uses SHAP values, which are the Shapley values [70] of a conditional expectation function of the original model. Similar to LIME, (kernel) SHAP uses a linear model to approximate the blackbox model. Note that the explanation model is a linear function of binary variables $E(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$ where $z' = \{0, 1\}^M$ and M is the number of features in the simplified input space. In this approach, $\{\phi_i\}_{i=1}^{M}$ are Shapley values, which are the only possible solution to the optimization problem that satisfies local accuracy, missingness and consistency [70].

B.1.2 Global Explanation Models. Global explanation models generate explanations by learning an interpretable model that imitates blackbox model predictions for a set of datapoints. In our experiments, we consider GAM [48] and a sparse decision tree, which are popular methods that build surrogate models.

Generalized Additive Models (GAM). A generalized additive model (GAM) is a linear model in which the output variable is modelled as a linear combination of on unknown smooth functions of some predictor variables or features. Training involves the inference of these smooth functions.

For a D-dimensonal input $x \in X$, with predicted variable $y, i \in [1, D]$ a GAM can be written as: $\theta(y) = \phi_0 + \sum_{i=1}^{D} \phi_i(x_i)$. where θ is a link function such as logit or identity. GAMs are interpretable because the function asso of each feature on the prediction can be visualized. Following past work [105], we use GAMs to predict or distill decisions made by blackbox models.

Sparse decision tree (Tree). We also fit low-depth, sparse, decision trees to predict blackbox decisions. This is because decision trees are simple to understand and to interpret, and can easily be visualised.

There are three primary reasons why these models may approximate a blackbox poorly: (1) disparities in model classes, (2) variance between training datasets for the blackbox and explainability models, and (3) choices of features used for explanation—some methods create explainable surrogate models by using fewer features.

B.2 Datasets

We conduct experiments using the four tabular binary classification datasets described in Table 1. Each dataset includes ground-truth labels for each instance's group membership, in this case either their sex or their race. All four datasets are commonly used by prior work on fairness in machine learning [58, 104]. Given the extremely imbalanced nature of the `lsac` dataset (95% positive class), we find that even highly-optimized blackbox models using the 6 features (without any fairness criterion) tend to produce less than 1% negative class predictions. As a result, the AUROC becomes very unstable across consecutive runs. To avoid this, we balance the label-proportions for both classes by randomly upsampling the minority class while training the blackbox model (i.e., 50-50 class balance to train the blackbox model) unless specified otherwise. Proportions of positive-class in groundtruth as well as blackbox model predictions are in Table 5 for all datasets.

Following Aïvodji et al. [5], we randomly split each dataset into four subsets: a training set for blackbox models (50%), a training set for explanation models (30%), a validation set for explanation models (10%), and a held-out test set for evaluating both blackbox and explanation models (10%). In our experiments, we omit the sensitive attribute from all models to allow for closer comparison to previous works.

B.3 Training Regimes and Hyperparameter Tuning

B.3.1 Training Blackbox Models. We train two different types of blackbox models on each dataset: Logistic Regression, and a Neural Network. All models are implemented using the *sklearn* python package [82]. We tune hyperparameters using a grid-search with five-fold cross validation. We report results for the Logistic Regression (LR) and Neural Network (NN) models in the main text, and report hyperparameter details and performance measures for each in the Appendix. We one-hot encode categorical features, and standardize continuous features to zero-mean and unit variance. The held-out test sets are used to estimate the accuracy of the blackbox models as well as the generalization of the explanation models. For local explanations, the performance is evaluated only on the query point of interest in each case (and not the synthetic data generated).

B.3.2 Training Explanation Models. We use the default settings for LIME [93][6] and SHAP [70][7] unless specified otherwise. For LIME, we use 5000 synthetic data points for each query point. For SHAP, we use 50 random samples. Note that SHAP requires access to a sample of the training set of models, as well as preprocessing steps for datasets with categorical variables. For global explanation models, we code categorical features as factors in logistic GAM models (continuous features were fit using splines). All features are scaled, and numeric features are standardized before fitting global models. For local models, we consider probabilistic outputs from the blackbox models and for global methods we threshold these predictions at 0.5 (following Aïvodji et al. [5]).

B.3.3 Choice of Evaluation Metrics. As introduced in Section 3.4, we measure fidelity gaps between subgroups using two key metrics. For both metrics (Definitions 3.3 and 3.4), we must also select a performance metric L. We use L to be the AUROC since it is threshold-independent, but also report results using Accuracy as L, following past work in explainability [4]. Given the imbalanced nature of the datasets considered, AUROC is used for hyperparameter tuning unless specified otherwise. Note that a non-zero fidelity gap across subgroups measured using both AUROC and accuracy indicates varying degrees of class separability for subgroups in the explanation model. However, the value of the accuracy fidelity gap is also a function of the decision threshold. Here, we select a threshold value of 0.5 for demonstration purposes, and use the same threshold for all groups. For accurate estimates of overall and subgroup fidelity (gaps), we average across multiple seeds (e.g., prior to performing the *max* operation for Δ_L computation).

C TRAINING BLACKBOX MODELS

C.1 Hyperparameter Search Space

We perform 5-fold grid-search hyperparameter tuning to maximize the AUROC of blackbox models in predicting the groundtruth label. We also found that tuning for F1-score macro-averaged at a threshold of 0.5 led to same hyperparameter settings (i.e., calibrating the blackbox models at a threshold of 0.5 is justified).

[6] https://github.com/marcotcr/lime
[7] https://github.com/slundberg/shap

- Logistic Regression (LR): We implement this using scikit-learn [82], and use the linear solver with L2 regularization. Grid-search hyperparameter tuning is performed for regularization constants between 1e-5 to 1 (over 25 evenly-spaced values).
- Neural Network (NN): We implement a neural network with one hidden layer, and vary the number of units from 50, 100, and 200 with grid-search hyperparameter tuning. The relu activation function is used at each hidden layer, and the network is trained using Adam for up to 200 epochs with other default parameters in Pedregosa et al. [82].

C.2 Performance of Blackbox Models in Predicting Groundtruth

We ensure that all blackbox models have an average accuracy@chosen threshold of 0.5 greater than 0.68 and an average AUROC greater than 0.7, as observed in Table 4. For all datasets blackbox models are well-calibrated with Brier scores [15] ranging from 0.06-0.29 across the subgroups. We observe that several blackbox models are unfair: both in terms of varying performance across groups as well as their Demographic Parity Gaps.

Dataset	Blackbox	AUROC	Acc.	Δ_{AUROC}	$\Delta_{Acc.}$	Δ_{AUROC}^{group}	$\Delta_{Acc.}^{group}$	$\lvert DP \rvert$
adult	lr	90.1% ± 0.0%	85.3% ± 0.0%	2.7% ± 0.0%	3.8% ± 0.0%	6.1% ± 0.0%	11.1% ± 0.0%	17.5% ± 0.0%
	nn	90.8% ± 0.2%	85.8% ± 0.1%	2.4% ± 0.0%	3.8% ± 0.1%	5.3% ± 0.1%	11.3% ± 0.4%	16.1% ± 0.3%
lsac	lr	88.8% ± 0.0%	78.5% ± 0.2%	23.0% ± 0.1%	29.0% ± 0.5%	14.2% ± 0.2%	16.4% ± 0.3%	25.9% ± 0.2%
	nn	82.9% ± 1.3%	88.8% ± 1.1%	21.8% ± 1.6%	17.7% ± 3.0%	14.0% ± 3.5%	11.7% ± 2.0%	13.7% ± 0.8%
mimic	lr	80.6% ± 0.0%	89.4% ± 0.0%	2.0% ± 0.0%	0.9% ± 0.0%	3.6% ± 0.0%	1.6% ± 0.0%	1.4% ± 0.0%
	nn	78.4% ± 0.7%	88.1% ± 0.3%	1.1% ± 1.0%	1.3% ± 0.4%	2.0% ± 1.7%	2.3% ± 0.7%	2.0% ± 0.7%
recidivism	lr	72.5% ± 0.0%	68.5% ± 0.0%	2.1% ± 0.0%	0.9% ± 0.0%	3.5% ± 0.0%	1.5% ± 0.0%	22.1% ± 0.0%
	nn	71.8% ± 0.1%	67.5% ± 0.3%	1.8% ± 0.1%	1.5% ± 0.3%	3.0% ± 0.1%	2.4% ± 0.4%	21.8% ± 0.2%

Table 4. Blackbox Model Performance in Predicting Groundtruth

C.3 Prevalence of Positive Class

The datasets we study have varying degrees of class imbalance in groundtruth as well as in predictions of trained blackbox models (at a calibrated decision threshold of 0.5; see Table 5). We see that fidelity gaps are observed across these different prevalence rates, though class imbalance could have an impact on the sensitivity of the *metrics*. This underscores the importance of choosing the evaluation metric based on deployment context carefully. Controlled experiments where we oversampled the minority label class led to similar degrees of fidelity gaps for all datasets (see Table 11 in Appendix).

D FIDELITY GAPS EXIST IN LOCAL AND GLOBAL EXPLANATION MODELS

We observe that significant gaps in fidelity between data subgroups. The largest gaps are generally observed for the `lsac` dataset, and this could be because there are annotations available for more sensitive subgroups (5 race subgroups).

We also verify that fidelity gaps exist for the `lsac` dataset even without label-balanced blackbox models (as described Section B.2). These results are summarized in Table 8.

E GAPS EXIST IN BLACKBOX MODELS TRAINED ON BALANCED TRAIN SETS

We observe that even with training sets balanced by protected group labels – by randomly oversampling the minority group – significant, non-zero fidelity gaps still persist (see Table 9 and Table 10).

Dataset	Outcome Variable	Majority Class%	Blackbox	Predicted Majority Class%
adult [39]	Income > 50K	78	lr	81
			nn	80
lsac [114]	Student passes the bar	95	lr	75
			nn	88
mimic [47]	Patient dies in ICU	88	lr	96
			nn	92
recidivism [87]	Defendant re-offends	55	lr	65
			nn	62

Table 5. Binary classification datasets used in our experiments. The prevalence of majority class (or degree of imbalance) on the test set is detailed above.

Dataset	Blackbox	Expl	$Fidelity^{AUROC}$	Δ_{AUROC}	$\Delta_{Acc.}$	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$
adult	lr	LIME	99.9% ± 0.0%	0.0% ± 0.0%	0.8% ± 0.0%	0.1% ± 0.0%	2.4% ± 0.1%
	nn	LIME	95.7% ± 1.2%	1.1% ± 0.5%	6.9% ± 0.7%	3.0% ± 1.2%	20.6% ± 2.0%
lsac	lr	LIME	100.0% ± 0.0%	0.0% ± 0.0%	2.0% ± 1.0%	0.0% ± 0.0%	1.5% ± 0.5%
	nn	LIME	93.5% ± 0.5%	9.8% ± 3.2%	21.4% ± 4.4%	6.6% ± 1.2%	12.2% ± 2.2%
mimic	lr	LIME	84.2% ± 1.3%	1.1% ± 1.3%	0.4% ± 0.6%	3.0% ± 1.8%	1.1% ± 0.3%
	nn	LIME	92.6% ± 3.7%	0.9% ± 0.8%	0.8% ± 0.4%	1.7% ± 1.5%	1.4% ± 0.7%
recidivism	lr	LIME	100.0% ± 0.0%	0.0% ± 0.0%	0.1% ± 0.1%	0.0% ± 0.0%	0.3% ± 0.2%
	nn	LIME	99.0% ± 0.2%	0.3% ± 0.1%	0.9% ± 0.3%	0.7% ± 0.3%	2.4% ± 0.7%

Table 6. Gaps in Local Explanation Models; all models have fidelity greater than 85%

Similarly, we observe that significant fidelity gaps occur even on oversampling labels during training both blackbox and explanation models. We show the impact of label balancing on during training for recidivism and lsac on Δ^{g}_{Acc} for two global models below to illustrate this (Table 11). We note that class imbalance – and varying degrees of class imbalance for data subgroups – may be an important factor that affects the model training as well as metric computation. We leave the estimation of the effect of balancing degrees of class imbalance as well as balancing of data subgroups on explanation fairness for future work – our preliminary results indicates that fidelity gaps still might persist with this.

F GAPS IN FIDELITY WITH VARYING FEATURES

Similar to varying AUROC-fidelity gaps (Δ^{group}_{AUROC}), the accuracy fidelity gaps also: (1) depend on number of features used in the explanation model, (2) are often higher when fewer features are used, especially in local explanation models (see Fig.8). We also observe similar trends in the adult dataset (Fig.7).

G WORST-CASE FIDELITY

We observe that the groups with worst-case fidelity varies depending on dataset, type of explanation model, and type of blackbox model. Generally, these groups are the minoritized subgroups based on a protected or demographic attribute in the dataset (see Table 12).

Dataset	Blackbox	Expl	$Fidelity^{AUROC}$	Δ_{AUROC}	$\Delta_{Acc.}$	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$
adult	lr	GAM	100.0% ± 0.0%	0.0% ± 0.0%	0.1% ± 0.0%	0.0% ± 0.0%	0.3% ± 0.0%
	lr	Tree	96.1% ± 0.2%	1.3% ± 0.0%	1.5% ± 0.1%	2.9% ± 0.4%	4.5% ± 0.2%
	nn	GAM	98.8% ± 0.2%	0.2% ± 0.1%	0.8% ± 0.2%	0.5% ± 0.3%	2.4% ± 0.5%
	nn	Tree	95.9% ± 0.3%	0.8% ± 0.6%	1.1% ± 0.1%	0.6% ± 0.4%	3.4% ± 0.2%
lsac	lr	GAM	100.0% ± 0.0%	0.0% ± 0.0%	0.9% ± 0.9%	0.0% ± 0.0%	0.6% ± 0.4%
	lr	Tree	98.3% ± 0.1%	2.0% ± 0.4%	3.7% ± 3.1%	1.1% ± 0.4%	2.8% ± 0.7%
	nn	GAM	93.2% ± 0.6%	8.5% ± 1.8%	13.5% ± 0.9%	5.2% ± 1.2%	7.3% ± 1.0%
	nn	Tree	91.1% ± 0.8%	5.0% ± 1.9%	11.5% ± 2.7%	5.8% ± 2.1%	7.4% ± 1.2%
mimic	lr	GAM	99.5% ± 0.1%	0.2% ± 0.1%	0.5% ± 0.1%	0.4% ± 0.1%	0.9% ± 0.1%
	lr	Tree	84.6% ± 0.4%	3.7% ± 0.4%	0.6% ± 0.0%	8.1% ± 0.8%	1.2% ± 0.1%
	nn	GAM	90.6% ± 4.4%	0.4% ± 1.1%	1.2% ± 0.3%	1.8% ± 1.2%	2.2% ± 0.6%
	nn	Tree	73.4% ± 4.0%	1.4% ± 0.7%	1.1% ± 0.5%	3.0% ± 1.5%	2.0% ± 0.9%
recidivism	lr	GAM	99.9% ± 0.0%	0.0% ± 0.0%	0.1% ± 0.0%	0.1% ± 0.0%	0.3% ± 0.0%
	lr	Tree	98.6% ± 0.0%	0.4% ± 0.0%	0.0% ± 0.0%	0.4% ± 0.0%	0.1% ± 0.0%
	nn	GAM	99.3% ± 0.1%	0.2% ± 0.1%	0.2% ± 0.2%	0.4% ± 0.2%	0.6% ± 0.6%
	nn	Tree	98.9% ± 0.6%	0.4% ± 0.3%	0.9% ± 0.3%	1.0% ± 0.9%	2.3% ± 0.7%

Table 7. Gaps in Global Explanation Models; all models have fidelity greater than 72%

Blackbox	Expl	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$
lr	LIME	0.1% ± 0.0%	4.6% ± 0.0%
	GAM	0.1% ± 0.0%	2.6% ± 0.0%
	Tree	3.9% ± 0.0%	2.7% ± 0.0%
nn	LIME	2.2% ± 1.5%	2.9% ± 0.8%
	GAM	4.9% ± 5.4%	3.0% ± 0.8%
	Tree	21.7% ± 5.2%	3.9% ± 1.2%

Table 8. Gaps in Explanation Models: lsac dataset when the blackbox model is trained without label-balancing described in Section B.2.

H EFFECT OF VARYING SAMPLING VARIANCE FOR LIME

We observe that non-zero fidelity gaps in AUROC and accuracy persist across a range of sampling variances in LIME – i.e., when σ varies from 1e-3 to 20 as seen in Fig. 9. The behaviour is as expected with low fidelity for very low variance (sampling around very close to the test instance) and high variance (sampling from the whole distribution; not necessarily local). These make sense, as LIME employs a weighted ridge regression, where sample errors are weighed inversely in proportion to their distance from test instance. When sampling variance is high, a lot of sampled data can lie far away (as per standard LIME settings of "local" neighborhood point with variance 1). And since weighting is inversely proportional to distance in loss function, the prediction for the test instance is always correct (we observe that fidelity is 100% for all datapoints at $\sigma = 100$). When sigma is low, sample weights for all points are very high. If the blackbox model is stable to perturbations around points, then the fidelity should be high as well. However, if blackbox predictions fluctuate a lot with small perturbations to inputs, then fidelity gaps might exist (especially in Accuracy which is a threshold-based measure).

Dataset	Blackbox	Expl	$Fidelity^{AUROC}$	Δ_{AUROC}	$\Delta_{Acc.}$	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$
adult	lr	LIME	99.9% ± 0.0%	0.0% ± 0.0%	0.8% ± 0.1%	0.1% ± 0.0%	2.3% ± 0.2%
	lr	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	nn	LIME	92.1% ± 1.9%	1.6% ± 0.2%	6.4% ± 1.4%	4.6% ± 0.7%	19.1% ± 4.2%
	nn	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
lsac	lr	LIME	100.0% ± 0.0%	0.2% ± 0.2%	5.9% ± 1.3%	0.2% ± 0.1%	3.4% ± 0.8%
	lr	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	nn	LIME	88.9% ± 0.9%	11.5% ± 3.3%	12.6% ± 1.6%	8.6% ± 1.7%	6.7% ± 1.1%
	nn	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
mimic	lr	LIME	82.7% ± 0.6%	1.1% ± 1.1%	0.4% ± 0.6%	2.2% ± 2.2%	1.0% ± 1.0%
	lr	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	nn	LIME	93.5% ± 1.4%	0.8% ± 0.7%	1.2% ± 0.6%	1.6% ± 1.4%	2.2% ± 1.0%
	nn	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
recidivism	lr	LIME	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.1%	0.0% ± 0.0%	0.1% ± 0.2%
	lr	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	nn	LIME	98.2% ± 0.5%	0.5% ± 0.1%	1.7% ± 0.7%	1.0% ± 0.3%	4.6% ± 1.9%
	nn	SHAP	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%

Table 9. Gaps in Local Explanation Models when Blackbox Models are Trained on Balanced Training Sets

Dataset	Blackbox	Expl	$Fidelity^{AUROC}$	Δ_{AUROC}	$\Delta_{Acc.}$	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$
adult	lr	GAM	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.1% ± 0.1%
	lr	Tree	96.2% ± 0.3%	1.2% ± 0.1%	1.4% ± 0.0%	2.8% ± 1.0%	4.2% ± 0.1%
	nn	GAM	97.8% ± 0.7%	0.3% ± 0.2%	1.2% ± 0.3%	0.4% ± 0.4%	3.6% ± 0.9%
	nn	Tree	95.1% ± 0.7%	0.8% ± 0.4%	1.7% ± 0.6%	1.0% ± 1.2%	5.0% ± 1.7%
lsac	lr	GAM	100.0% ± 0.0%	0.1% ± 0.1%	1.4% ± 1.1%	0.0% ± 0.0%	0.9% ± 0.7%
	lr	Tree	98.2% ± 0.4%	4.3% ± 1.6%	7.3% ± 2.8%	2.7% ± 0.6%	3.9% ± 1.4%
	nn	GAM	89.5% ± 1.3%	12.7% ± 3.1%	14.4% ± 2.6%	9.4% ± 2.7%	8.5% ± 2.2%
	nn	Tree	85.2% ± 2.5%	15.4% ± 2.7%	17.4% ± 1.9%	10.8% ± 2.7%	9.1% ± 0.5%
mimic	lr	GAM	99.7% ± 0.1%	0.1% ± 0.0%	0.4% ± 0.2%	0.2% ± 0.1%	0.7% ± 0.3%
	lr	Tree	87.2% ± 1.0%	2.8% ± 0.3%	0.6% ± 0.1%	5.8% ± 0.7%	1.2% ± 0.1%
	nn	GAM	90.7% ± 1.9%	0.4% ± 0.8%	1.1% ± 0.4%	1.0% ± 1.2%	1.9% ± 0.7%
	nn	Tree	72.7% ± 3.5%	0.0% ± 1.7%	1.1% ± 0.8%	2.4% ± 2.3%	1.9% ± 1.5%
recidivism	lr	GAM	99.8% ± 0.2%	0.1% ± 0.1%	0.2% ± 0.1%	0.3% ± 0.3%	0.5% ± 0.2%
	lr	Tree	97.9% ± 0.5%	0.1% ± 0.2%	0.5% ± 0.5%	0.8% ± 0.9%	1.4% ± 1.1%
	nn	GAM	99.3% ± 0.1%	0.2% ± 0.1%	0.5% ± 0.6%	0.5% ± 0.2%	1.8% ± 1.1%
	nn	Tree	98.6% ± 0.6%	0.3% ± 0.3%	0.6% ± 0.1%	1.0% ± 0.8%	1.7% ± 0.4%

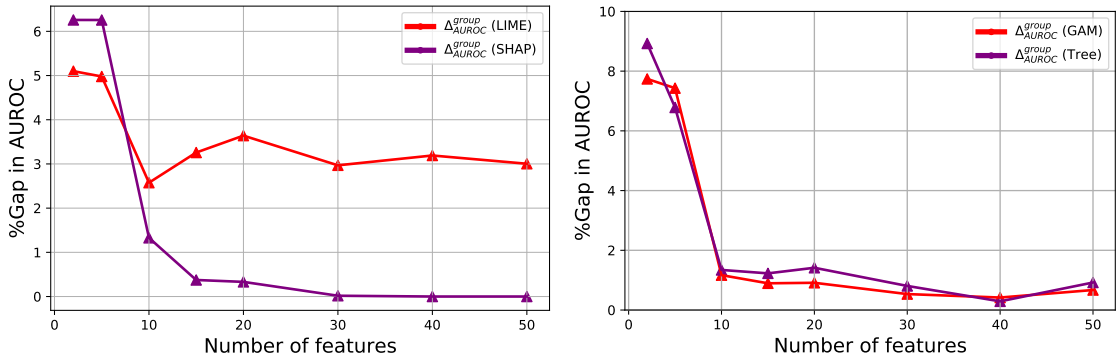
Table 10. Gaps in Global Explanation Models when Blackbox Models are Trained on Balanced Training Sets

I COMPARING FORMULATIONS OF EMPIRICAL RISK MINIMIZATION AND JUST TRAIN TWICE

Empirical risk minimization (ERM). Empirical risk minimization minimizes the loss averaged across training points [65]. Following formulation from Liu et al. [65], given a loss function $\ell(x, y; \theta) : \mathcal{X} \times \mathcal{Y} \times \Theta \to \mathbb{R}$ (*e.g.* cross-entropy loss),

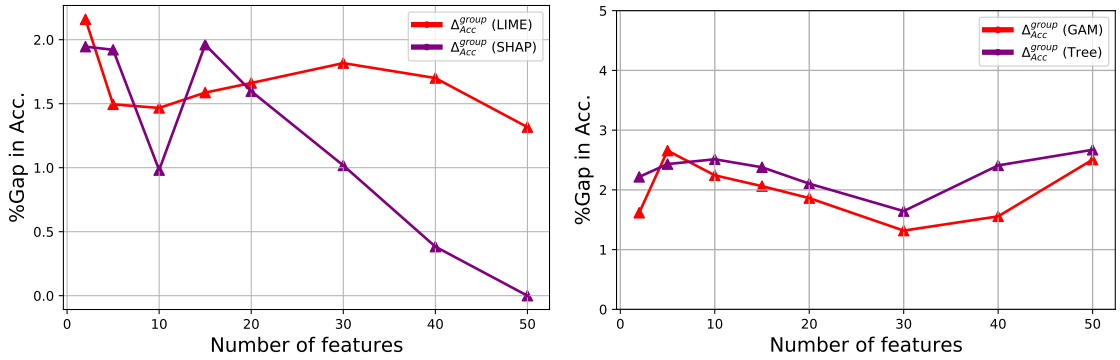
Dataset	Expl	$Fidelity^{Acc.}$	$\Delta^{\text{group}}_{AUROC}$	$\Delta^{\text{group}}_{Acc.}$
`lsac`	GAM	92.7% ± 0.7%	6.7% ± 1.5%	9.5% ± 1.4%
	Tree	90.5% ± 1.3%	5.8% ± 2.4%	9.6% ± 2.2%
`recidivism`	GAM	99.2% ± 0.4%	0.5% ± 0.2%	1.5% ± 0.9%
	Tree	98.9% ± 0.4%	0.6% ± 0.5%	1.1% ± 0.5%

Table 11. Fidelity gaps with oversampling class labels (data subgroups still imbalanced) for both blackbox and explanation models.



(a) `adult` + NN + AUROC-fidelity gap for local explanations (b) `adult` + NN + AUROC-fidelity gap for global explanations

Fig. 7. Similar to Fig. 2, we observe larger fidelity gaps across subgroups with sparser models, i.e., fewer features in local explanation model in AUROC fidelity gaps. The plots shown above are for the *adult* dataset.



(a) `mimic` + NN + Acc.-fidelity gap for local explanations (b) `mimic` + NN + Acc.-fidelity gap for global explanations

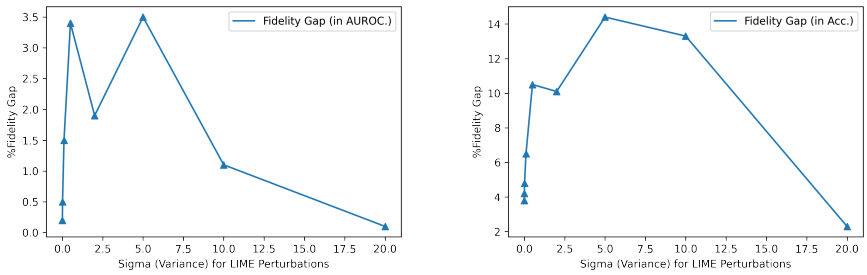
Fig. 8. We often observe larger gaps (or larger mean approximation error) with sparser models, i.e., fewer features in explanation model even in terms of accuracy-based fidelity.

ERM minimizes the following objective:

$$J_{\text{ERM}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i; \theta). \tag{5}$$

Dataset	Blackbox	Expl	Worst-Case Group
adult	LR	GAM	Sex: Majority
	LR	Tree	Sex: Majority
	NN	GAM	Sex: Majority
	NN	Tree	Sex: Minority
lsac	LR	GAM	Race: Majority
	LR	Tree	Race: Neither
	NN	GAM	Race: Neither
	NN	Tree	Race: Neither
mimic	LR	GAM	Sex: Minority
	LR	Tree	Sex: Minority
	NN	GAM	Sex: Minority
	NN	Tree	Sex: Minority
recidivism	LR	GAM	Race: Majority
	LR	Tree	Race: Minority
	NN	GAM	Race: Majority
	NN	Tree	Race: Majority

Table 12. The subgroup with lowest fidelity for all global explanation models. The Worst-Case Group is in the form of 'Protected Group: Majority/Minority/Neither Majority nor Minority'



(a) Perturbation variance σ vs Δ^{group}_{AUROC} (b) Perturbation variance σ vs $\Delta^{group}_{Acc.}$

Fig. 9. Fidelity gaps in AUROC and accuracy persist across a range of sampling variances

LIME fits a locally accurate model fit on sampled synthetic datapoints around a test instance [93] with weighted risk-minimization. However, since synthetic data is used in LIME, the group information of these synthetic instances is not known. Hence, the goal is to achieve good worst-group error at test time *without training group annotations*. We adopt a recent method proposed for rebalancing and training robust models without group information: Just Train Twice (JTT) [65]. While LIME is generally trained by minimizing the weighted mean-squared-error between blackbox model and local linear model predictions with Empirical Risk Minimization (ERM), JTT relies on robust training that upsamples the error-prone or difficult regions to predict during optimization.

Just Train Twice (JTT). JTT [65] is a two-stage approach that does not require group annotations at training time. First, it trains an identification model $\hat{f}_{\text{id}}$ via ERM and then identifies an error set $E = \{(x_i, y_i) \text{ s.t. } \hat{f}_{\text{id}}(x_i) \neq y_i\}$ of training examples that $\hat{f}_{\text{id}}$ misclassifies. Then, it trains a final model $\hat{f}_{\text{final}}$ by upsampling the points in the identified

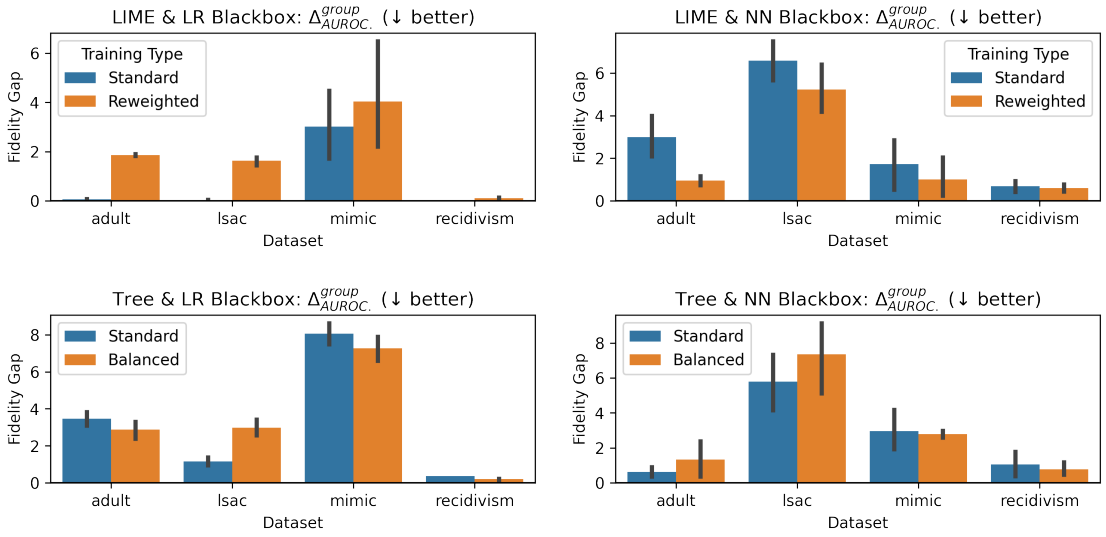


Fig. 10. AUROC Fidelity gaps across subgroups with and without robust training for LIME and Tree-based Models. We observe some improvements with balanced and robust training, but results depend on dataset and/or model type. Error bars indicate 95% confidence intervals around mean fidelity gap in each case.

error set [65].

$$J_{\text{up-ERM}}(\theta, E) = \left(\lambda_{\text{up}} \sum_{(x,y)\in E} \ell(x, y; \theta) + \sum_{(x,y)\notin E} \ell(x, y; \theta)\right), \tag{6}$$

To practically implement this we LIME, we multiply the local distance-based weights with λ_{up} for training samples in the error set. We also threshold the outputs of $\hat{f}_{\text{id}}$ and y_i to binary values for identifying the error set.

J IMPACT OF ROBUST TRAINING

In Fig. 10, we show the impact of robust and balanced training on all four datasets, two blackbox models, and two explanation (one local, one global) models.

K SIMULATION OF REAL-WORLD IMPACT

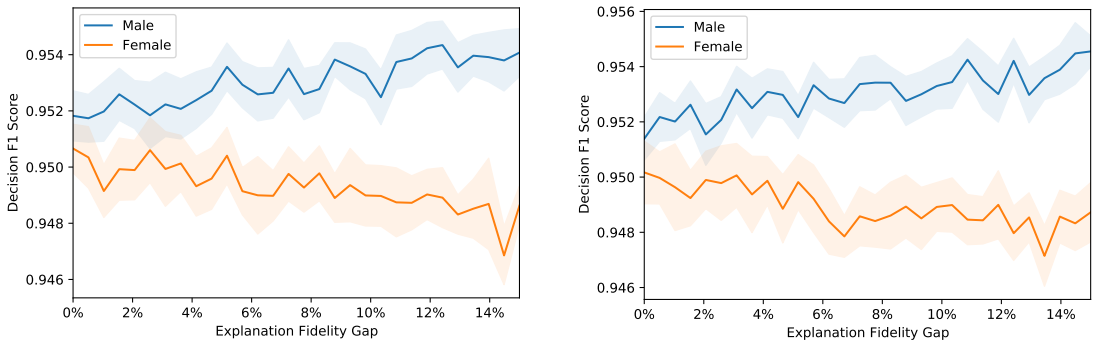
We assume the following simulation parameters for whether the user makes a correct decision based on the correctness of the blackbox model and the fidelity explanation.

- $P(\text{User} = \text{Correct} \mid \text{BlackBox} = \text{Incorrect}, \text{Explanation} = \text{Good}) = 0.6497$. This was selected based on results from Bansal et al. [10].
- $P(\text{User} = \text{Correct} \mid \text{BlackBox} = \text{Incorrect}, \text{Explanation} = \text{Poor}) = 0.68$, as it is reasonable that a poor explanation would more likely reveal the error in the blackbox model to the user, resulting in greater likelihood of a correct decision due to lower trust in the model [79].
- $P(\text{User} = \text{Correct} \mid \text{BlackBox} = \text{Correct}, \text{Explanation} = \text{Good}) = 0.9281$. This was selected based on results from Bansal et al. [10].

- $P(\text{User} = \text{Correct} \mid \text{BlackBox} = \text{Correct}, \text{Explanation} = \text{Poor}) = 0.90$, as it is reasonable that a poor explanation could lead the user astray into believing that the blackbox was incorrect, leading to an incorrect decision [10].

Parameters specified above were obtained from the user study conducted by Bansal et al. [10] for LIME explanations with the *beer* sentiment review dataset (aggregated across question-participant data points). Note that these are *subgroup-agnostic parameters*. While we use these parameters as reasonable estimates of real-world performance for our simulation, a detailed user-study is required to assess if these are accurate estimates and assumptions specifically for our case: this is a limitation of our simulation setup. Another assumption inherent in our simulation is that low fidelity explanations are poor. Thus our results are strongly influenced by the assumptions involved, but we believe still hold value as a proof-of-concept.

For the blackbox model, we use the logistic regression and neural network models described in Section B. For the explanation model, we assume that the average fidelity between males and females is 85%, and vary the maximum fidelity gap between the mean and two groups from 0% to 15%. We report results with final decision-making accuracy in the main text, and also obtained similar results on computing the F1-score instead (see Fig. 11a and 11b). The results shown here are all for blackbox models trained on a random 50% `lsac` data split without label balancing via oversampling, tested on the remaining 50%. Note that very similar results are obtained for models trained with balanced class labels via oversampling.



(a) Logistic Regression

(b) Neural Network

Fig. 11. F1-score of an admission officer's decision versus the fidelity gap of the explanation method between males and females using (a) logistic regression and (b) neural network as the blackbox model. We simulate the real-world impact of a deployed blackbox machine learning model along with an explanation method with varying levels of fidelity between males and females. We find that larger gap in explanation fidelity leads to a larger decision accuracy gap between the groups. Error bars are 95% confidence intervals derived from 20 runs of the simulation.

L FAIR BLACKBOX MODEL

We use an adversarial debiasing approach [118] to train fairer neural network blackbox models for two datasets: `lsac` and `mimic`. Optimal parameters are chosen based on lowest demographic fairness gap during validation. Note that we do not oversample datasets to deal with class imbalance, and rely on standard fair training approaches. Test performance is shown in Table 13, where we observe that models are fair.

Dataset	Blackbox	AUROC	Acc.	Δ_{AUROC}	$\Delta_{Acc.}$	Δ^{group}_{AUROC}	$\Delta^{group}_{Acc.}$	$\lvert DP \rvert$
lsac	nn	62.4% ± 3.2%	91.8% ± 1.4%	8.0% ± 1.3%	10.3% ± 2.1%	20.4% ± 1.0%	6.0% ± 0.7%	9.0% ± 1.3%
mimic	nn	81.3% ± 0.3%	89.2% ± 0.3%	1.0% ± 0.4%	0.9% ± 0.3%	1.8% ± 0.7%	1.6% ± 0.5%	0.6% ± 0.2%

Table 13. Fair Neural-Network Blackbox Model Performance in Predicting Groundtruth

M REMOVING FEATURES PREDICTIVE OF PROTECTED ATTRIBUTE

On removing these features, the results for `mimic` are shown in Table 15. The fidelity gaps using accuracy are not significantly higher than zero based on a one-sided Wilcoxon signed-rank test ($p > 0.05$) for the logistic regression blackbox model, and are all less then 1. These values are also reduced in comparison to the accuracy-based fidelity gaps on using all features. Blackbox and explanation models trained on similar versions of `lsac` dataset produce single class predictions due to high correlations between protected attributes and class label. We note that however that training models on these data representations lead to highly imbalanced blackbox model predictions (3% positive class prevalence for `mimic`). We observed that oversampling the minority class during blackbox model training to mitigate the effects of this imbalance led to similar results for the `mimic` dataset (i.e., significant reduction in accuracy-based fidelity metrics).

Blackbox	Expl	$Fidelity^{AUROC}$	$\Delta_{Acc.}$	$\Delta^{\text{group}}_{AUROC}$	$\Delta^{\text{group}}_{Acc.}$
LR	GAM	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	Tree	77.2% ± 0.0%	0.0% ± 0.0%	6.6% ± 0.0%	0.1% ± 0.0%
NN	GAM	97.0% ± 0.8%	0.2% ± 0.2%	1.9% ± 1.3%	0.5% ± 0.3%
	Tree	74.2% ± 7.4%	0.2% ± 0.3%	6.1% ± 3.2%	0.6% ± 0.2%

Table 14. Fidelity gaps on `mimic` dataset when all features predictive of protected attribute label are removed

For the `lsac` dataset, we oversample the minority class label (i.e., data subgroups still imbalanced) during blackbox model training with the reduced representation. This helps avoid the single class prediction problem. Similar to the results for `mimic`, we find that fidelity gaps are significantly reduced.

Blackbox	Expl	$Fidelity^{AUROC}$	$\Delta^{\text{group}}_{AUROC}$	$\Delta^{\text{group}}_{Acc.}$
NN	Tree	100.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%
	GAM	99.8% ± 0.1%	0.6% ± 0.5%	0.8% ± 0.6%

Table 15. Fidelity gaps on `lsac` dataset when all features predictive of protected attribute label are removed

However, the overall AUROC of the underlying NN blackbox classifier is low (0.62) with this representation. This indicates that class imbalance – and varying degrees of class imbalance for data subgroups – may be an important factor to consider. Our findings indicate that fidelity gaps persist across a range of class-imbalance ratios, but we leave the estimation of the effect of varying degrees of class imbalance (or positive-class prevalence) across subgroups with varying sample sizes on explanation fairness for future work.