Explanation Difference: Bridging Procedural and Distributional Fairness

Joe Germino[1], **Yuying Zhao**[2], **Tyler Derr**[2], **Nuno Moniz**[1], **Nitesh V. Chawla**[1]

[1]Lucy Family Institute for Data & Society, University of Notre Dame, Notre Dame, IN, USA 46556
[2]Vanderbilt University, Nashville, TN, USA 37235
jgermino@nd.edu, yuying.zhao@vanderbilt.edu, tyler.derr@vanderbilt.edu, nuno.moniz@nd.edu, nchawla@nd.edu

Abstract

Fairness in Machine Learning (Fair ML) is often presented as a trade-off between predictive performance and equality of predicted values. This view of fairness, commonly referred to as *distributional fairness*, fails to consider *how* a model arrives at its predictions. This may lead to Fair ML models evaluating protected groups on differing criteria, creating incentive structures that further perpetuate societal biases. Alternatively, *procedural fairness* attempts to ensure a fair decision-making process, but often does so at the expense of distributional fairness. In this paper, we propose a new procedural fairness measure, **Explanation Difference (EDiff)**, and further illustrate the importance of treating fairness as a multi-objective optimization problem considering distributional and procedural fairness, and predictive performance. We conduct an extensive experimental evaluation showing 1) the shortcomings of solely optimizing for distributional or procedural fairness, and that 2) our multi-objective approach utilizing **EDiff** can build fair ML models in both distributional and procedural fairness while retaining strong predictive performance.

Data and Code — https://github.com/joegermino/EDiff

Introduction

Since its inception, Fair Machine Learning (Fair ML) has been primarily focused on *distributional fairness*, evaluating the consistency of predicted outcomes between protected groups (Fragkathoulas et al. 2024; Decker, Wegner, and Leicht-Scholten 2025; Pfeiffer et al. 2023). Classic group fairness measures focus on the differences in probability of positive predictions between the privileged and unprivileged group (Dwork et al. 2012; Agarwal et al. 2018). Similarly, counterfactual fairness (Kusner et al. 2017) focuses on differences in the predicted value between two similar cases. This focus fails to consider an essential aspect of fairness: model explanations, also a critical dimension of humans' fairness judgment (Zhou, Chen, and Holzinger 2020). Striving for fairer models, we must consider *how* a model makes its prediction with model explanations (Dai et al. 2021), i.e., *procedural fairness* (Thibaut and Walker 1975).

We demonstrate this problem in the artificial scenario in Figure 1. In this example, there are two persons, one male and one female, applying for a loan. They are identical in all other relevant categories and the black-box model approves both of them. Using a counterfactual fairness approach, this outcome would be considered fair. However, the model explanations show that the female is being punished more harshly for her accumulated debt than the male. Similarly, the male is being rewarded much more for his income level than education while it is the opposite for the female.

While this model may be correcting for existing soci-

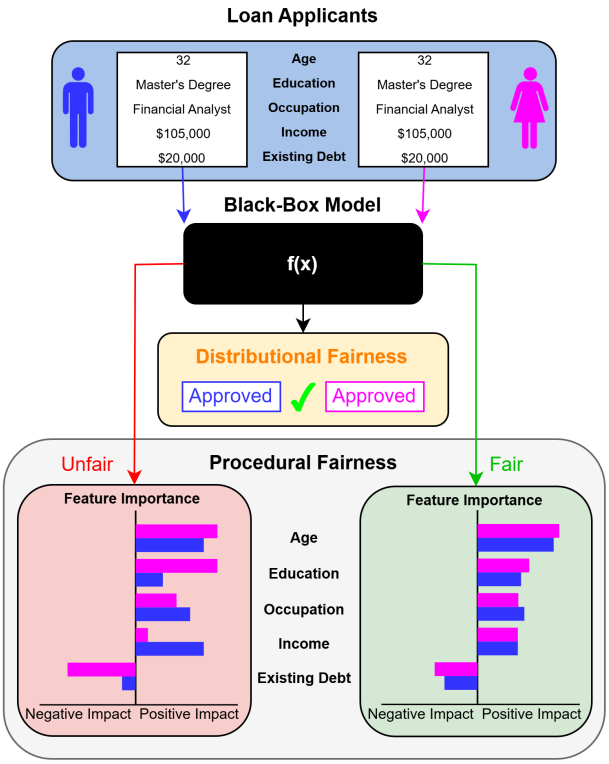

Figure 1: A synthetic example illustrating two loan applicants who are identical except for sex. While both applicants are approved for a loan (distributional fairness) there may be considerable disparity between the model's explanations for each suggesting the model is procedurally unfair.

etal biases in its decision-making and thus achieving distributional fairness, it is also creating incentives that reinforce these biases (Zhang, Khalili, and Liu 2020; Fu et al. 2022). Females evaluated by this model have an incentive to avoid debt while maximizing their education even at the expense of income, a difficult proposition. Instead, the goal of the model should be to achieve distributional fairness while simultaneously maintaining similar treatment for similar members of different protected groups. The goal of such a model would be to converge towards *equity* for members of different protected groups.

Furthermore, fair ML is frequently portrayed as a trade-off between accuracy and distributional fairness (Ma, Wang, and Liu 2022; Maity et al. 2020; Wick, Tristan et al. 2019; Wei and Niethammer 2022; Zliobaite 2015; Rueda et al. 2024). Correcting for inherent biases in a model typically degrades predictive performance but optimizing for one without regard for the other can limit its utility (Carvalho, Moniz, and Antunes 2023). Models that are fair but inaccurate do not offer more than random prediction. Meanwhile, accurate but unfair models may perpetuate stereotypes and harmful societal biases. Optimizing for only a single value leads to impractical models in high-risk domains.

We argue that this multi-objective view of fairness is still insufficient. Instead, we posit that fairness should be considered a three-pronged multi-objective problem weighing accuracy, prediction equality, and explanation equality. In our view, each of these three factors are necessary to achieving equity and it is improper to focus on one or two at the expense of the others. A model is equitable when it achieves fair outcomes through a fair decision-making process while retaining predictive performance.

Contributions. In this paper, we address the urgent need to expand our view of fairness to incorporate model explanations. Our main contributions are as follows:

- We consider fairness as a multi-objective problem focusing on both distributional and procedural fairness as well as predictive performance.
- We propose a new fairness measure **Explanation Difference (EDiff)** to measure the difference in model explanations between counterfactual samples. We further demonstrate how **EDiff** can be used efficiently in an optimization problem alongside traditional fairness and predictive performance measures.
- Our results demonstrate that considering **EDiff** can lead to a more equitable model, achieving the best trade-off across all possible user preferences (w.r.t. balancing distributional, procedural fairness, and utility) 59% of the time, nearly 3.5x better than the second-best method.

Related Work

Fairness is commonly measured by looking at the disparate treatment in protected groups (Pessach and Shmueli 2022). Popular fairness measures in classification include Statistical Parity (SP) (Dwork et al. 2012), which measures the difference in probability of positive prediction across a single protected attribute, and Equalized Odds (EO) (Agarwal et al. 2018), which measures the difference in True Positive Rate and False Positive Rate between groups. Similarly, in regression, group fairness measures such as Statistical Parity for regression (Agarwal, Dudik, and Wu 2019) and Equalized Odds (Hardt, Price, and Srebro 2016) are most popular.

While group fairness is useful for looking at average treatment over a larger sample, fairness can also be measured on an individual or sample-wise basis (Dwork et al. 2012; Mukherjee et al. 2020; Ilvento 2019; Petersen et al. 2021). Kusner et al. (2017) proposed the notion of counterfactual fairness, which uses the tools from causal inference to establish a prediction as fair if an individual's prediction remains the same with changing protected attributes. However, Fleisher (2021) has argued that individual or counterfactual fairness should not be used in isolation because they are insufficient for achieving group fairness.

Our proposal utilizes both group and counterfactual fairness. Prior work has demonstrated the viability of this approach. Anthis and Veitch (2023) explored the connection between the two using causal graphs. Germino, Moniz, and Chawla (2024) proposed a Mixture of Experts framework, which achieved group fairness by incorporating counterfactual fairness in the optimization process. Sharifi-Malvajerdi, Kearns, and Roth (2019) proposed using the average individual fairness over a larger sample, where the protected attribute did not necessarily define groups.

Model Explainability

Explainable AI (XAI) models are usually categorized into inherently interpretable and post-hoc explainable models (Arrieta et al. 2020). While inherently interpretable models are often preferred for high-stakes decision making (Rudin 2019; Carvalho, Pereira, and Cardoso 2019), black-box models are popular for their improved predictive capabilities. Post-hoc explanation models such as Permutation Feature Importance (Breiman 2001), SHAP (Lundberg and Lee 2017), and LIME (Ribeiro, Singh, and Guestrin 2016) are fit on an underlying black-box model, estimating individual features' importance at a local or global level.

A main drawback of these techniques is the computational complexity required to permutate each feature and calculate importances (Chuang et al. 2023a). SHEAR (Wang et al. 2022) accelerates SHAP by using only a subset of the input features in the calculation. Alternatively, CoRTX (Chuang et al. 2023b) uses contrastive learning to generate real-time explanations. In our proposal, we utilize FastSHAP (Jethani et al. 2021), capable of generating SHAP values through a single forward pass of a model trained through stochastic gradient descent without a ground truth value.

Explainable Fairness

While explainability and fairness are often studied independently, fair model explanations should be designed with specific protected groups in mind (Menon et al. 2024; Bisconti et al. 2024; Begley et al. 2020). Lünich and Keller (2024) shows that explanation simplicity directly affects fairness perceptions. Waller, Rodrigues, and Cocarascu (2024) design a technique to identify why a model is biased using

model explanations. Other approaches, e.g., the Comprehensive Fairness Algorithm, focus on ensuring fairness in the quality of explanations (Zhao, Wang, and Derr 2023).

Wang, Huang, and Yao (2024) proposed Group Procedural Fairness (GPF_{FAE}) as a measure of procedural fairness using feature attribute explanation (FAE) by measuring the distance in explanations for two similar members of different classes. Grgić-Hlača et al. (2018) used feature selection to detect and mitigate procedural fairness, relying on human feedback. Wang and Wu (2024) proposed a notion of equalized explainability, measuring the differences in model explanations. Significantly, their approach differs from ours in that they do not consider counterfactual cases or traditional fairness measures in the optimization process.

Multi-Objective Optimization

Multi-objective Optimization (MOO) refers to the process of finding the optimal solution value subject to multiple goals simultaneously (Gunantara 2018). MOO allows a model to optimize for a trade-off between multiple, often conflicting objectives. MOO has wide-ranging applications in fields such as finance (Tapia and Coello 2007), economics (Mardle, Pascoe, and Tamiz 2000), and mechanics (Deb and Datta 2012). These problems are typically complex and frequently rely upon evolutionary algorithms (Tian et al. 2021; Sharma and Kumar 2022).

MOO solutions are usually divided into two methods (De Weck 2004). Pareto fronts, originally proposed by Vilfredo Pareto, search for dominant solutions in which no other solution is superior to the dominant solution in every objective. Alternatively, scalarization incorporates multiple objectives into a single loss function, utilizing user-specified weights for each objective (Murata, Ishibuchi, and Tanaka 1996; Dodgson et al. 2009). Our proposal uses scalarization with equal weights assigned to each objective.

Explanation Difference

In this paper, we propose **Explanation Difference (EDiff)** to measure the unfairness of a model's decision-making process towards identical members of different protected groups. **EDiff** computes the absolute difference in feature importance between two counterfactual samples. **EDiff** can be measured using any model that provides local explanations; in this paper, we measure it using FastSHAP (Jethani et al. 2021) due to its demonstrated speed and faithful explanations w.r.t. state-of-the-art methods (Chuang et al. 2023a).

The steps to calculate **EDiff** are described in Algorithm 1. We create a counterfactual sample for each sample in the data that is identical in all features except the protected feature. Using a trained explainer model, we estimate the feature importance of all features for each sample and take the absolute difference per feature. Finally, we average such differences across all samples and sum all features to return a single score. The ideal value for **EDiff** is 0.

We demonstrate the calculation of **EDiff** in Figure 2 using a single sample, though **EDiff** would typically be the average value over the entire dataset. We assume two models using the synthetic example in Figure 1, where Model A has

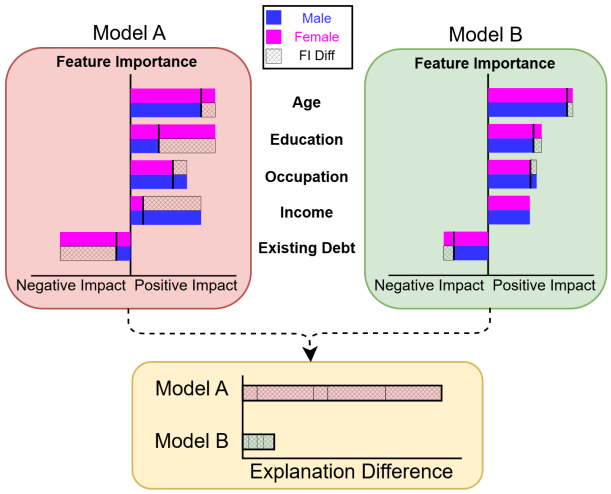

Figure 2: A synthetic example illustrating how **Explanation Difference (EDiff)** is calculated using a single counterfactual sample. The lightly shaded areas show the feature's important difference (FI Diff) between male and female. **EDiff** is the sum of these differences.

Algorithm 1 Explanation Difference

Require: ϕ: Explainer Model, X: Input Data, A: Protected Attribute

$EDiff = 0$
for X_i in X **do**
  $E_0 = \phi(X_i|A = 0)$
  $E_1 = \phi(X_i|A = 1)$
  $EDiff_i = |E_0 - E_1|$
  $EDiff$ += $EDiff_i$
end for
$EDiff = EDiff/|X|$
$EDiff = \Sigma(EDiff)$
return $EDiff$

the same prediction for both males and females, but with large differences between the importance of each feature; and Model B similarly maintains counterfactual fairness but with more equal feature importance. **EDiff** is calculated by summing the absolute difference between the importance of each feature. Therefore, Model A has a much higher **EDiff** than Model B, suggesting it is less fair.

By focusing on differences between counterfactual values, **EDiff** ensures that the difference in explanations is attributable to the value of the protected attribute. In optimizing for fairness, it is expected that there should be some difference in the treatment between protected groups to correct for societal biases, a principle underlying traditional measures such as Statistical Parity. However, we argue that this disparity should exist to correct for statistical differences between protected groups but not result in disparate treatment for similar members of different protected classes. We aim to find the minimal disparity necessary to achieve fair results.

Ethical Positioning

The procedural justice literature identifies several key dimensions contributing to fair-process perceptions. Rawls (1971) argues that justice requires fair outcomes and procedures that treat similar individuals similarly, i.e., models should evaluate individuals using consistent criteria regardless of protected attributes. Thibaut and Walker (1975) distinguished between two fundamental types of justice: distributive (outcome fairness) and procedural justice (process fairness). Their research demonstrated that people are willing to accept unfavorable outcomes when they perceive the process used to reach those outcomes as fair.

While fairness metrics tend to focus on distributional justice, we propose to combine this with procedural justice theories that emphasize fairness in decision-making processes. Formally, this would translate to scenarios where, given any two individuals x_1 and x_2 who differ only in a protected attribute a w.r.t. which resource allocation must be fair, a fair model should not only produce similar outcomes $f(x_1) \approx f(x_2)$, but these outcomes should come from similar decision processes, $|\Phi(x_1) - \Phi(x_2)| \approx 0$, where Φ represents the explanation function.

While **EDiff** offers a technical approach to a specific aspect of algorithmic fairness, we acknowledge the limitations of technical solutions to deeply rooted social inequalities: even well-designed fairness metrics may mask broader systemic issues if deployed without consideration of social context (Binns 2018). In addition, related literature (Green 2020; Benjamin 2019) has demonstrated that algorithmic fairness tools risk reinforcing existing power structures if deployed without consideration of broader social contexts.

We position **EDiff** not as a complete solution to discrimination, but as one component in a holistic approach to ethical AI that must include participatory design practices, appropriate governance structures, and ongoing human oversight. We aim to find the minimal disparity necessary to achieve fair results while avoiding disparate treatment for similar members of different protected classes.

Optimization

We propose an optimization procedure incorporating **EDiff** while optimizing for traditional fairness and predictive performance objectives. In our experiments, we implement this loss using a Multilayer Perceptron (MLP) (Rosenblatt 1958), though it can be used with any loss-based algorithm. We describe our proposal in Algorithm 2. To achieve our multi-objective optimization goal, we use a three-part loss function with components for F1 score ($\mathcal{L}_{F1}$), Statistical Parity ($\mathcal{L}_{SP}$), and Explanation Difference ($\mathcal{L}_{EDiff}$):

$$\mathcal{L}_{total} = \mathcal{L}_{F1} + \mathcal{L}_{SP} + \mathcal{L}_{EDiff} \tag{1}$$

F1 Loss This first term optimizes for predictive performance. A fundamental issue in fairness research is that a classifier can achieve fairness by predicting the same value for all samples, but this classifier would serve no utility. Also, due to the imbalanced nature of many fairness datasets, these classifiers can appear to have high accuracy despite having low utility. We improve model utility by optimizing for the F1 score to address this. AnyLoss proposed by Han, Moniz, and Chawla (2024) provides a framework to optimize any confusion matrix-based performance measure. AnyLoss first uses an amplifying factor L to push predicted values $\hat{y}$ towards extreme values of 1 or 0.

$$yh_i = \frac{1}{1 + e^{-L(\hat{y}_i - 0.5)}} \tag{2}$$

This amplified predicted value can then be used in a traditional F1 formula to create a differentiable loss function:

$$\mathcal{L}_{F1} = 1 - \frac{2(\sum_{i=1}^{n} y_i \cdot yh_i)}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} yh_i}, \tag{3}$$

where y_i is the ground truth label, and L is a hyperparameter chosen in validation.

SP Loss The second term optimizes for distributional fairness. By including a group fairness measure, the model should avoid situations where explanations may be similar but the final predictions are still unfair, i.e., when the model treats everybody equally regardless of protected attributes, but does not correct for societal biases captured in the data. We define a group fairness loss function based on Statistical Parity (Hardt, Price, and Srebro 2016):

$$\mathcal{L}_{SP} = |P(\hat{y} = 1|A = 0) - P(\hat{y} = 1|A = 1)|, \tag{4}$$

where $\hat{y}$ is the predicted value and A the protected attribute.

EDiff Loss The final term of the loss function optimizes directly for **EDiff** (Algorithm 1). Here, we calculate feature importance using FastSHAP for each sample and its counterfactual and sum the average absolute difference across features. The loss function is defined as:

$$\mathcal{L}_{EDiff} = \frac{\sum_{i=1}^{n} ||\phi(x_i|A = 0) - \phi(x_i|A = 1)||_1}{n}, \tag{5}$$

where $X \in \mathbb{R}^{n \times m}$ is the dataset, A is the protected attribute, and ϕ is the trained model explainer which returns a vector of length m indicating the importance of each feature towards the predicted value.

Algorithm 2 Multi-Objective Optimization

Require: X: Train Data, A: Protected Attribute, y: Ground Truth, n: Number Epochs, ϕ: FastSHAP Explainer

for epoch in n **do**
  $s =$ Random sample of features in X
  $\mathcal{L}_{FS} = (f(s) - f(0) - s^T\hat{\phi})^2$
  Update ϕ using $\mathcal{L}_{FS}$
  $\hat{y} = f(X)$
  $\mathcal{L}_{F1} =$ AnyLoss$(\hat{y}, y)$
  $\mathcal{L}_{SP} = |P(\hat{y} = 1|A = 0) - P(\hat{y} = 1|A = 1)|$
  $\mathcal{L}_{EDiff} = \frac{\sum_{i=1}^{|X|} ||\phi(x_i|A=0) - \phi(x_i|A=1)||_1}{|X|}$
  $\mathcal{L}_{total} = \mathcal{L}_{F1} + \mathcal{L}_{SP} + \mathcal{L}_{EDiff}$
  Update learner f using $\mathcal{L}_{total}$
end for

Name	Prediction Task	Cases	Features	Imbalance Ratio	Privileged Groups
Adult (Dua and Graff 2017)	Annual income exceeds $50,000	45222	94	0.20	Male, White
German Credit (Dua and Graff 2017)	Bank Account is high credit risk	1000	47	2.33	Male
Dutch Census (Center 2013)	Person's occupation is prestigious	60420	50	1.10	Male
Bank Marketing (Moro, Cortez, and Rita 2014)	Client subscribes with deposit	45211	42	0.13	Married
Credit Card Clients (Yeh and Lien 2009)	Client will default in next month	30000	82	0.28	Male, Single
OULAD (Kuzilek, Hlosta, and Zdrahal 2017)	Student will pass class	21562	40	2.12	Male
Lawschool (Wightman 1998)	Student will pass bar on first attempt	20798	18	8.07	Male, White
Diabetes (Strack et al. 2014)	Patient readmits high credit risk	173	272	0.25	Male
COMPAS	Likelihood of Recidivism	6172	18	0.94	Caucasian

Table 1: Datasets used in the experimental evaluation.

FastSHAP In order to optimize for **EDiff**, faithful explanations are required at each epoch of training. Standard explanation methods, such as SHAP, would require training a new explanation model at each step, which is prohibitively expensive. We propose incorporating FastSHAP (Jethani et al. 2021) into the training process to maintain efficiency. During each epoch, we calculate the loss for the FastSHAP explainer model and the loss for our multi-objective MLP. By doing so, we can simultaneously train the explainer and the MLP, eliminating the need to fit a new model at each epoch and greatly improving efficiency. Our method is not dependent upon FastSHAP, and other efficient model explanation techniques can be used.

Curriculum Learning Even with FastSHAP, calculating the feature importance for all samples at every iteration is computationally expensive. We address this through curriculum learning (Bengio et al. 2009). In our proposal, we calculate $\mathcal{L}_{EDiff}$ and $\mathcal{L}_{FS}$ on a small subset of the samples, the size of which increases as the model trains: we divide the training process into fourths with sample sizes of 5%, 10%, 20%, and 50%, respectively. The intuition of using curriculum learning in this context is that in the early stages of training, the model can detect large patterns within the data with small samples. As it becomes more complex, slight differences between samples become more consequential, and therefore, more data is needed to optimize properly.

Experimental Evaluation

Our experimental evaluation aims to answer the following research questions:

RQ1 Does optimizing for **EDiff** create a more equitable model in terms of fair outcomes and fair explanations than existing fairness algorithms?

RQ2 How does our optimization proposal compare to the state-of-the-art when treating fairness as a multi-objective task?

RQ3 Does removing components of $\mathcal{L}_{All}$ affect its ability to balance multiple objectives?

RQ4 Does incorporating FastSHAP during the training process allow for efficient training comparable with other fairness approaches?

Parameter	Search Space
L (AnyLoss Parameter)	[5, 25, 75]
Hidden Size	[10, 50, 150]
Learning Rate	[.001, .0001]
Batch Size	[128, 512, 1024]
Number Epochs	[50, 100, 250]

Table 2: Hyperparameters searched during cross-validation.

Data

We use nine fairness-oriented and publicly available datasets. We followed the pre-processing steps, protected class definitions, and privileged groups for all datasets based on prior literature (Le Quy et al. 2022). Pre-processing included removing samples that had missing data and dropping non-predictive columns. When necessary, the target variable was converted to a binary value, and categorical variables were one-hot encoded. We restricted our analysis to binary protected attributes and fit models using one protected attribute at a time. Privileged groups were defined using the definitions from Le Quy et al. (2022). When undefined, the majority class was designated as the privileged group. Details regarding the datasets can be found in Table 1.

Algorithms

We compare our proposal against an MLP optimized using AnyLoss (MLP_{F1}) and five fairness-aware algorithms. The first solution proposed by Hardt, Price, and Srebro (2016) is a post-processing algorithm that optimizes a model for Equalized Odds. The next proposed by Agarwal et al. (2018) is an in-processing algorithm that reduces a fairness classification task to a series of cost-sensitive classification problems, where the outcome is a randomized classifier optimized for the most accurate classifier subject to fairness constraints. The third, proposed by Feldman et al. (2015), is a pre-processing method to remove disparate impact from data before a model is trained. We refer to these solutions by the authors' names. The remaining two algorithms are xFAIR (Peng, Chakraborty, and Menzies 2022), which aims to mitigate bias by relabeling protected attributes in test data, and Adversarial Debiaser (Zhang, Lemoine, and Mitchell 2018) (Adv. Deb.), which uses adversarial learning to address fairness concerns. These fairness-aware algorithms focus on distributional fairness, not procedural fairness.

	AUC	F1	Accuracy	Precision	Recall	Statistical Parity	Equalized Odds	GPF_{FAE}	Explanation Difference
MLP_{F1}	2.67 ± 0.15	$\mathbf{2.06 \pm 0.11}$	$\mathit{3.34 \pm 0.18}$	$\mathit{3.42 \pm 0.15}$	4.09 ± 0.15	6.59 ± 0.17	6.09 ± 0.18	4.72 ± 0.21	6.05 ± 0.14
Agarwal	6.65 ± 0.12	5.09 ± 0.15	5.15 ± 0.15	5.47 ± 0.15	4.41 ± 0.21	$\mathit{3.91 \pm 0.16}$	$\mathit{3.61 \pm 0.15}$	4.57 ± 0.22	6.96 ± 0.17
Hardt	6.83 ± 0.13	7.75 ± 0.14	7.39 ± 0.15	6.63 ± 0.26	6.36 ± 0.29	$\mathbf{3.67 \pm 0.21}$	$\mathbf{3.41 \pm 0.20}$	5.38 ± 0.19	5.10 ± 0.29
Feldman	$\mathit{2.35 \pm 0.16}$	$\mathit{2.29 \pm 0.14}$	3.52 ± 0.18	3.82 ± 0.15	3.91 ± 0.14	6.50 ± 0.19	6.63 ± 0.17	4.74 ± 0.20	5.12 ± 0.16
Adv. Deb.	6.88 ± 0.17	6.08 ± 0.17	5.51 ± 0.25	5.20 ± 0.24	5.69 ± 0.22	6.43 ± 0.20	6.69 ± 0.24	6.77 ± 0.25	7.45 ± 0.22
xFAIR	$\mathbf{2.27 \pm 0.12}$	5.60 ± 0.20	$\mathbf{2.56 \pm 0.21}$	$\mathbf{2.18 \pm 0.19}$	7.40 ± 0.09	6.13 ± 0.19	6.41 ± 0.19	4.47 ± 0.20	5.47 ± 0.17
MLP_{EDiff}	7.92 ± 0.15	7.86 ± 0.16	7.19 ± 0.23	7.95 ± 0.15	5.77 ± 0.31	$\mathbf{3.67 \pm 0.26}$	3.76 ± 0.26	6.53 ± 0.16	$\mathbf{2.49 \pm 0.15}$
MLP_{All}	4.67 ± 0.12	4.15 ± 0.09	5.20 ± 0.12	5.18 ± 0.11	$\mathbf{3.67 \pm 0.13}$	3.92 ± 0.12	4.15 ± 0.14	$\mathbf{4.02 \pm 0.18}$	3.21 ± 0.15
$MLP_{Curriculum}$	4.76 ± 0.12	4.11 ± 0.09	5.16 ± 0.11	5.15 ± 0.11	$\mathit{3.69 \pm 0.13}$	$\mathit{3.91 \pm 0.12}$	4.16 ± 0.14	$\mathit{4.15 \pm 0.17}$	$\mathit{3.15 \pm 0.13}$

Table 3: Average and Standard Error of performance rankings for all data sets. Algorithms are grouped by fairness-agnostic, fairness-aware, and our proposal. Lower numbers indicate better performance. **Best** and *second-best* results marked.

Hardt, Agarwal, and AdvDeb algorithms are implemented using the Fairlearn Python package (Bird et al. 2020). For xFAIR, we used a Decision Tree as the extrapolation model and a Random Forest as the classification model, as suggested in the original paper (Peng, Chakraborty, and Menzies 2022). We used an MLP trained with AnyLoss to consistently compare Hardt, Agarwal, and Feldman. All experiments were run using a single protected attribute at a time.

We evaluate three versions of our method. The first, MLP_{EDiff}, is an MLP trained using only $\mathcal{L}_{EDiff}$ as its loss function. This version demonstrates a model's ability to optimize for **EDiff** without regard for utility or traditional fairness. The next version, MLP_{All}, uses the full 3-part loss function defined above. This solution demonstrates our optimization procedure in a multi-objective setting. The final solution, $MLP_{Curriculum}$, is our full proposal optimizing for the 3-part loss function while using curriculum learning.

Results

For each dataset, we split the data into train and test sets using a 75/25 split. Models were fit on the train set using 3-fold cross-validation. The hyperparameters used in our search are listed in Table 2. Each of the algorithms require a different subset of these parameters. We compared the results in predictive performance (Accuracy, ROC-AUC, Precision, Recall, F1), group fairness (SP, EO), and procedural fairness (GPF_{FAE}, **EDiff**). The solutions were ranked by their performance in each measure. We ran the experiment ten times using different train and test splits. The results are in Table 3.

As the results demonstrate, using **EDiff** as a loss function is effective in optimizing for **EDiff**, as our proposals are top three in **EDiff**, but optimizing strictly for **EDiff** without considering predictive performance results in poor utility, the worst performer across all baselines. Meanwhile, standard fairness-aware algorithms such as Agarwal and Adversarial Debiasing are effective in optimizing for group fairness but struggle with procedural fairness, trailing even the fairness-agnostic MLP. These findings support our hypothesis that equal predictions might not be indicative of an equitable decision-making process. While it is possible to achieve strong performance in both fairness approaches, they require separate optimization processes or the model may learn to neglect one in favor of the other. With regards to **RQ1**, we find that optimizing for **EDiff** is able to create a more equitable model than the existing approaches.

Multi-Objective Analysis

Our proposed approach using a three-part loss function is most effective when looking at each objective collectively. Our proposal ranks third in F1 and SP and second in **EDiff**. All other fairness-aware algorithms perform poorly in at least one of these categories. This multi-objective performance can be seen in Figure 3, with plots showing average performance across all utility and fairness measures for each fairness-aware algorithm. Our proposal consistently outperforms other measures in **EDiff**, maintaining competitive performance with the leading algorithm in each measure.

In real-world decision-making, the choice of model often depends on the relative importance of different objectives. For example, in some settings, predictive performance may take precedence over fairness, while in others, procedural and/or distributional fairness may be prioritized (and at varying levels). To systematically explore how robust a method is to such preference variability, we simulate the process of model selection under all convex combinations of the objectives – that is, all possible non-negative weightings of F1 (utility performance), SP (statistical parity), and **EDiff** (explanation difference) that sum to one. Formally, for a given weight vector $\mathbf{w} = (w_{\text{F1}}, w_{\text{SP}}, w_{\text{EDiff}})$ we compute a weighted score:

$$w_{\text{avg}} = w_{\text{F1}} \cdot \text{F1} + w_{\text{SP}} \cdot \text{SP} + w_{\text{EDiff}} \cdot \text{EDiff}$$
$$\text{s.t. } w_{F1} + w_{SP} + w_{EDiff} = 1 \text{ and } w_{F1}, w_{SP}, w_{EDiff} \geq 0. \quad (6)$$

This defines the 2D simplex of weights–a triangle in 3D weight space where each point corresponds to a specific trade-off among the three objectives. For example, a corner such as (0,0,1) reflects exclusive emphasis on **EDiff**, while a point like (0.33, 0.33, 0.34) represents a near-equal weight across all objectives. By uniformly sampling this triangle, we simulate all reasonable user preferences.

At each sampled weight combination, we determine which fixed model (among our proposed and baseline methods) achieves the highest weighted performance. This allows us to answer the central question: *"Across all user-defined trade-offs, how often is each method the best choice?"*

We visualize this analysis by plotting the 2D simplex of weight combinations, where each point within the triangle corresponds to a specific trade-off among the three objectives (F1, SP, **EDiff**). Each point is colored according to the best method under that weighting. The resulting plot provides an interpretable partitioning of the preference

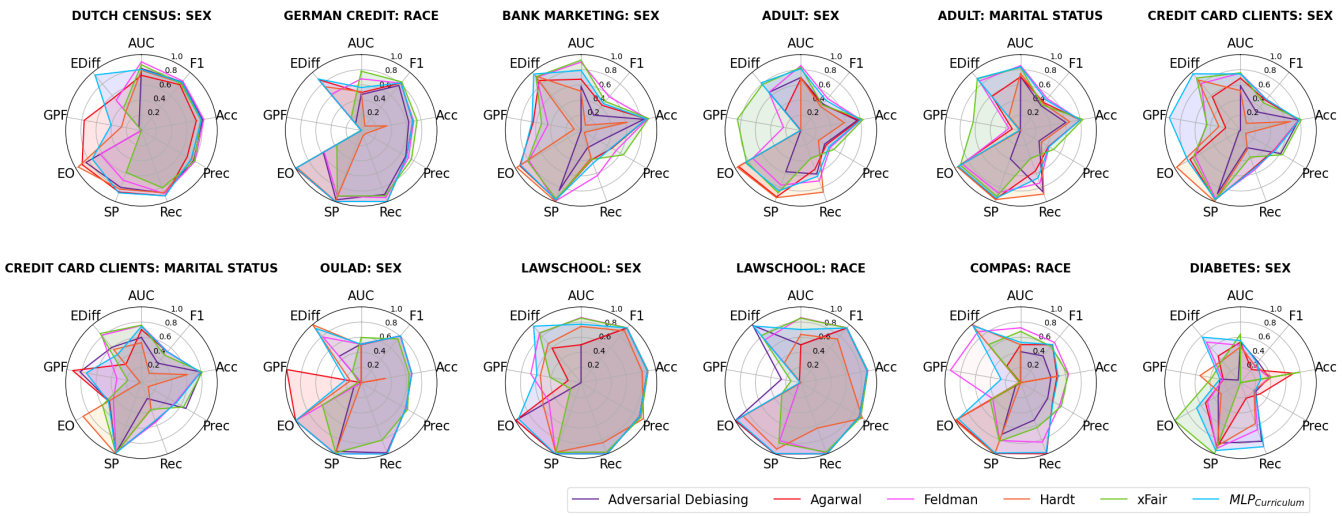

Figure 3: Radar plots showing average performance of fairness-aware algorithms in all measures across ten experiments. Measures were scaled from 0 to 1, so larger values were always better.

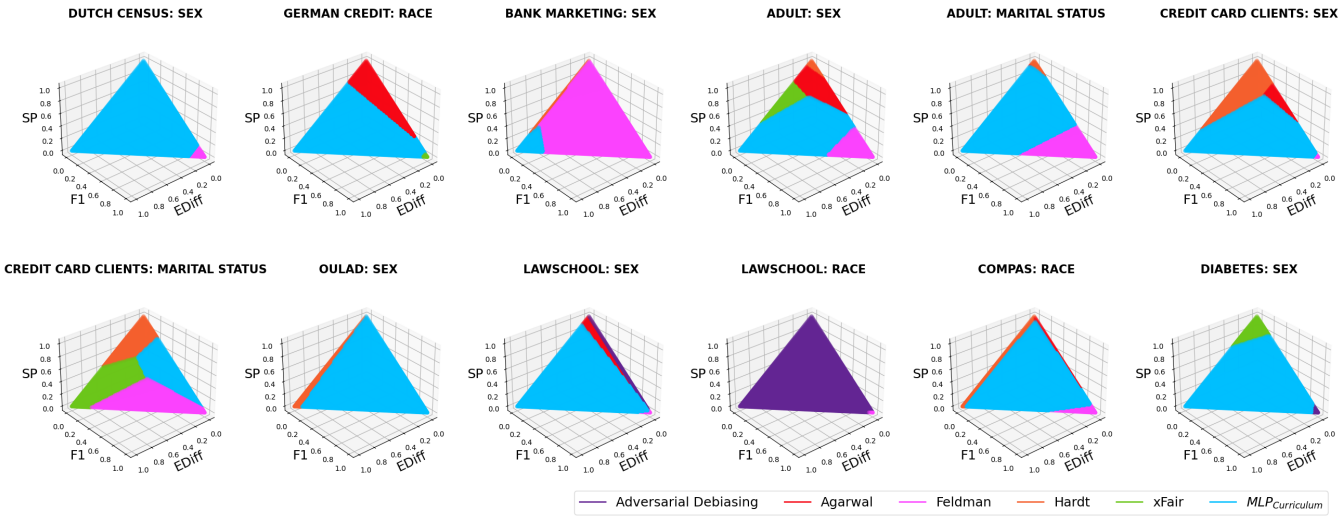

Figure 4: Best performing model at each triplet of weights assigned to **EDiff**, SP, and F1. For example, the point in the bottom left corner of the triangle shows the best model when only considering performance in **EDiff**, and the top corner is the best model when only considering SP. Results were split by dataset, and the average was over ten trials.

space, showing which regions of user preferences are best served by each method.

Figure 4 illustrates this for all datasets, where our proposed method occupies the largest triangle share in 7 out of 12 datasets. Notably, the center of the triangle (equal weighting) is also best served by our method in most datasets, highlighting its balanced performance.

Table 4 quantifies this analysis. On average, our proposed method accounts for over 50% of the total simplex coverage, nearly 3.5x more often than the second-best method. These results strongly suggest that our approach is the most versatile across a wide range of user preferences, including cases where predictive accuracy or statistical parity dominate and EDiff is the primary concern. Concerning our above question, we find that our model competes with and often exceeds state-of-the-art methods in this multi-objective setting.

Ablation Study

Our proposed optimization approach uses a three-part loss function to balance for each of our objectives. We conduct an ablation study to demonstrate how each component is necessary to balance the objectives properly. Following the same experimental design from above, we compared the seven different combinations of our loss function (F1, SP, EDiff, F1+SP, F1+EDiff, SP+EDiff, F1+SP+EDiff) and the curriculum learning approach with an MLP over ten different trials.

Algorithm	Percentage
$\text{MLP}_{Curriculum}$	$.59 \pm .03$
Feldman	$.17 \pm .02$
Agarwal	$.07 \pm .01$
Adversarial Debiasing	$.07 \pm .02$
Hardt	$.06 \pm .01$
xFair	$.04 \pm .01$

Table 4: Average proportion of the convex weight simplex over F1, SP, and EDiff representing the multi-objective user preference trade-offs, indicating how often each method is optimal across all datasets. Our proposed solution is highlighted in blue.

Similar to above, Table 5 shows the average rank performance comparing the MLP solutions. Notably MLP_{SP}, MLP_{EDiff}, and $\text{MLP}_{SP+EDiff}$ are the worst performers in AUC, F1, Accuracy, and Precision while MLP_{F1} is the best in each of these demonstrating the trade-off between fairness and predictive performance. While MLP_{SP} is second in **EDiff**, it achieves this while offering no predictive utility. Meanwhile, optimizing for F1 and SP provides strong predictive performance and traditional group fairness but is the second worst in Explanation Difference. Meanwhile, $\text{MLP}_{Curriculum}$ performs similarly to MLP_{All} despite the condensed training procedure. While neither $MLP_{Curriculum}$ nor MLP_{All} is best when looking at individual measures, they are also never the worst. Their performance is consistently in the middle, demonstrating their capability to balance each goal.

Figure 5 details the performance of MLP_{F1}, MLP_{SP}, MLP_{EDiff}, and $\text{MLP}_{Curriculum}$ across the 9 datasets in AUC, SP, and EDiff. MLP_{F1} considerably outperforms the others in F1 while trailing in both fairness measures. On the other hand, MLP_{SP} and MLP_{EDiff} generally have similar performance to each other. $\text{MLP}_{Curriculum}$ typically stands alone with higher F1, SP, and EDiff than MLP_{SP} and MLP_{EDiff} but lower than MLP_{F1}. $\text{MLP}_{Curriculum}$ is properly balancing the objectives instead of optimizing for one or two at the expense of the others. With regards to **RQ3** we find that each of the components of $\mathcal{L}_{All}$ is necessary to balance the three separate objectives properly.

Efficiency

Finally, we look to see how including FastSHAP in the optimization procedure affects the overall efficiency of our approach. Instead of fitting a new explainer model at each epoch to calculate the EDiff Loss, we use a FastSHAP model, which is being updated each epoch along with the MLP. The logic behind this is to minimize the considerable cost associated with fitting a new explainer model each epoch. Figure 6 illustrates how FastSHAP loss changes over time compared to the other three components of the proposed loss function.

In 5 of the 9 datasets, the FastSHAP loss improves from its starting point. The FastSHAP loss stabilizes within 100 epochs in all datasets while the **EDiff** loss improves considerably. These findings illustrate that our proposal, which optimizes an explainer model while also using the model's predictions, is meritorious. Importantly, in our earlier results, **EDiff** was measured by a separate FastSHAP model fit on the final MLP. Even when the FastSHAP loss did not improve internally, our proposal still improved **EDiff** using an external explainer model.

Table 6 shows how this approach affects the processing time required to train an MLP. Our proposal scales with the number of samples more than with the number of features, despite requiring feature explanations. The most efficient algorithm is Adv. Deb. However, our results indicate that its performance is inconsistent. Our proposal uses more hyperparameters than the fairness baselines, leading to more total training time. Despite this, MLP_{All} is still faster than Agarwal in every dataset in total time and on a per-model basis, where it is 89%

Significantly, incorporating curriculum learning into the training procedure dramatically increases the efficiency without a significant change in performance. As Table 5 shows, $\text{MLP}_{Curriculum}$ and MLP_{All} have similar performance across all fairness and performance measures. However, $\text{MLP}_{Curriculum}$ has an average processing time that is 66% quicker than MLP_{All}. Our results demonstrate that even a simple implementation of curriculum learning can optimize for procedural fairness, distributional fairness, and predictive performance efficiently. Overall, regarding **RQ4**, we find that incorporating FastSHAP during the training process allows for comparative training times with SOTA approaches while effectively improving the model's procedural fairness.

Discussion

Despite demonstrating **EDiff**'s usefulness in optimization, the question remains regarding its real-world significance as a fairness measure. Figure 7 shows examples of good and bad **EDiff** from the experimental evaluation. We used the Lawschool dataset with sex as the protected attribute, and we picked the three features with the most considerable explanation disparity for both the best and worst performing models (six total features). The plots show the feature importance for all these features across both models as measured by a separate FastSHAP model. These are from a single trial of the experiment and are not meant to represent model performance but rather illustrate the practical implications of strong or weak performance in **EDiff**.

Our proposal MLP_{All}, with an **EDiff** of 0.03, is the best performing model while Adv. Deb. with an **EDiff** of 0.43 is the worst. Also, our proposal has an SP of .0027 compared to .0003 for Adv. Deb., indicating both are fair by distributional fairness. In this case, Adv. Deb. has a larger gap in importance, even in some of the features, in which our proposal performs the worst. This is especially true in the protected attribute, Sex, where it is evident that males are being punished to make up for overperformance in other attributes.

EDiff is measured using counterfactuals, so each of these differences indicates an unfairness in prediction beyond the features known to the model. Using a model with unfair explanations could be detrimental to the long-term goals of fairness. Using the examples above, a law school considering a student's likelihood of success on the bar exam will ad-

	AUC	F1	Accuracy	Precision	Recall	Statistical Parity	Equalized Odds	GPF$_{FAE}$	Explanation Difference
MLP$_{F1}$	$\mathbf{1.87 \pm 0.15}$	$\mathbf{2.07 \pm 0.13}$	$\mathbf{2.71 \pm 0.16}$	$\mathbf{2.33 \pm 0.14}$	4.52 ± 0.16	6.67 ± 0.14	6.36 ± 0.15	4.21 ± 0.19	7.28 ± 0.12
MLP$_{SP}$	6.23 ± 0.16	6.11 ± 0.12	5.72 ± 0.17	6.15 ± 0.13	$\mathit{4.05 \pm 0.22}$	$\mathbf{2.64 \pm 0.12}$	$\mathbf{2.63 \pm 0.12}$	5.18 ± 0.16	$\mathit{2.23 \pm 0.15}$
MLP$_{EDiff}$	6.81 ± 0.14	6.80 ± 0.09	5.96 ± 0.18	6.67 ± 0.12	5.48 ± 0.19	4.22 ± 0.21	4.30 ± 0.22	5.75 ± 0.14	3.82 ± 0.15
MLP$_{F1+SP}$	$\mathit{2.95 \pm 0.15}$	$\mathit{2.53 \pm 0.13}$	$\mathit{3.04 \pm 0.13}$	$\mathit{2.82 \pm 0.13}$	$\mathbf{3.95 \pm 0.14}$	5.33 ± 0.14	5.09 ± 0.15	4.17 ± 0.18	6.11 ± 0.17
MLP$_{F1+EDiff}$	3.41 ± 0.16	3.50 ± 0.11	4.08 ± 0.17	3.74 ± 0.13	4.17 ± 0.16	5.65 ± 0.17	5.29 ± 0.18	4.66 ± 0.18	5.32 ± 0.15
MLP$_{SP+EDiff}$	6.45 ± 0.14	6.67 ± 0.11	5.73 ± 0.18	6.38 ± 0.14	5.20 ± 0.22	$\mathit{2.90 \pm 0.14}$	$\mathit{2.97 \pm 0.15}$	5.05 ± 0.18	$\mathbf{2.08 \pm 0.14}$
MLP$_{All}$	4.10 ± 0.10	4.17 ± 0.08	4.40 ± 0.13	3.99 ± 0.10	4.30 ± 0.17	4.30 ± 0.10	4.67 ± 0.12	$\mathbf{3.47 \pm 0.15}$	4.62 ± 0.13
MLP$_{Curriculum}$	4.19 ± 0.10	4.15 ± 0.07	4.35 ± 0.13	3.93 ± 0.10	4.32 ± 0.17	4.29 ± 0.10	4.70 ± 0.12	$\mathit{3.51 \pm 0.15}$	4.54 ± 0.12

Table 5: Ablation study comparing average and standard error of performance rankings for all data sets. All solutions are MLP using different combinations of the loss function components. Lower numbers indicate better performance. **Best** and *second-best* results marked.

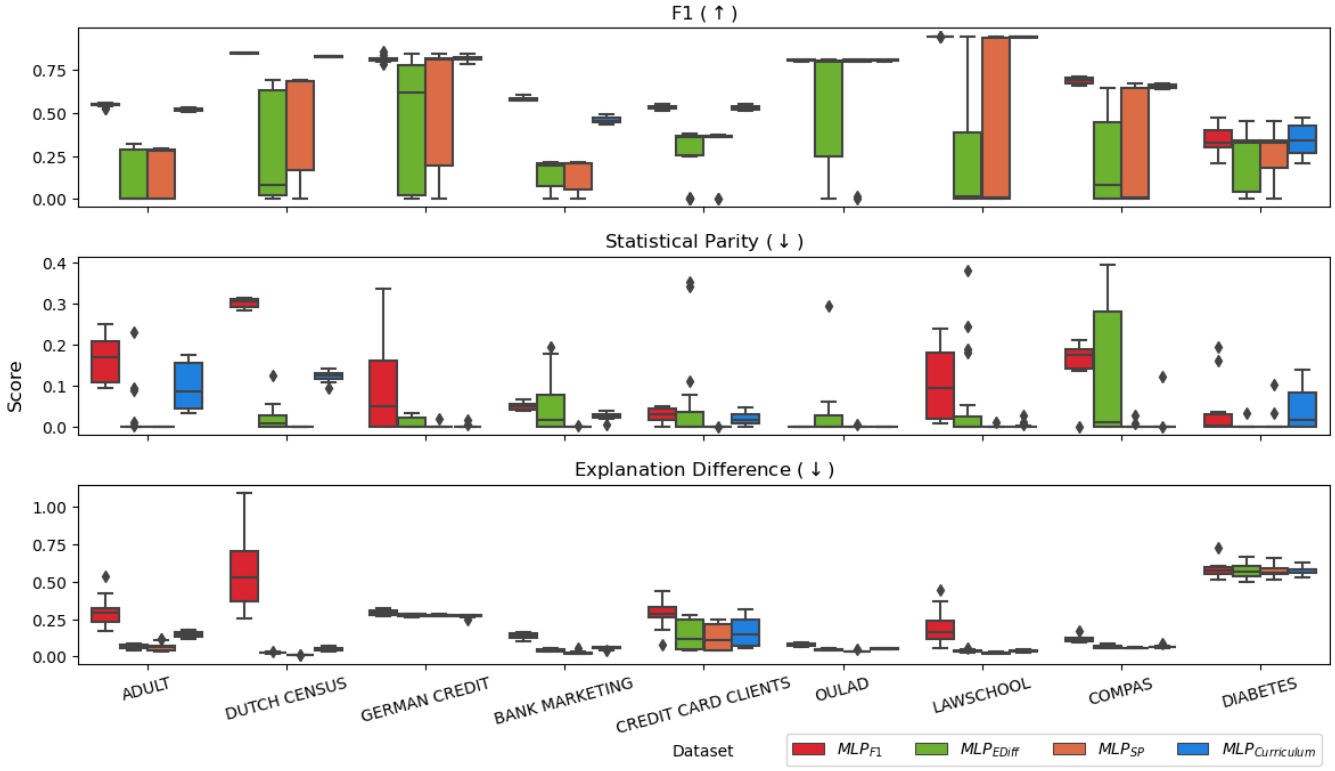

Figure 5: Box plots comparing the performance of four MLP loss functions in F1, SP, and EDiff across 10 trials in 9 datasets.

mit many more males than females with identical undergraduate GPAs. This can lead to a situation where females desiring to be accepted into law school must be over-qualified in terms of GPA compared to similar males. While the school may have fair results in admitting an equal number of males and females, its admission process is inherently unfair towards females. In contrast, our proposal promotes similar treatment while maintaining overall fairness.

The goal of a fairness model should be to converge towards a reality in which males and females are admitted on an equal basis for similar criteria. The guiding principle of our work is that males and females who are otherwise identical should be accepted at the same rate. Our proposal exemplifies this. In this case, the differences in explanations for males and females are much smaller, and, as our Experimental Evaluation demonstrates, this is accomplished while still achieving group fairness.

Limitations Our study was limited to classification tasks with single, binary protected attributes without consideration for multi-level attributes or the intersectionality of multiple protected attributes. Future work will look to expand our proposal to be adaptable to these more complex fairness problems. Additionally, our proposal relies on the fidelity of the chosen explainer to the underlying model. Explainer models cannot fully approximate a blackbox model such as an MLP and may contain their own biases. We used FastSHAP throughout this study for consistency and efficiency but different methods could yield different results. We envision future work exploring the impact of different model explanation techniques on both distributional and procedural fairness within the scope of our proposal.

	Dutch Census	Adult	Bank Marketing	Credit Card Clients	OULAD	Lawschool	COMPAS	German Credit	Diabetes
Samples — Features	60420 — 50	45222 — 94	45211 — 42	30000 — 82	21562 — 40	20798 — 18	6172 — 18	1000 — 47	173 — 272
Agarwal	1787.76 *(33.11)*	1476.11 *(27.34)*	1243.58 *(23.03)*	1262.92 *(23.39)*	437.32 *(8.10)*	1055.51 *(19.55)*	247.67 *(4.59)*	9.62 *(0.18)*	27.18 *(0.50)*
Hardt	87.51 *(1.62)*	21.68 *(0.40)*	14.69 *(0.27)*	19.34 *(0.36)*	7.75 *(0.14)*	5.94 *(0.11)*	8.56 *(0.16)*	0.22 *(0.00)*	0.23 *(0.00)*
Feldman	2.68 *(0.01)*	3.44 *(0.01)*	1.79 *(0.00)*	3.20 *(0.01)*	0.67 *(0.00)*	0.50 *(0.00)*	0.10 *(0.00)*	0.04 *(0.00)*	0.05 *(0.00)*
Adv. Deb.	0.24 *(0.24)*	0.18 *(0.18)*	0.17 *(0.17)*	0.12 *(0.12)*	0.09 *(0.09)*	0.08 *(0.08)*	0.03 *(0.03)*	0.01 *(0.01)*	0.01 *(0.01)*
xFair	8.82 *(8.82)*	7.77 *(7.77)*	8.65 *(8.65)*	10.11 *(10.11)*	3.02 *(3.02)*	4.19 *(4.19)*	0.95 *(0.95)*	0.40 *(0.40)*	0.26 *(0.26)*
MLP_{F1}	36.25 *(0.07)*	45.86 *(0.09)*	36.72 *(0.08)*	24.16 *(0.05)*	9.13 *(0.02)*	26.23 *(0.05)*	8.63 *(0.02)*	1.81 *(0.00)*	0.37 *(0.00)*
MLP_{SP}	18.68 *(0.12)*	13.70 *(0.08)*	13.57 *(0.08)*	9.52 *(0.06)*	6.57 *(0.04)*	6.36 *(0.04)*	2.41 *(0.01)*	0.73 *(0.00)*	0.16 *(0.00)*
MLP_{EDiff}	50.11 *(0.10)*	43.38 *(0.09)*	37.32 *(0.08)*	27.75 *(0.06)*	17.44 *(0.04)*	16.02 *(0.03)*	5.09 *(0.01)*	1.38 *(0.00)*	0.40 *(0.00)*
MLP_{All}	1544.82 *(9.54)*	892.47 *(5.51)*	887.71 *(5.48)*	433.09 *(2.67)*	68.74 *(0.42)*	91.62 *(0.57)*	18.07 *(0.11)*	1.62 *(0.01)*	1.08 *(0.01)*
$\text{MLP}_{Curriculum}$	224.56 *(1.39)*	145.46 *(0.90)*	142.65 *(0.88)*	85.46 *(0.53)*	17.42 *(0.11)*	21.34 *(0.13)*	7.57 *(0.05)*	1.17 *(0.01)*	0.84 *(0.01)*

Table 6: Average total processing time in seconds to train models. The number in parentheses shows the average time to fit a single model, normalized for the size of the hyper-parameter grid search. Training times under 0.05 seconds are listed as 0.0.

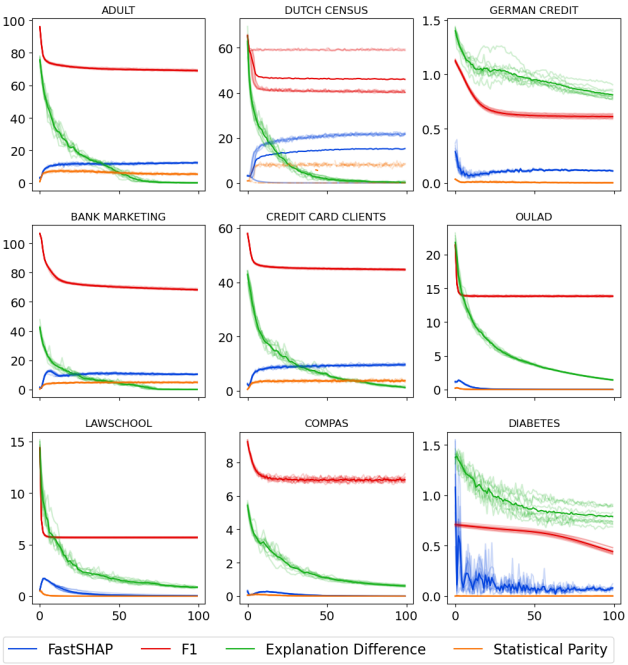

Figure 6: Evolution of the 4 different loss functions in $\text{MLP}_{Curriculum}$. 10 models were fit for 100 epochs. Each faded light represents a single model. The dark lines show the average loss at each epoch.

Conclusion

We propose a new method of measuring fairness to address an urgent gap in existing fairness research. By focusing solely on predicted outcomes, fair machine learning methods overlook the impact of how decisions were made. Instead, we argue for a multi-objective approach to fairness focusing on predictive performance, group fairness, and procedural fairness. Through our experimental evaluation, we demonstrate that our proposed method is superior to the state-of-the-art in optimizing each of these three objectives simultaneously without a significant increase in processing time. Our results illustrate the need to include explanation fairness as another dimension in fairness analysis. Our proposal paves the way for future work evaluating fairness as a multi-objective optimization task. We envision expanding our proposal to consider more complex fairness situations

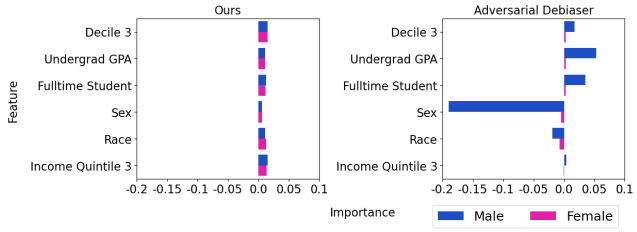

Figure 7: Example of SHAP explanations for the best and worst performing model in the Lawschool dataset. Selected features are the union of the three largest feature importance differences for each of the models (six total).

such as non-binary protected attributes or intersectionality and further improving its efficiency. We make all data and code including an appendix with additional results publicly available.

Acknowledgments

The authors would like to thank the reviewers for their insights and suggestions. Additionally, the authors would like to thank the Lucy Family Institute for Data and Society at the University of Notre Dame for their support.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International conference on machine learning*, 60–69. PMLR.

Agarwal, A.; Dudik, M.; and Wu, Z. S. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 120–129. PMLR.

Anthis, J.; and Veitch, V. 2023. Causal context connects counterfactual fairness to robust prediction and group fairness. *Advances in Neural Information Processing Systems*, 36: 34122–34138.

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and

challenges toward responsible AI. *Information fusion*, 58: 82–115.

Begley, T.; Schwedes, T.; Frye, C.; and Feige, I. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Benjamin, R. 2019. *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, 149–159.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Bisconti, P.; Aquilino, L.; Marchetti, A.; and Nardi, D. 2024. A Formal Account of Trustworthiness: Connecting Intrinsic and Perceived Trustworthiness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 131–140.

Breiman, L. 2001. Random forests. *Machine learning*, 45: 5–32.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.

Carvalho, T.; Moniz, N.; and Antunes, L. 2023. A Three-Way Knot: Privacy, Fairness, and Predictive Performance Dynamics. In Moniz, N.; Vale, Z.; Cascalho, J.; Silva, C.; and Sebastião, R., eds., *Progress in Artificial Intelligence*, 55–66. Cham: Springer Nature Switzerland. ISBN 978-3-031-49008-8.

Center, M. P. 2013. Integrated Public Use Microdata Series International.

Chuang, Y.-N.; Wang, G.; Yang, F.; Liu, Z.; Cai, X.; Du, M.; and Hu, X. 2023a. Efficient xai techniques: A taxonomic survey. *arXiv preprint arXiv:2302.03225*.

Chuang, Y.-N.; Wang, G.; Yang, F.; Zhou, Q.; Tripathi, P.; Cai, X.; and Hu, X. 2023b. Cortx: Contrastive framework for real-time explanation. *arXiv preprint arXiv:2303.02794*.

Dai, J.; Upadhyay, S.; Bach, S. H.; and Lakkaraju, H. 2021. What will it take to generate fairness-preserving explanations? *arXiv preprint arXiv:2106.13346*.

De Weck, O. L. 2004. Multiobjective optimization: History and promise. In *Invited Keynote Paper, GL2-2, The Third China-Japan-Korea Joint Symposium on Optimization of Structural and Mechanical Systems, Kanazawa, Japan*, volume 2, 34.

Deb, K.; and Datta, R. 2012. Hybrid evolutionary multi-objective optimization and analysis of machining operations. *Engineering Optimization*, 44(6): 685–706.

Decker, M. C.; Wegner, L.; and Leicht-Scholten, C. 2025. Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics and Information Technology*, 27(1): 1.

Dodgson, J. S.; Spackman, M.; Pearman, A.; and Phillips, L. D. 2009. Multi-criteria analysis: a manual.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Fleisher, W. 2021. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 480–490.

Fragkathoulas, C.; Papanikou, V.; Karidi, D. P.; and Pitoura, E. 2024. On Explaining Unfairness: An Overview. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, 226–236.

Fu, R.; Aseri, M.; Singh, P. V.; and Srinivasan, K. 2022. "Un" fair machine learning algorithms. *Management Science*, 68(6): 4173–4195.

Germino, J.; Moniz, N.; and Chawla, N. V. 2024. FairMOE: counterfactually-fair mixture of experts with levels of interpretability. *Machine Learning*, 113(9): 6539–6559.

Green, B. 2020. The false promise of risk assessments: Epistemic reform and the limits of fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 594–606.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Gunantara, N. 2018. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1): 1502242.

Han, D.; Moniz, N.; and Chawla, N. V. 2024. AnyLoss: Transforming Classification Metrics into Loss Functions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 992–1003.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Ilvento, C. 2019. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*.

Jethani, N.; Sudarshan, M.; Covert, I. C.; Lee, S.-I.; and Ranganath, R. 2021. Fastshap: Real-time shapley value estimation. In *International conference on learning representations*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Kuzilek, J.; Hlosta, M.; and Zdrahal, Z. 2017. Open university learning analytics dataset. *Scientific data*, 4(1): 1–8.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3): e1452.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Lünich, M.; and Keller, B. 2024. Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1031–1042.

Ma, X.; Wang, Z.; and Liu, W. 2022. On the tradeoff between robustness and fairness. *Advances in Neural Information Processing Systems*, 35: 26230–26241.

Maity, S.; Mukherjee, D.; Yurochkin, M.; and Sun, Y. 2020. There is no trade-off: enforcing fairness can improve accuracy. *stat*, 1050: 6.

Mardle, S.; Pascoe, S.; and Tamiz, M. 2000. An investigation of genetic algorithms for the optimization of multi-objective fisheries bioeconomic models. *International Transactions in Operational Research*, 7(1): 33–49.

Menon, A. V.; Omar, Z. A.; Nahar, N.; Papademetris, X.; Fiellin, L. E.; and Kästner, C. 2024. Lessons from Clinical Communications for Explainable AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 958–970.

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

Mukherjee, D.; Yurochkin, M.; Banerjee, M.; and Sun, Y. 2020. Two simple ways to learn individual fairness metrics from data. In *International conference on machine learning*, 7097–7107. PMLR.

Murata, T.; Ishibuchi, H.; and Tanaka, H. 1996. Multi-objective genetic algorithm and its applications to flowshop scheduling. *Computers & industrial engineering*, 30(4): 957–968.

Peng, K.; Chakraborty, J.; and Menzies, T. 2022. FairMask: Better Fairness via Model-based Rebalancing of Protected Attributes. *IEEE Transactions on Software Engineering*, 1–14.

Pessach, D.; and Shmueli, E. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44.

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34: 25944–25955.

Pfeiffer, J.; Gutschow, J.; Haas, C.; Möslein, F.; Maspfuhl, O.; Borgers, F.; and Alpsancar, S. 2023. Algorithmic fairness in AI: an interdisciplinary view. *Business & Information Systems Engineering*, 65(2): 209–222.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass: The Belknap press of Harvard University Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6): 386.

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

Rueda, J.; Rodríguez, J. D.; Jounou, I. P.; Hortal-Carmona, J.; Ausín, T.; and Rodríguez-Arias, D. 2024. "Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI & society*, 39(3): 1411–1422.

Sharifi-Malvajerdi, S.; Kearns, M.; and Roth, A. 2019. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32.

Sharma, S.; and Kumar, V. 2022. A comprehensive review on multi-objective optimization techniques: Past, present and future. *Archives of Computational Methods in Engineering*, 29(7): 5605–5633.

Strack, B.; DeShazo, J. P.; Gennings, C.; Olmo, J. L.; Ventura, S.; Cios, K. J.; Clore, J. N.; et al. 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014.

Tapia, M. G. C.; and Coello, C. A. C. 2007. Applications of multi-objective evolutionary algorithms in economics and finance: A survey. In *2007 IEEE congress on evolutionary computation*, 532–539. IEEE.

Thibaut, J.; and Walker, L. 1975. *Procedural Justice: A Psychological Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tian, Y.; Si, L.; Zhang, X.; Cheng, R.; He, C.; Tan, K. C.; and Jin, Y. 2021. Evolutionary large-scale multi-objective optimization: A survey. *ACM Computing Surveys (CSUR)*, 54(8): 1–34.

Waller, M.; Rodrigues, O.; and Cocarascu, O. 2024. Identifying Reasons for Bias: An Argumentation-Based Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21664–21672.

Wang, G.; Chuang, Y.-N.; Du, M.; Yang, F.; Zhou, Q.; Tripathi, P.; Cai, X.; and Hu, X. 2022. Accelerating shapley explanation via contributive cooperator selection. In *International Conference on Machine Learning*, 22576–22590. PMLR.

Wang, S.; and Wu, Y. 2024. Achieving Equalized Explainability Through Data Reconstruction. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

Wang, Z.; Huang, C.; and Yao, X. 2024. Procedural fairness in machine learning. *arXiv preprint arXiv:2404.01877*.

Wei, S.; and Niethammer, M. 2022. The fairness-accuracy Pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3): 287–302.

Wick, M.; Tristan, J.-B.; et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.

Wightman, L. F. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, X.; Khalili, M. M.; and Liu, M. 2020. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3): 7–11.

Zhao, Y.; Wang, Y.; and Derr, T. 2023. Fairness and explainability: Bridging the gap towards fair model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11363–11371.

Zhou, J.; Chen, F.; and Holzinger, A. 2020. Towards explainability for AI fairness. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 375–386. Springer.

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