FAIRER: Fairness as Decision Rationale Alignment

Tianlin Li [1] **Qing Guo** [2 3] **Aishan Liu** [4] **Mengnan Du** [5] **Zhiming Li** [1] **Yang Liu** [1]

Abstract

Deep neural networks (DNNs) have made significant progress, but often suffer from fairness issues, as deep models typically show distinct accuracy differences among certain subgroups (*e.g.*, males and females). Existing research addresses this critical issue by employing fairness-aware loss functions to constrain the last-layer outputs and directly regularize DNNs. Although the fairness of DNNs is improved, it is unclear how the trained network makes a fair prediction, which limits future fairness improvements. In this paper, we investigate fairness from the perspective of decision rationale and define the *parameter parity score* to characterize the fair decision process of networks by analyzing neuron influence in various subgroups. Extensive empirical studies show that the unfair issue could arise from the unaligned decision rationales of subgroups. Existing fairness regularization terms fail to achieve decision rationale alignment because they only constrain last-layer outputs while ignoring intermediate neuron alignment. To address the issue, we formulate the fairness as a new task, *i.e.*, *decision rationale alignment* that requires DNNs' neurons to have consistent responses on subgroups at both intermediate processes and the final prediction. To make this idea practical during optimization, we relax the naive objective function and propose *gradient-guided parity alignment*, which encourages gradient-weighted consistency of neurons across subgroups. Extensive experiments on a variety of datasets show that our method can significantly enhance fairness while sustaining a high level of accuracy and outperforming other approaches by a wide margin.

1. Introduction

Deep neural networks (DNNs) are increasingly being used in high-stakes applications in our society. However, as deep learning is increasingly adopted for many applications that have brought convenience to our daily lives (He et al., 2016; Devlin et al., 2019; Deng et al., 2013), DNNs still suffer from the fairness problem and often exhibit undesirable discrimination behaviors (News, 2021; 2020). For example, for an intelligent task (*e.g.*, salary prediction), a trained DNN easily presents distinct accuracy values in different subgroups (*e.g.*, males and females). The discriminatory behaviors contradict people's growing demand for fairness, which would cause severe social consequences. To alleviate such fairness problems, a line of mitigation strategies has been constantly proposed (Zemel et al., 2013; Sarhan et al., 2020; Wang et al., 2019).

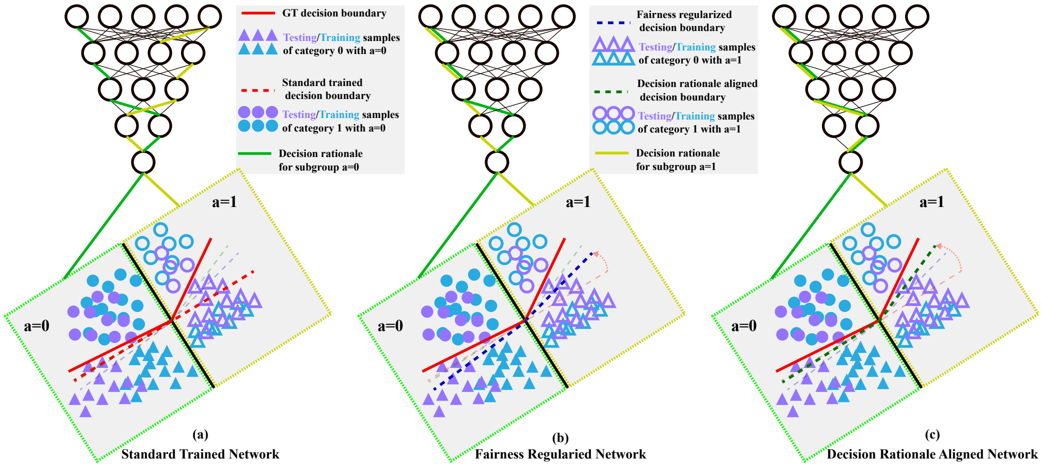
A direct regularization method to improve fairness is to relax fairness metrics as constraints in the training process (Madras et al., 2018). This regularization method is designed to reduce the disparities between different subgroups in the training and testing data (See Fig. 1 (a) vs. (b)). Although this method easily improves the fairness of DNN models, it is still unclear how the trained network makes a fair decision[1]. For example, we do not know *how the fairness regularization terms actually affect the final network parameters and let them make a fair prediction.* Without such an understanding, we would not know the effective direction for further fairness enhancement. Existing work does not address this question and the majority of them concentrate on the last-layer outputs (*i.e.*, predictions) while ignoring the internal process. In this work, we propose to study the fairness from the perspective of decision rationale and analyze existing fairness-regularized methods through a *decision-rationale-aware analysis* method. The term 'decision rationale' is known as the reason for making a decision and could be represented as the influence of neurons in a DNN (Khakzar et al., 2021). Specifically, for each intermediate neuron (*i.e.*, a parameter of the DNN [2]), we can calculate the loss change on a subgroup before and after

[1] Nanyang Technological University, Singapore [2] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research, Singapore [3] Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research, Singapore [4] Beihang University, China [5] New Jersey Institute of Technology, USA. Corresponding author: Qing Guo <tsingqguo@ieee.org>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1] The 'decision' here means the prediction results of the DNN regarding given inputs. The name follows the interpretable works (Du et al., 2019; Wang et al., 2018b; Khakzar et al., 2021).

[2] We follow Molchanov et al. (2016; 2019) to use the terms "neuron" and "parameter" interchangeably.


GT decision boundary
Testing/Training samples of category 0 with a=0
Standard trained decision boundary
Testing/Training samples of category 1 with a=0
Decision rationale for subgroup a=0
Fairness regularized decision boundary
Testing/Training samples of category 0 with a=1
Decision rationale aligned decision boundary
Testing/Training samples of category 1 with a=1
Decision rationale for subgroup a=1
a=1
a=0
a=1
a=0
a=1
a=0

(a) Standard Trained Network

(b) Fairness Regularied Network

(c) Decision Rationale Aligned Network

Figure 1: Schematic diagrams of two existing solutions and the proposed one. (a) and (b) represent results of the standard trained network and the regularized fairness network. (c) show the results of the decision rationale-aligned network. The previous work, *i.e.*, fairness regularization-based method, adds a regularization term to the final loss function to make the trained network have similar predictions on the two subgroups, which makes the "decision rationales" of the trained network on the two subgroups become partially similar (See the green solid line for the subgroup and yellow solid line for the subgroup in (b)). In contrast, our method is to add a "decision rationale" alignment explicitly and make "decision rationales" on the two subgroups consistent.

removing the neuron. As a result, we can characterize the decision rationale of a network on the subgroup by collecting the loss changes of all neurons. For example, the solid green and yellow lines in Fig. 1 represent the neurons leading to high loss changes at each layer and characterize the decision rationales of the two subgroups. Then, we define the *parameter parity score* as the decision rationale shifting across different subgroups, which actually reveals the influences of intermediate neurons (*i.e.*, parameters) to the decision rationale changes. With the new analysis tool, we find that the network fairness is directly related to the consistency of the decision rationales on different subgroups, and existing fairness regularization terms could only partially achieve this goal, which restricts the fairness improvement (Compare the solid lines in Fig. 1 (b)) since they only add constraints to the final outputs. Intuitively, we could define new regularization terms to minimize parity scores of all neurons and encourage them to have similar influence across subgroups. We name this new task as the *decision rationale alignment* that requires DNNs to have consistent decision rationales as well as final predictions on different subgroups. Although straightforward, the task is challenging for two reasons: *First*, the decision rationale and parity score are defined based on a dataset and it is impractical to calculate them at each iteration during the training process. *Second*, different neurons have different effects on fairness and such differences should be carefully considered.

To address the above two challenges, we propose the *gradient-guided parity alignment* method by relaxing the calculation of decision rationale from the dataset-based strategy to the sample-based one. As a result, the corresponding regularization term is compatible with the epoch-based training process. Moreover, we use the first-order Taylor expansion to approximate the parity score between decision rationales, and the effects of different neurons on the fairness are weighted via their gradient magnitudes automatically. Overall, the proposed method can achieve much higher fairness than state-of-the-art methods. In summary, the work makes the following contributions:

1. To understand how a network makes a fair decision, we define *parameter parity score* to characterize the decision rationales of the network on different subgroups. We reveal that the fairness of a network is directly related to the consistency of its decision rationales on different subgroups and existing regularization terms cannot achieve this goal.

2. To train a fairer network, we formulate the *decision rationale alignment* task and propose the *gradient-guided parity alignment* method to solve it by addressing the complex optimization challenges.

3. Extensive experiments on three public datasets, *i.e.*, Adult, CelebA, and Credit, demonstrate that our method can enhance the fairness of DNNs effectively and outperform others largely.

2. Preliminaries

2.1. Problem Formulation

In general, given a dataset $\mathcal{D}$ containing data samples (*i.e.*, $\mathbf{x} \in \mathcal{X}$) and corresponding labels (*i.e.*, $y \in \mathcal{Y}$), we can train a DNN to predict the labels of input samples, *i.e.*, $\hat{y} = \mathrm{F}(\mathbf{x})$ with $\hat{y} \in \mathcal{Y}$ being the prediction results. In the real world, the samples might be divided into subgroups according to some sensitive attributes $a \in \mathcal{A}$ such as gender and race. Without loss of generality, we consider the binary classification and binary attribute setup, *i.e.*, $y \in \{0, 1\}$ and $a \in \{0, 1\}$. For example, $a = 0$ and $a = 1$ could represent males and females, respectively. A fair DNN (*i.e.*, $\mathrm{F}(\cdot)$) is desired to obtain a similar accuracy in the two subgroups.

2.2. Fairness Regularization

Among fairness evaluation measures, Demographic Parity (DP) and Equalized Odds (EO) are most frequently adopted in deep learning fairness research. Specifically, Feldman et al. (2015) develop the DP metric to encourage the predicted label to be independent of the sensitive attribute (*i.e.*, a), that is, $P(\hat{y}|a = 0) = P(\hat{y}|a = 1)$ which means that the probability distribution of $\hat{y}$ condition on $a = 0$ should be the same as the condition on $a = 1$. Hardt et al. (2016a) further propose the EO metric to consider the ground truth label y and make the prediction and sensitive attribute conditionally independent w.r.t. y, *i.e.*, $P(\hat{y}|a = 0, y) = P(\hat{y}|a = 1, y)$. Although straightforward, it is difficult to optimize the above measures and existing fairness works (Madras et al., 2018; Chuang & Mroueh, 2021) focus on designing fairness regularization terms and adding them to the loss function, which encourages the targeted DNN to predict similar results across subgroups. Madras et al. (2018) propose relaxed counterparts:

$$\Delta\mathrm{DP}(\mathrm{F}) = \left|\mathrm{E}_{\mathbf{x}\sim P_0}(\mathrm{F}(\mathbf{x})) - \mathrm{E}_{\mathbf{x}\sim P_1}(\mathrm{F}(\mathbf{x}))\right|, \qquad (1)$$

where $P_0 = P(\mathbf{x}|a = 0)$ and $P_1 = P(\mathbf{x}|a = 1)$ are the distributions of $\mathbf{x}$ condition on $a = 0$ and $a = 1$, respectively, and the function $\mathrm{E}(\cdot)$ is to calculate the expectation under the distributions.

$$\Delta\mathrm{EO}(\mathrm{F}) = \sum_{y\in\{0,1\}} \left|\mathrm{E}_{\mathbf{x}\sim P_0^y}(\mathrm{F}(\mathbf{x})) - \mathrm{E}_{\mathbf{x}\sim P_1^y}(\mathrm{F}(\mathbf{x}))\right|, \quad (2)$$

where $P_0^1 = P(\mathbf{x}|a = 0, y = 1)$ denotes the distribution of $\mathbf{x}$ condition on the $a = 0$ and $y = 1$, and we have similar notations for P_0^0, P_1^1, P_1^0 if we set the DNN for a binary classification task and have the label $y \in 0, 1$. We can add Eq. (1) and Eq. (2) to the classification loss (*e.g.*, cross-entropy loss) to regularize the fairness of the targeted DNN, respectively, and obtain the whole loss function

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\mathrm{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda\mathcal{L}_{\mathrm{fair}}(\mathrm{F}), \qquad (3)$$

where P denotes the joint distribution of $\mathbf{x}$ and y, $\mathcal{L}_{\mathrm{cls}}$ is the classification loss, and the term $\mathcal{L}_{\mathrm{fair}}$ could be $\Delta\mathrm{DP}(\mathrm{F})$ or $\Delta\mathrm{EO}(\mathrm{F})$ defined in Eq. (1) or Eq. (2). We can minimize the above loss function and get fairness-regularized DNNs. Although effective, the above method presents some generalization limitations. To alleviate this issue, (Chuang & Mroueh, 2021) embed the data augmentation strategy into the fairness regularization method and propose FairMixup with novel DP- and EP-dependent regularization terms. Please refer to Chuang & Mroueh (2021) for details.

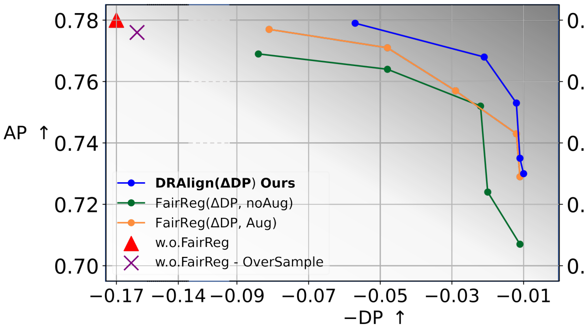


Figure 2: Accuracy and fairness comparison of five different methods on the Adult dataset. The hyperparameter λ increases from 0.2 to 0.6 along the $-$DP axis as it becomes larger.

Overall, we get several fairness regularization methods via different regularization terms. Specifically, we denote the methods without augmentation as $\mathrm{FairReg}(\Delta\mathrm{DP}, \mathrm{noAug})$ and $\mathrm{FairReg}(\Delta\mathrm{EO}, \mathrm{noAug})$ based on regularization functions (*i.e.*, Eq. (1) and Eq. (2)). We denote the methods equipped with data augmentation as $\mathrm{FairReg}(\Delta\mathrm{DP}, \mathrm{Aug})$ and $\mathrm{FairReg}(\Delta\mathrm{EO}, \mathrm{Aug})$, respectively.

2.3. Observations

We conduct an experiment on the Adult dataset (Dua & Graff, 2017a) with a neural network with 3-layer MLPs. Specifically, we train the network with two fairness regularization methods (*i.e.*, $\mathrm{FairReg}(\Delta\mathrm{DP}, \mathrm{noAug})$ and $\mathrm{FairReg}(\Delta\mathrm{DP}, \mathrm{Aug})$ [3]) and five different $\lambda \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, that is, for each method, we get five trained networks. Then, we can calculate the accuracy scores and fairness scores of all networks on the testing dataset. We employ average precision for the accuracy score and $-$DP for the fairness score since a smaller DP means better fairness. For each method, we can draw a plot w.r.t. different λ. Besides, we also train a network without the fairness regularization term and denote it as w.o.FairReg. Based on w.o.FairReg, we can conduct oversampling on the training samples to balance the samples across different subgroups (Wang et al., 2020) and denote it as w.o.FairReg-Oversample. As shown in Fig.2, we see that: ❶ The standard trained network via w.o.FairReg presents an obvious fairness issue and the oversampling solution has limited capability to fix it. ❷ When we use the regularization methods and gradually increase the weight λ in Eq. (3) from 0.2 to 0.6, $\mathrm{FairReg}(\Delta\mathrm{DP}, \mathrm{noAug})$ is able to generate fairer networks with higher fairness scores (*i.e.*, higher -DP) than the one from w.o.FairReg. However, the corresponding accuracy decreases by a large margin, that is, *existing methods could hardly generate enough fair networks under similar accuracy.* ❸ The data augmentation-

[3] We have similar observations on the ΔEO-based methods and remove them for a clear explanation.

based method (*i.e.*, FairReg(ΔDP, Aug)) can alleviate such an issue to some extent and achieves higher fairness than FairReg(ΔDP, noAug) under similar accuracy.

Such fairness regularization methods neglect the decision-making process and might generate sub-optimal models. Although intuitively having a consistent decision process among various groups could enhance model performance in terms of fairness, we still empirically explore the connection between the decision-making process and fairness. We provide an analysis method by extending the decision rationale-aware explainable methods in Sec. 3. Specifically, instead of using the final fairness metrics, we define the parameter parity score for each parameter of a network that measures whether the parameter is fair, that is, whether it has consistent responses to different subgroups.

3. Decision Rationale-aware Fairness Analysis

In recent years, decision rationale-aware explainable methods are developed and help understand how a trained network makes a decision (Khakzar et al., 2021; Wang et al., 2018a). In these works, the decision rationale is represented by measuring the importance of intermediate neurons. Inspired by this idea, to understand a fair decision, we study the decision process of networks by analyzing their neuron influence under different subgroups, and define the decision rationales for different subgroups. Then, we define the *parity score* for a network that actually measures whether the decision rationales on different subgroups are consistent. Besides, we can use the parity score to compare the networks trained with different regularization terms.

3.1. Parameter Parity Score

Inspired by recent work on understanding the importance of the neuron for the classification loss (Molchanov et al., 2019), we define the parameter parity score based on the independent assumption across neurons (*i.e.*, parameters)[4]. When we have a trained network $\mathrm{F}(\cdot)$ with its parameters $\mathcal{W} = \{w_0, \ldots, w_K\}$, we can calculate classification losses on samples from two distributions $P_0 = P((\mathbf{x}, y)|a = 0)$ and $P_1 = P((\mathbf{x}, y)|a = 1)$ which correspond to the training subsets of two subgroups (*i.e.*, $a = 0$ and $a = 1$), and get the losses $\mathcal{J}(\mathrm{F}, P_0)$ and $\mathcal{J}(\mathrm{F}, P_1)$, respectively. Meanwhile, we can modify $\mathrm{F}(\cdot)$ by removing a specific parameter w_k and denote the new counterpart as $\mathrm{F}_{w_k=0}$, and we can also obtain losses via $\mathcal{J}(\mathrm{F}_{w_k=0}, P_0)$ and $\mathcal{J}(\mathrm{F}_{w_k=0}, P_1)$. Then, for each subgroup (*i.e.*, P_0 or P_1), we calculate the loss change before and after removing the parameter w_k by

$$c_k^{a=i} = \mathrm{C}(\mathrm{F}, w_k, P_i) = |\mathcal{J}(\mathrm{F}, P_i) - \mathcal{J}(\mathrm{F}_{w_k=0}, P_i)|^2, \quad (4)$$
$$\forall i \in \{0, 1\}, k \in [0, K],$$

where the function $\mathcal{J}(\mathrm{F}, P_i)$ is to calculate the classification loss (*i.e.*, $\mathcal{L}_{\mathrm{cls}}$ in Eq. (3)) of examples in P_i with $\forall i \in \{0, 1\}$ based on the network F. With a subgroup P_i and a K-neuron network F, we can get $\mathbf{c}_{\mathrm{F}}^{a=i} = [c_0^{a=i}, c_1^{a=i}, \ldots, c_K^{a=i}]$ that is regarded as a representation of the decision rationale on the subgroup P_i (Khakzar et al., 2021).

Then, we define the parity score of the parameter w_k as the difference between $c_k^{a=0} = \mathrm{C}(\mathrm{F}, w_k, P_0)$ and $c_k^{a=1} = \mathrm{C}(\mathrm{F}, w_k, P_1)$, *i.e.*,

$$d_k = |\mathrm{C}(\mathrm{F}, w_k, P_0) - \mathrm{C}(\mathrm{F}, w_k, P_1)|^2. \quad (5)$$

Intuitively, if the network F is fair to a kind of sensitive attribute, each parameter should have consistent responses to different subgroups, and the changes before and after removing the parameter should be the same. As a result, a smaller d_k means that the parameter w_k is less sensitive to the attribute changes. For the entire network with K neurons, we get K parity scores and $\mathbf{c}_{\mathrm{F}}^{a=i} = [c_0^{a=i}, c_1^{a=i}, \ldots, c_K^{a=i}]$, and can represent the network with $\mathbf{d}_{\mathrm{F}} = [d_0, d_1, \ldots, d_K]$ and aggregate all scores for a network-level parity score, *i.e.*, $d_{\mathrm{F}} = \sum_{k=0}^{K} d_k = |\mathbf{c}_{\mathrm{F}}^{a=0} - \mathbf{c}_{\mathrm{F}}^{a=1}|_1$, which measures whether the decision rationales on the two subsets are consistent (*i.e.*, properly aligned).

3.2. Relationship between Parity Score and Fairness

With the *parameter parity score*, we conduct an empirical study based on the Adult dataset and a neural network with 3-layer MLPs. Specifically, we train six networks with the regularization terms defined in Sec. 2.3, *e.g.*, the ΔDP-based regularization terms with six different weights (*i.e.*, FairReg(ΔDP, noAug) with $\lambda \in \{0.0, 0.2, 0.3, 0.4, 0.5, 0.6\}$). Note that, FairReg($\Delta$DP, noAug) with $\lambda = 0.0$ represents the standard trained network without fair regularization terms (*i.e.*, w.o.FairReg). Then, for each method, we can train a neural network and calculate the parity score, *i.e.*, $d_{\mathrm{F}} = \sum_{k=0}^{K} d_k = |\mathbf{c}_{\mathrm{F}}^{a=0} - \mathbf{c}_{\mathrm{F}}^{a=1}|_1$ to measure the decision rationale shifting across subgroups and the fairness score defined by $-$DP. As reported in Table 1, we see that: ❶ the parity score of the network gradually decreases as the DP becomes smaller, which demonstrates that the fairness of a network is highly related to the decision rationale shifting across subgroups. ❷ adding the fairness regularization term on the last-layer outputs (*i.e.*, $\lambda > 0$) can decrease the decision rationale shifting to some extent. However, such an indirect way could hardly achieve the optimized results and a more effective way is to actively align the decision rationale explicitly. Note that we can observe similar results on other regularization methods and focus on FairReg(ΔDP, noAug) due to the limited space. We conclude that the existing fairness regularization-based methods can only encourage the consistency between decision rationales of the network on different subgroups to some extent. This inspires our

[4] More details about this assumption are deferred to A.13.

Table 1: Parity scores, fairness scores, and the first-order Taylor approximation of the parity scores of networks trained via FairReg(ΔDP, noAug) with different λ in Eq. (3). For each network, we train 10 runs with different seeds and the average results are reported.

	FairReg(ΔDP, noAug)					
	$\lambda = 0.0$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$
Parity score (d_{F})	0.624	0.391	0.101	0.070	0.046	0.039
Fairness ($-$DP)	-0.160	-0.084	-0.048	-0.022	-0.020	-0.010
Approx. ($-\sum_{l=0}^{L}\cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1})$)	-0.670	-1.382	-1.530	-1.629	-1.631	-1.800

method in Sec.4 that conducts alignment of the decision rationales of different subgroups explicitly. It is worth noting that the parameter parity score is the most straightforward way to measure whether the parameter has consistent responses to different subgroups and represents the degree of the decision rationale alignment.

4. Decision Rationale Alignment

4.1. Formulation and Challenges

According to Eq. (5), we can achieve a fairer network by aligning the decision rationales of subgroups and a straightforward way is to set the parity score $d_{\mathrm{F}} = \sum_{k=0}^{K} d_k$ as an extra loss function and minimize it directly, that is, we can add a new loss to Eq. (3) and have

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\mathrm{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda \mathcal{L}_{\mathrm{fair}}(\mathrm{F}) + \beta \sum_{k=0}^{K} d_k, \tag{6}$$

where d_k is the parity score of the kth neuron and calculated by Eq. (5). Such a loss should calculate parity scores for all neurons and all samples in a dataset, leading to a high cost and is not practical.

4.2. Gradient-guided Parity Alignment

To address the challenges, we relax Eq. (4) to the sample-based counterpart

$$\begin{aligned} c_k^{a=i} = \mathrm{C}(\mathrm{F}, w_k, P_i) =& |\mathrm{E}_{(\mathbf{x},y)\sim P_i}(\mathcal{L}_{\mathrm{cls}}(\mathrm{F}(\mathbf{x}), y) - \\ & \mathrm{E}_{(\mathbf{x},y)\sim P_i}(\mathcal{L}_{\mathrm{cls}}(\mathrm{F}_{w_k=0}(\mathbf{x}), y))|^2, \\ & \forall i \in \{0,1\}, k \in [0, K]. \end{aligned} \tag{7}$$

We use the first-order Taylor expansion to approximate $c_k^{a=i}$ similar to Molchanov et al. (2019) and get

$$\hat{c}_k^{a=i} = \hat{\mathrm{C}}(\mathrm{F}, w_k, P_i) = (g_k^{a=i} \cdot w_k)^2,\ \forall i \in \{0,1\}, k \in [0, K]. \tag{8}$$

where $g_k^{a=i}$ denotes the gradient of the kth neuron (*i.e.*, w_k) w.r.t. the loss function on the examples sampled from the distribution of the ith subgroup (*i.e.*, P_i). Intuitively, the above definition means that we should pay more attention to the neurons with higher gradients and make them have similar responses to different subgroups. However, neurons (*i.e.*, parameters) of different layers may have different score ranges. To avoid this influence, we further normalize $\hat{c}_k^{a=i}$ by $\frac{\hat{c}_k^{a=i}}{|\hat{\mathbf{c}}_l^{a=i}|}$ $\forall i \in \{0,1\}, k \in \mathcal{K}_l$, where $\mathcal{K}_l$ contains the indexes of the neurons in the lth layer, and parity scores of neurons in the lth layer (*i.e.*, $\{\hat{c}_k^{a=i} | k \in \mathcal{K}_l\}$) form a vector $\hat{\mathbf{c}}_l^{a=i} = \mathrm{vec}(\{\hat{c}_k^{a=i} | k \in \mathcal{K}_l\})$. Then, we can get a new vector for the lth layer $\vec{\mathbf{c}}_l^{a=i} = \mathrm{vec}(\{\frac{\hat{c}_k^{a=i}}{|\hat{\mathbf{c}}_l^{a=i}|} | k \in \mathcal{K}_l\})$ by normalizing each element. Then, we can update Eq. (6) by minimizing the distance between $\vec{\mathbf{c}}_l^{a=0}$ and $\vec{\mathbf{c}}_l^{a=1}$ $\forall l \in [0, L]$, *i.e.*,

$$\begin{aligned} \mathcal{L} =& \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\mathrm{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda \mathcal{L}_{\mathrm{fair}}(\mathrm{F}) \\ & - \beta \sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1}), \end{aligned} \tag{9}$$

where L denotes the number of layers in the network, and the function $\cos(\cdot)$ is the cosine similarity function. The last two terms are used to align the final predictions and the responses of the intermediate neurons across subgroups, respectively. To validate the approximation (*i.e.*, $-\sum_{l=0}^{L}\cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1})$) can reflect the decision rationale alignment degree like the parity score $\sum_{k=0}^{K} d_k$, we conduct an empirical study on FairReg(ΔDP, noAug) as done in Sec. 3.2 and calculate the value of $-\sum_{l=0}^{L}\cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1})$ for all trained networks. From Table 1, we see that the approximation has a consistent variation trend with the parity score under different λ.

4.3. Implementation Details

We detail the whole training process in Algorithm 1. In particular, given a training dataset $\mathcal{D}$, we first sample two groups of samples (ie, $(\mathbf{X}_0, \mathbf{Y}_0)$ and $(\mathbf{X}_1, \mathbf{Y}_1)$) from the two subgroups in the dataset, respectively (See lines 4 and 5). Then, we calculate the cross-entropy loss for both sample groups (See line 7) and calculate the fairness regularization loss (*i.e.*, $\mathcal{L}_{\mathrm{fair}} = \Delta\mathrm{DP}(\mathbf{F}, \mathbf{X}_0, \mathbf{X}_1)$). After that, we can calculate the gradient of each parameter (*i.e.*, neuron w_k) w.r.t. the classification loss (See lines 11 and 12) and calculate the decision rationale for each neuron and layer (See lines 16 and 17). Finally, we calculate the cosine similarity between $\vec{\mathbf{c}}_l^{a=0}$ and $\vec{\mathbf{c}}_l^{a=1}$ and use the whole loss to update the parameters. We defer the algorithm depiction for the EO metric to the Appendix (A.2).

5. Experiments

5.1. Experimental Setup

Datasets. In our experiments, we use two tabular benchmarks (**Adult** and **Credit**) and one image dataset (**CelebA**) that are all for binary classification tasks: ❶ Adult (Dua & Graff, 2017a). The original aim of the dataset Adult is to determine whether a person makes salaries over 50K a

Algorithm 1 Gradient-guided Parity Alignment

```
Input: Network F with parameters $\mathcal{W} = \{w_0, \dots, w_K\}$, epoch index set $\mathcal{E}$, training data $\mathcal{D}$, batch size $B$, network layers $L$, neurons in the $l$th layer $\mathcal{K}_l$, hyper-parameters $\lambda$ and $\beta$, learning rate $\eta$
for $e \in \mathcal{E}$ do
    // Sampling $B$ samples from subgroups in $\mathcal{D}$
    $(\mathbf{X}_0, \mathbf{Y}_0) \leftarrow \text{Sample}(\mathcal{D}, a = 0, B)$
    $(\mathbf{X}_1, \mathbf{Y}_1) \leftarrow \text{Sample}(\mathcal{D}, a = 1, B)$
    // Calculating loss and updating the model
    $\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X}_0), \mathbf{Y}_0) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X}_1), \mathbf{Y}_1)$
    $\mathcal{L}_{\text{fair}} = \Delta\text{DP}(\mathbf{F}, \mathbf{X}_0, \mathbf{X}_1)$
    for $l \in \mathcal{L}$ do
        for $k \in \mathcal{K}_l$ do
            $g_k^{a=0} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_0}), \mathbf{Y_0}))}{\partial w_k}$
            $g_k^{a=1} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_1}), \mathbf{Y_1}))}{\partial w_k}$
            $\hat{c}_k^{a=0} = (g_k^{a=0} \cdot w_k)^2$
            $\hat{c}_k^{a=1} = (g_k^{a=1} \cdot w_k)^2$
        end for
        $\vec{\mathbf{c}}_l^{a=0} = [\hat{c}_0^{a=0}, \hat{c}_1^{a=0}, ..., \hat{c}_{|\mathcal{K}_l|}^{a=0}]$
        $\vec{\mathbf{c}}_l^{a=1} = [\hat{c}_0^{a=1}, \hat{c}_1^{a=1}, ..., \hat{c}_{|\mathcal{K}_l|}^{a=1}]$
    end for
    $\mathcal{L}_{d_F} = -\sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1})$;
    $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{fair} + \beta\mathcal{L}_{d_F}$;
    $w = w - \eta\nabla_w\mathcal{L}, \forall w \in \mathcal{W}$.
end for
```

year. We consider *gender* as the sensitive attribute, and the vanilla training will lead the model to predict females to earn less salaries. ❷ CelebA (Liu et al., 2015). The CelebFaces Attributes dataset is to predict the attributes of face. We split into two subgroups according to the attribute *gender*. Here we consider two attributes classification tasks. For the task to predict whether the hair in an image is *wavy* or not, the standard training will show discrimination towards the male group; when predicting whether the face is *attractive*, the standard training will result in a model prone to predict males as less attractive. ❸ Credit (Dua & Graff, 2017b). This dataset is to give an assessment of credit based on personal and financial records. In our paper, we take the attribute *gender* as the sensitive attribute.

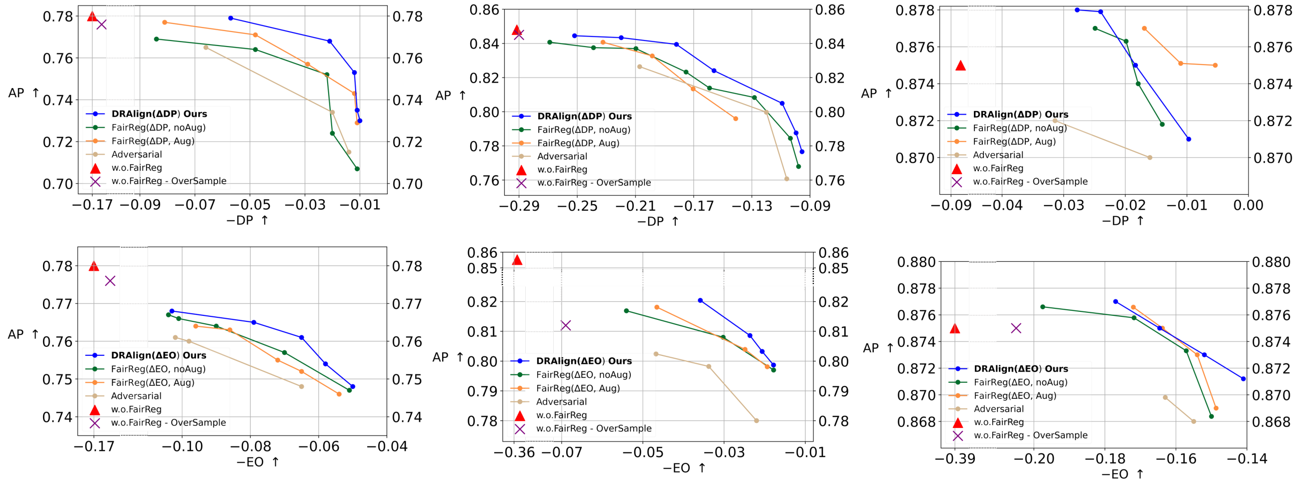
Models. For tabular benchmarks, we use the MLP (multi-layer perception) (Bishop, 1996) as the classification model, which is commonly adopted in classifying tabular data. For the CelebA dataset, we use AlexNet (Krizhevsky et al., 2012) and ResNet-18 (He et al., 2016), both of which are popular in classifying image data (Alom et al., 2018). We mainly show the experimental results of predicting *wavy hair* using AlexNet. More results are in the Appendix (A.3).

Metrics. For fairness evaluation, we take two group fairness metrics DP and EO as we introduced in the Sec. 2.2 and define $-$DP and $-$EO as fairness scores since smaller DP and EO mean better fairness. We use the average precision (AP) for classification accuracy evaluation. A desired fairness method should achieve smaller DP or EO but higher AP (*i.e.*, the top left corner in Fig. 3). Consistent with the previous work (Chuang & Mroueh, 2021; Du et al., 2021), we consider the DP and EO metrics in our work. Moreover, we also explore the Equality of Opportunity (Hardt et al., 2016b) and Predictive Parity (Chouldechova, 2017) . The details are deferred to the Appendix (A.11).

Baselines. Following the common setups in Chuang & Mroueh (2021), we compare our method with several baselines which are shown to be among the most effective and typical methods: ❶ Standard training based on empirical risk minimization (ERM) principle (*i.e.*, w.o.FairReg). DNNs are trained only with the cross entropy loss. ❷ Oversample (*i.e.*, w.o.FairReg-Oversample) (Wang et al., 2020). This method samples from the subgroup with rare examples more often, making a balanced sampling in each epoch. ❸ FairReg(ΔDP or ΔEO, noAug) (Madras et al., 2018). This method is to directly regularize the fairness metrics, *i.e.*, ΔDP or ΔEO. ❹ FairReg(ΔDP or ΔEO, Aug) (*i.e.*, FairMixup) (Chuang & Mroueh, 2021). This method regularizes the models on paths of interpolated samples between subgroups to achieve fairness. ❺ Adversarial (Zhang et al., 2018). This method minimizes the adversary's ability to predict sensitive attributes.

5.2. Fairness Improvement Performance

As shown in Fig. 3, we have following observations: ❶ With the Adult and CelebA datasets, our method (*i.e.*, DRAlign) achieves higher fairness (*i.e.*, higher -DP or -EO scores) than all baseline methods when they have similar AP scores. In particular, on the Adult dataset, DRAlign has relative 41.6% DP improvement over the second best method (*i.e.*, FairReg(ΔDP, Aug)) when both get around 0.770 AP. Overall, our method can enhance the fairness significantly with much less precision sacrifice. ❷ Data augmentation method does not always improve DNN's fairness. For example, on the dataset CelebA, FairReg(ΔDP, noAug) presents slightly higher fairness score (*i.e.*, higher -DP) than FairReg(ΔDP, Aug). A potential reason is that the augmented data becomes less realistic due to the rich information in the image modality, which leads to less effective learning. ❸ Although oversampling could improve fairness to some extent, it is less effective than the fairness regularization-based methods (*i.e.*, FairReg($*$, noAug)). For example, on the CelebA dataset, w.o.FairReg-Oversample only obtains -0.069 -EO score with the 0.812 AP score, while FairReg(ΔEO, noAug) achieves the -0.054 -EO score with 0.817 AP score. The networks trained by FairReg(ΔEO, noAug) are not only fairer but also of higher accuracy. On the tabular dataset,



(a) Adult **(b) CelebA** **(c) Credit**

Figure 3: Comparing different methods on AP vs. (-DP/-EO). According to the common setups, we evaluate ΔDP-based and ΔEO-based methods via -DP and -EO, respectively. The plot is drawn by adjusting the hyperparameter λ and β in Eq. 6 and Eq. 9. The detailed hyperparameter settings are in the Appendix (A.1). We train networks with the compared methods for 10 times and the averaging results are reported. We show our results are statistically significant via t-test (A.10).

w.o.FairReg-Oversample outperforms the w.o.FairReg by a small margin. ❹ On the Credit dataset, FairReg(ΔDP, Aug) achieves better results than DRAlign under the DP metric although our method still outperforms the regularization-based one. A potential reason is that the data size of the Credit is small (*i.e.*, 500 training samples) and the data augmentation can present obvious advantages by enriching the training data significantly. The data augmentation and our decision rationale alignment are two independent ways to enhance fairness. Intuitively, we can combine the two solutions straightforwardly. We did further experiments and found that our DRAlign could still improve FairReg(ΔDP, Aug). In addition, our experiments show that the decision rationale alignment itself could still slightly improve fairness when the fairness regularization item ($\mathcal{L}_{\text{fair}}(\text{F})$) is removed. More details are put in the Appendix (A.8,A.9).

5.3. Discussion and Analysis

Connection with over-parameterization.

To better understand the cause of the decision rationale misalignment, we further investigate the connection between decision rationale misalignment and model over-parameterization. We conduct an empirical study on the Adult dataset using 3-layer MLP networks based on $\text{FairReg}(\Delta\text{DP}, \text{noAug})$. Specifically, we explore 4 MLP architectures, where the hidden sizes are set as 10, 20, 50, and 200, respectively. The corresponding parameter sizes of the 4 networks are 1331, 2861, 8651, and 64601. For each architecture, we draw a plot w.r.t. different λ for $\text{FairReg}(\Delta\text{DP}, \text{noAug})$ to show the decision rationale similarity score (*i.e.*, $\sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1})$ in Sec. 4.2). We denote the four trained models as FairReg(ΔDP,noAug) (c10), FairReg(noAug) (c20), FairReg(ΔDP,noAug) (c50), and FairReg(ΔDP,noAug) (c200), respectively, according to their hidden sizes. *Accuracy performance of these models and more results under ΔEO metric are put in the Appendix (A.6,A.4).* With Fig. 4 (a), we have the following observations: ❶ The decision rationale similarity consistently ascends when λ increases. When λ becomes 0.5, decision rationale similarities of FairReg(ΔDP,noAug) (c10), FairReg(ΔDP,noAug) (c20)and FairReg(ΔDP,noAug) (c50) reach the same maximum score (*i.e.*, 3.0 for any 3-layer MLP network). We conclude that larger λ (stricter fairness constraint) results in a higher decision rationale similarity. ❷ The misalignment of decision rationale is more likely to occur in the over-parameterized networks. For the largest network FairReg(ΔDP,noAug) (c200), even when the λ is set as 0.6 for a strict fairness constraint, the decision rationale similarity score only reaches 1.8 which is much smaller than the values on other variants and infers that the decision rationale is still far from being aligned.

Furthermore, we also report the results of augmentation-based method, *i.e.*, FairReg(ΔDP,Aug). We find that data augmentation can generally mitigate the misalignment but still fails to completely align the decision rationale (See the plot of FairReg(ΔDP,noAug) (c200)). Our method DRAlign is able to achieve the maximum similarity on all λ settings even on the architecture with hidden size 200. (See the plot of DRAlign (c200)) This enlightens us that common methods such as data augmentation that aims to address over-parameterization can not completely solve the misalignment, while our gradient-guided parity alignment method can directly improve the alignment.

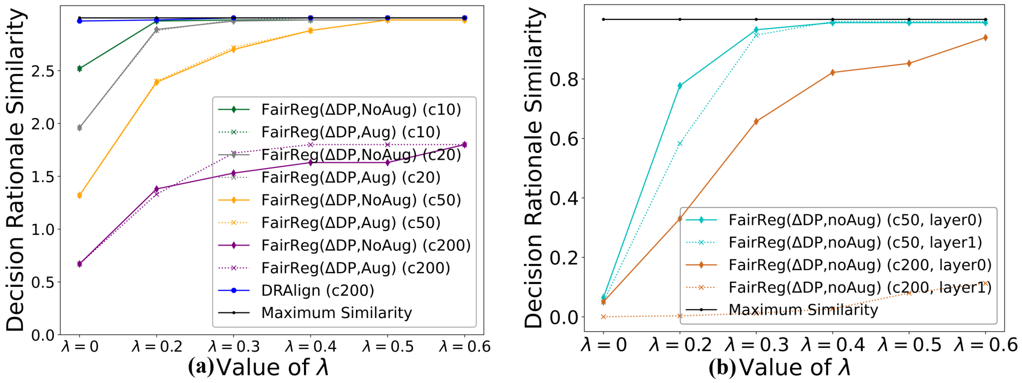


Figure 4: (a) : correlation between λ and decision rationale similarity score. (b) layer-wise analysis for correlation between λ and decision rationale similarity score.

Layer-wise decision rationale alignment analysis.

We further conduct a layer-wise analysis to understand which layer owns better decision rationale alignment. We calculate the decision rationale similarity for the 1st and 2nd layer (*i.e.*, $\cos(\vec{\mathbf{c}}^{a=0}_{l=0}, \vec{\mathbf{c}}^{a=1}_{l=0})$ and $\cos(\vec{\mathbf{c}}^{a=0}_{l=1}, \vec{\mathbf{c}}^{a=1}_{l=1})$). From Fig. 4 (b), we see that: for both layers, the layer-wise similarity score ascends when λ increases. This is consistent with the observation that stricter fairness constraint results in a higher decision rationale similarity. As we compare the 1st and 2nd layers, we can observe that the similarity score of the first layer is generally higher. Moreover, we can see that for smaller models (*i.e.*, models with hidden size 50), the similarity gap between the first layer and the second layer is relatively trivial. However, for models with hidden size 200, the similarity score of the second layer is rather low (*i.e.*, the score is 0.113 even when the λ is 0.6). Thus, the misalignment of the deeper layer is severer.

6. Related Work

Fairness in Deep Learning. Deep learning models, while potent, often display questionable behavior in relation to critical issues such as robustness, privacy, and trustworthiness (Goodfellow et al., 2014; Kurakin et al., 2018; Liu et al., 2019; 2020b;a; 2021; 2023; Guo et al., 2023; Huang et al., 2023). Among these concerns, discrimination stands out as a highly influential factor with considerable social implications. There are different methods to evaluate fairness in deep learning, among which individual fairness (Zhang et al., 2020b; 2021; George John et al., 2020; Xiao et al., 2023), group fairness (Louppe et al., 2016; Moyer et al., 2018; Gupta et al., 2021; Garg et al., 2020), and counterfactual fairness (Kusner et al., 2017) are the mainstream. We focus on group fairness which is derived by calculating and comparing the predictions for each group There is a line of work dedicated to alleviating unjustified bias. For example, Wang et al. (2020) compare mitigation methods including oversampling, adversarial training, and other domain-independent methods. Some work proposes to disentangle unbiased representations to ensure fair DNNs. On the contrary, Du et al. (2021) directly repair the classifier head even though the middle representations are still biased. To improve fairness, it is also popular to constrain the training process by imposing regularization. Woodworth et al. (2017) regularize the covariance between predictions and sensitive attributes. Madras et al. (2018) relax the fairness metrics for optimization. Although such methods are easy to be implemented and integrated into the training process, these constraints suffer from overfitting (Cotter et al., 2019). The model with a large number of parameters could memorize the training data, which causes the fairness constraints to fit well only in the training process. Chuang & Mroueh (2021) ensure better generalization via data augmentation (*e.g.*, mix-up) to reduce the trade-off between fairness and accuracy. However, these methods barely pay attention to the rationale behind the fair decision results. Besides, some studies propose measures for procedural fairness that consider the input features used in the decision process and evaluate the moral judgments of humans regarding the use of these features (Grgic-Hlaca et al., 2016; Grgić-Hlača et al., 2018). They focus on feature selection for procedurally fair learning. In this paper, we further analyze the decision rationales behind the fair decision results in the training process and reveal that ensuring the fair decision rationale could further improve fairness.

Understanding DNNs Decision Rationale. There are some interpretable methods enabling DNNs models to present their behaviors in understandable ways to humans (Zhang & Zhu, 2018; Fong & Vedaldi, 2017; Koh & Liang, 2017; Liang et al., 2019; Zhang et al., 2020a; Li et al., 2021). Specifically, there is a line of work that depicts the decision rationale of DNNs via neuron behaviors analysis. Routing paths composed of the critical nodes (*e.g.* neurons with the most contribution to the final classification on each layer) can be extracted in a learnable way to reflect the network's semantic information flow regarding to a group of data (Khakzar et al., 2021). Conquering the instability existing in the learnable method, Qiu et al. (2019) propose an activation-based back-propagation algorithm to decompose the entire DNN model into multiple components composed of structural neurons. Meanwhile, Xie et al. (2022) base the model function analysis on the neuron contribution calculation and reveal that the neuron contribution patterns of OOD samples and adversarial samples are different from that of normal samples, resulting in wrong classification results. Zheng et al. (2022) analyze neurons sensitive to individual discrimination and generate testing cases according to sensitive neuron behaviors. However, these methods analyze neuron behaviors via static analysis or in a learnable way. These analysis methods result in huge time overhead, making their integration into the training process difficult, which restricts their applications in optimizing the training.

In our paper, we follow the spirit of analyzing neuron behaviors to understand the model decision rationale. Unlike previous methods, our method successfully simplifies the estimation process of neuron contribution and can be easily integrated into the training process to optimize the model.

7. Conclusions and Future Work

In this work, we have studied the fairness issue of deep models from the perspective of decision rationale and defined the *parameter parity score* to characterize the decision rationale shifting across subgroups. We observed that such a decision rationale-aware characterization has a high correlation to the fairness of deep models, which means that a fairer network should have aligned decision rationales across subgroups. To this end, we formulated fairness as the decision rationale alignment (DRAlign) and proposed the *gradient-guided parity alignment* to implement the new task. The results on three public datasets demonstrate the effectiveness and advantages of our methods and show that DRAlign is able to achieve much higher fairness with less precision sacrifice than all existing methods.

Although promising, our method also presents some drawbacks: (1) it requires the computation of second-order derivatives; and (2) the gradient-guided parity alignment method is limited to the layer-wise DNN architecture. In the future, we are interested in solving these limitations.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-022T). It is also supported by A*STAR Centre for Frontier AI Research, the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019), National Satellite of Excellence in Trustworthy Software System No. NRF2018NCR-NSOE003-0001, NRF Investigatorship No. NRF-NRFI06-2020-0001, and the National Natural Science Foundation of China 62206009. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC).

References

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., and Asari, V. K. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., USA, 1996. ISBN 0198538499.

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

Chuang, C.-Y. and Mroueh, Y. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=DNl5s5BXeBn.

Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1397–1405. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cotter19b.html.

Deng, L., Hinton, G., and Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599–8603, 2013. doi: 10.1109/ICASSP.2013.6639344.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Du, M., Liu, N., and Hu, X. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1): 68–77, 2019.

Du, M., Mukherjee, S., Wang, G., Tang, R., Awadallah, A. H., and Hu, X. Fairness via representation neutralization. In *NeurIPS*, 2021.

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

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

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.371. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.371.

Garg, P., Villasenor, J., and Foggo, V. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3662–3666, Los Alamitos, CA, USA, dec 2020. IEEE Computer Society. doi: 10.1109/BigData50022.2020.9378025. URL https://doi.ieeecomputersociety.org/10.1109/BigData50022.2020.9378025.

George John, P., Vijaykeerthy, D., and Saha, D. Verifying individual fairness in machine learning models. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 749–758. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/george-john20a.html.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, pp. 11. Barcelona, Spain, 2016.

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

Guo, J., Bao, W., Wang, J., Ma, Y., Gao, X., Xiao, G., Liu, A., Dong, J., Liu, X., and Wu, W. A comprehensive evaluation framework for deep model robustness. *Pattern Recognition*, 2023.

Gupta, U., Ferber, A., Dilkina, B., and Steeg, G. V. Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation. 2021.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NIPS*, 2016a.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Huang, Y., Cao, Y., Li, T., Juefei-Xu, F., Lin, D., Tsang, I. W., Liu, Y., and Guo, Q. On the robustness of segment anything, 2023.

Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13528–13538, 2021.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1885–1894. JMLR.org, 2017.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/

a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Li, T., Liu, A., Liu, X., Xu, Y., Zhang, C., and Xie, X. Understanding adversarial robustness via critical attacking route. *Information Sciences*, 547:568–578, 2021.

Liang, R., Li, T., Li, L., Wang, J., and Zhang, Q. Knowledge consistency between neural networks and beyond. *arXiv preprint arXiv:1908.01581*, 2019.

Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., and Tao, D. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1028–1035, 2019.

Liu, A., Huang, T., Liu, X., Xu, Y., Ma, Y., Chen, X., Maybank, S., and Tao, D. Spatiotemporal attacks for embodied agents. In *ECCV*, 2020a.

Liu, A., Wang, J., Liu, X., Cao, b., Zhang, C., and Yu, H. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020b.

Liu, A., Liu, X., Yu, H., Zhang, C., Liu, Q., and Tao, D. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*, 2021.

Liu, A., Guo, J., Wang, J., Liang, S., Tao, R., Zhou, W., Liu, C., Liu, X., and Tao, D. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. In *USENIX Security*, 2023.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Louppe, G., Kagan, M., and Cranmer, K. Learning to Pivot with Adversarial Networks. *ArXiv e-prints*, November 2016.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. Learning adversarially fair and transferable representations. In *ICML*, 2018.

Molchanov, P., Tyree, S., Karras, T., Aila, T., and Kautz, J. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016.

Molchanov, P., Mallya, A., Tyree, S., Frosio, I., and Kautz, J. Importance estimation for neural network pruning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11256–11264, 2019. doi: 10.1109/CVPR.2019.01152.

Moyer, D., Gao, S., Brekelmans, R., Steeg, G. V., and Galstyan, A. Invariant representations without adversarial training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9102–9111, Red Hook, NY, USA, 2018. Curran Associates Inc.

News, B. Ibm abandons 'biased' facial recognition tech, 2020. URL https://www.bbc.co.uk/news/technology-52978191.

News, B. Ai at work: Staff 'hired and fired by algorithm', 2021. URL https://www.bbc.com/news/technology-56515827.

Office, B. C. C. Compas recidivism risk score data and analysis. URL https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Qiu, Y., Leng, J., Guo, C., Chen, Q., Li, C., Guo, M., and Zhu, Y. Adversarial defense through network profiling based path extraction. pp. 4772–4781, 06 2019. doi: 10.1109/CVPR.2019.00491.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. Fairness by learning orthogonal disentangled representations. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, pp. 746–761, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58526-6.

Tyler, T. Social justice: Outcome and procedure. *International Journal of Psychology - INT J PSYCHOL*, 35: 117–125, 04 2000. doi: 10.1080/002075900399411.

Tyler, T. R. Procedural justice, legitimacy, and the effective rule of law. *Crime and Justice*, 30:283–357, 2003. ISSN 01923234. URL http://www.jstor.org/stable/1147701.

Wang, G., Du, M., Liu, N., Zou, N., and Hu, X. Mitigating algorithmic bias with limited annotations. *arXiv preprint arXiv:2207.10018*, 2022.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., and Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Wang, Y., Su, H., Zhang, B., and Hu, X. Interpret neural networks by identifying critical data routing paths. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8906–8914, 2018a. doi: 10.1109/CVPR.2018.00928.

Wang, Y., Su, H., Zhang, B., and Hu, X. Interpret neural networks by identifying critical data routing paths. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8906–8914, 2018b.

Wang, Z., Qinami, K., Karakozis, I., Genova, K., Nair, P., Hata, K., and Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8916–8925, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00894. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00894.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning non-discriminatory predictors. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1920–1953. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/woodworth17a.html.

Xiao, Y., Liu, A., Li, T., and Liu, X. Latent imitator: Generating natural individual discriminatory instances for black-box fairness testing. In *ISSTA*, 2023.

Xie, X., Li, T., Wang, J., Ma, L., Guo, Q., Juefei-Xu, F., and Liu, Y. Npc: Neuron path coverage via characterizing decision logic of deep neural networks. *ACM Trans. Softw. Eng. Methodol.*, 31(3), apr 2022. ISSN 1049-331X. doi: 10.1145/3490489. URL https://doi.org/10.1145/3490489.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

Zhang, C., Liu, A., Liu, X., Xu, Y., Yu, H., Ma, Y., and Li, T. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, 30:1291–1304, 2020a.

Zhang, L., Zhang, Y., and Zhang, M. Efficient white-box fairness testing through gradient search. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2021, pp. 103–114, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384599. doi: 10.1145/3460319.3464820. URL https://doi.org/10.1145/3460319.3464820.

Zhang, P., Wang, J., Sun, J., Dong, G., Wang, X., Wang, X., Dong, J. S., and Dai, T. White-box fairness testing through adversarial sampling. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, pp. 949–960, New York, NY, USA, 2020b. Association for Computing Machinery. ISBN 9781450371216. doi: 10.1145/3377811.3380331. URL https://doi.org/10.1145/3377811.3380331.

Zhang, P., Wang, J., Sun, J., Dong, G., Wang, X., Wang, X., Dong, J. S., and Dai, T. White-box fairness testing through adversarial sampling. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, pp. 949–960, New York, NY, USA, 2020c. Association for Computing Machinery. ISBN 9781450371216. doi: 10.1145/3377811.3380331. URL https://doi.org/10.1145/3377811.3380331.

Zhang, Q. and Zhu, S. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology and Electronic Engineering*, 19, 02 2018. doi: 10.1631/FITEE.1700808.

Zheng, H., Chen, Z., Du, T., Zhang, X., Cheng, Y., Ti, S., Wang, J., Yu, Y., and Chen, J. Neuronfair: Interpretable white-box fairness testing through biased neuron identification. In *2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE)*, pp. 1519–1531, 2022. doi: 10.1145/3510003.3510123.

A. Appendix

A.1. Training Details

Adult Dataset. The parameter setting of the Adult Dataset is shown in Table 2. We follow the settings in Chuang & Mroueh (2021) for data preprocessing. The hidden size of MLP is 200. We use Adam as the learning optimizer and the batch size is set as 1000 for the DP metric and 2000 for the EO metric following the setting in Chuang & Mroueh (2021).

Table 2: Setting for Adult Dataset training with MLP.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*, noAug)	FairReg(*, Aug)	DRAlign
Training Epochs for DP	20	20	20	20	20
Training Epochs for EO	20	20	20	20	20
Learning rate	0.001	0.001	0.001	0.001	0.001
Range of λ for DP	-	-	[0.2,0.3,0.4,0.5,0.6]	[0.2,0.3,0.4,0.5,0.6]	[0.1,0.2,0.3,0.4,0.5]
β for DP	-	-	-	-	$\lambda/10$
Range of λ for EO	-	-	[0.5~2.0]	[0.5~2.0]	[0.5~2.0]
β for EO	-	-	-	-	$\lambda/10$

CelebA Dataset. The parameter setting of the CelebA Dataset is shown in Table 3. We follow the settings in (Chuang & Mroueh, 2021) for data preprocessing. We use Adam as the learning optimizer and the batch size is set as 64 for the DP metric and 128 for the EO metric following the setting in Chuang & Mroueh (2021).

Table 3: Setting for CelebA Dataset training with AlexNet.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*, noAug)	FairReg(*, Aug)	DRAlign
Training Epochs for DP	15	15	15	30	15
Training Epochs for EO	30	30	30	60	30
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001
Range of λ for DP	-	-	[0.1~0.7]	[0.1~0.6]	[0.1~0.7]
β for DP	-	-	-	-	0.01
Range of λ for EO	-	-	[0.1,0.4,0.7]	[0.1,0.4,0.7]	[0.1,0.4,0.7,1.0]
β for EO	-	-	-	-	0.01

Credit Dataset. The parameter setting of the Credit Dataset is shown in Table 4. We follow the settings in (Zhang et al., 2020c) for data preprocessing. We use Adam as the learning optimizer and the batch size is set as 400 for the DP metric and 500 for the EO metric.

Table 4: Setting for Credit Dataset.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*, noAug)	FairReg(*, Aug)	DRAlign
Training Epochs for DP	20	20	20	20	20
Training Epochs for EO	20	20	20	20	20
Learning rate	0.001	0.001	0.001	0.001	0.001
Range of λ for DP	-	-	[0.2,0.8,1.0,2.0]	[0.2,0.8,2.0]	[0.8,1.0,2.0,3.0]
β for DP	-	-	-	-	0.005
Range of λ for EO	-	-	[0.2,0.4,0.6,0.8]	[0.2,0.4,0.8,1.0]	[0.6, 0.8,1.0,2.0]
β for EO	-	-	-	-	0.01

Algorithm 2 Gradient-guided Parity Alignment for The EO Metric

```
Input: Network F with parameters $\mathcal{W} = \{w_0, \dots, w_K\}$, epoch index set $\mathcal{E}$, training data $\mathcal{D}$, batch size $B$, network layers $L$, neurons in the $l$th layer $\mathcal{K}_l$, hyper-parameters $\lambda$ and $\beta$, learning rate $\eta$
for $e \in \mathcal{E}$ do
  // Sampling $B$ samples from subgroups in $\mathcal{D}$
  $[\mathbf{X_{00}}, \mathbf{Y_{00}}]$ = Sample($D$, a=0, y=0, $B$);
  $[\mathbf{X_{01}}, \mathbf{Y_{01}}]$ = Sample($D$, a=0, y=1, $B$);
  $[\mathbf{X_{10}}, \mathbf{Y_{10}}]$ = Sample($D$, a=1, y=0, $B$);
  $[\mathbf{X_{11}}, \mathbf{Y_{11}}]$ = Sample($D$, a=1, y=1, $B$);
  // Calculating loss and updating the model
  $\mathcal{L}_c = \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{00}}), \mathbf{Y_{00}})) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{01}}), \mathbf{Y_{01}}) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{10}}), \mathbf{Y_{10}})) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{11}}), \mathbf{Y_{11}});$
  $\mathcal{L}_{\text{fair}} = \Delta\text{EO}(\mathbf{F}, \mathbf{X}_{00}, \mathbf{X}_{01}, \mathbf{X}_{10}, \mathbf{X}_{11})$
  for $l \in \mathcal{L}$ do
    for $k \in \mathcal{K}_l$ do
      $g_k^{a=0,y=0} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{00}}),\mathbf{Y_{00}}))}{\partial w_k};$
      $g_k^{a=1,y=0} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{10}}),\mathbf{Y_{10}}))}{\partial w_k};$
      $g_k^{a=0,y=1} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{01}}),\mathbf{Y_{01}}))}{\partial w_k};$
      $g_k^{a=1,y=1} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{X_{11}}),\mathbf{Y_{11}}))}{\partial w_k};$
      $\hat{c}_k^{a=0,y=0} = (g_k^{a=0,y=0} \cdot w_k)^2;$
      $\hat{c}_k^{a=1,y=0} = (g_k^{a=1,y=0} \cdot w_k)^2;$
      $\hat{c}_k^{a=0,y=1} = (g_k^{a=0,y=1} \cdot w_k)^2;$
      $\hat{c}_k^{a=1,y=1} = (g_k^{a=1,y=1} \cdot w_k)^2;$
    end for
    $\vec{\mathbf{c}}_l^{a=0,y=0} = [\hat{c}_0^{a=0,y=0}, \hat{c}_1^{a=0,y=0}, ..., \hat{c}_{\mathcal{K}_l}^{a=0,y=0}];$
    $\vec{\mathbf{c}}_l^{a=1,y=0} = [\hat{c}_0^{a=1,y=0}, \hat{c}_1^{a=1,y=0}, ..., \hat{c}_{\mathcal{K}_l}^{a=1,y=0}];$
    $\vec{\mathbf{c}}_l^{a=0,y=1} = [\hat{c}_0^{a=0,y=1}, \hat{c}_1^{a=0,y=1}, ..., \hat{c}_{\mathcal{K}_l}^{a=0,y=1}];$
    $\vec{\mathbf{c}}_l^{a=1,y=1} = [\hat{c}_0^{a=1,y=1}, \hat{c}_1^{a=1,y=1}, ..., \hat{c}_{\mathcal{K}_l}^{a=1,y=1}];$
    $\mathcal{L}_{d_F} = -\sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0,y=0}, \vec{\mathbf{c}}_l^{a=1,y=0}) + \sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0,y=1}, \vec{\mathbf{c}}_l^{a=1,y=1});$
    $\mathcal{L} = \mathcal{L}_c + \lambda\mathcal{L}_{fair} + \beta\mathcal{L}_{d_F};$
    $w \leftarrow w - \eta\nabla_w\mathcal{L}, \forall w \in \mathcal{W}.$
  end for
end for
```

In our paper, we did a rough search for the hyper-parameter β. Taking CelebA dataset as an example, we mainly search β value in the range 0.001, 0.01, 0.1. When β is set as 0.001, the training process is close to that of FairReg, which means that our decision rationale alignment item is ignored in the training because β is too small. When β is 0.1, the training process will optimize the decision rationale alignment first and cause a detrimental influence on the optimization of other loss items. We finally choose 0.01 as the β value. We have found that the training process is stable with a proper β setting. We train the models for the CelebA dataset from scratch.

In our paper, we mainly consider the parameters of multiplicative weights (neurons/filters) in the convolution layers and the linear layers, because we focus on the decision rationale which could be defined by the influence of neuron/filter/parameter that is usually regarded as a semantic unit. We did not consider the biases parameters and the parameters in BN layers.

A.2. Algorithm of DRAlign When Training under The EO Metric

The training algorithm for EO metric is shown in Algorithm 2.

A.3. More Experimental Results

A.3.1. Classification for Attractive Attribute

In our paper, on the CelebA dataset, we show the results of predicting *wavy hair* attribute. Here, we also show the results of classifying *attractive* attribute adopting AlexNet. For better observation, we show our results in Table 5. We find that our method outperforms FairReg(noAug) both in AP and in the fairness metric.

Table 5: Comparison between DRAlign(ours) and FairReg(*, noAug) when classifying *attractive* attribute.

	−DP			−EO		
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.1$	$\lambda = 0.4$	$\lambda = 1.0$
AP$_{DRAlign}$	**0.8956**	**0.8895**	0.8783	**0.8760**	**0.8735**	**0.8717**
Fairness$_{DRAlign}$	**-0.3196**	**-0.2727**	**-0.2126**	**-0.0520**	**-0.0337**	**-0.0243**
AP$_{FairReg(*,noAug)}$	0.8942	0.8873	**0.8807**	0.8733	0.8707	0.8681
Fairness$_{FairReg(*,noAug)}$	−0.3305	−0.2819	−0.2377	−0.0533	−0.0374	−0.0273

A.3.2. Classification for Wavy Hair Based on ResNet-18

In our algorithm, we expect to reduce the parity score for all layers. However, for some larger architectures such as ResNet-18, it is relatively difficult to optimize all layers. To address such a problem, we here only align the last two layers. We find that only aligning the last two layers could also improve fairness. The loss function is revised as follows:

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\text{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda\mathcal{L}_{\text{fair}}(\mathrm{F}) - \beta \sum_{l=L-1}^{L} \cos(\vec{\mathbf{c}}_l^{a=0}, \vec{\mathbf{c}}_l^{a=1}), \tag{10}$$

The experimental results are shown in Table 6.

Table 6: Comparison between DRAlign(ours) and FairReg(*, noAug) when classifying *Wavy hair* attribute using ResNet-18.

	−DP			−EO		
	$\lambda = 0.1$	$\lambda = 5.0$	$\lambda = 10.0$	$\lambda = 0.2$	$\lambda = 5.0$	$\lambda = 10.0$
AP$_{DRAlign}$	**0.8578**	**0.8385**	**0.8179**	**0.8212**	**0.7965**	**0.7703**
Fairness$_{DRAlign}$	**-0.3011**	**-0.2723**	**-0.2481**	**-0.1294**	**-0.0495**	**-0.0446**
AP$_{FairReg(*,noAug)}$	0.8506	0.8355	0.8123	0.8063	0.7857	0.7560
Fairness$_{FairReg(*,noAug)}$	−0.3063	−0.2795	−0.2552	−0.1832	−0.0498	−0.0494

A.4. Connection With Over-parameterization under EO Metric

We here analyze the connection between decision rationale alignment and over-parameterization under EO metric. We show the results on the Adult dataset adopting 3-layer MLP models. The maximum alignment score is 6.0. Here we also conclude that over-parameterization might prevent the alignment of decision rationale and stricter fairness regularizations require fairer decision rationale.

A.5. Training Time Estimation

We here show the time consumption of different methods on the Adult dataset, CelebA dataset, and Credit dataset in Table 8, Table 9 and Table 10 respectively. Please be noted that the FairReg(*,Aug) method also requires the calculation of a second-order derivative. Moreover, as a method based on data augmentation, the FairReg(*,Aug) method requires more time to converge.

Table 7: Connection between decision rationale similarity and over-parameterization under EO metric.

	$\lambda = 0.5$	$\lambda = 0.8$	$\lambda = 1.0$	$\lambda = 2.0$	$\lambda = 3.0$
FairReg(ΔEO,noAug), (c10)	6.0	6.0	6.0	6.0	6.0
FairReg(ΔEO,noAug), (c20)	5.7	5.7	5.8	6.0	6.0
FairReg(ΔEO,noAug), (c50)	5.6	5.7	5.8	6.0	6.0
FairReg(ΔEO,noAug), (c200)	5.6	5.7	5.8	5.9	6.0
FairReg(ΔEO,Aug), (c10)	6.0	6.0	6.0	6.0	6.0
FairReg(ΔEO,Aug), (c20)	6.0	6.0	6.0	6.0	6.0
FairReg(ΔEO,Aug), (c50)	5.9	6.0	6.0	6.0	6.0
FairReg(ΔEO,Aug), (c200)	5.7	5.9	6.0	6.0	6.0

Table 8: Training time estimation when training with Adult dataset under the DP and EO metric.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*,noAug)	FairReg(*,Aug)	DRAlign
DP	8.2s	10.1s	10.5s	14.7s	14.5s
EO	12.5s	14.6s	15.0s	33.2s	30.1s

A.6. The AP Values of Various Model Architectures

Table 11 show the AP values of different model architectures. The model is chosen according to the performance on the validation dataset. We can see that larger models are prone to have higher APs.

A.7. Connection With Human Society

Our main idea is similar to human society where people are not only focusing on the *outcome justice* (Tyler, 2000) (e.g., fairness in the decision results) but pay increasing attention to the *procedural justice* (Tyler, 2003) (e.g., fairness in the decision rationale). The regularization method to improve fairness can be deemed as achieving the *outcome justice* directly. Our experiments/analysis show that *procedural justice* might be easily violated in DNN models. We propose decision rationale alignment to further achieve the *procedural justice* and improve fairness.

A.8. Combination With Data Augmentation

The data augmentation and our decision rationale alignment are two independent ways to enhance fairness. From Fig. 3 (main paper), we can see that on the Credit dataset, FairReg(ΔDP, Aug) achieves better results than DRAlign under the DP metric. Intuitively, we can combine the two solutions straightforwardly. For example, we can replace the second term in Eq.(6) (main paper) (i.e., L_{fair}) with the data augmentation-embedded term (See (Chuang & Mroueh, 2021) for more details) and have a new formulation of Eq.(6) (main paper).

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\text{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda \mathcal{L}_{\text{aug}}(\mathrm{F}) + \beta \sum_{k=0}^{K} d_k, \tag{11}$$

We denote the above method for DP regularization as $\text{DRAlign}(\Delta\text{DP}, \text{Aug})$. We evaluate this version and compare it to the method without augmentation (i.e., DRAlign(ΔDP) on the Credit dataset. We see that: the fairness score (i.e., -DP) increases from -0.0169 to -0.0155 while the average precision (AP) also increases from 0.877 to 0.881, which further demonstrates the scalability of our method.

A.9. Decision Rationale Alignment Without The Fairness Regularization.

Table 9: Training time estimation when training with CelebA dataset and AlexNet under the DP and EO metric.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*,noAug)	FairReg(*,Aug)	DRAlign
DP	611.3s	725.2s	811.6s	1995.3s	1397.8s
EO	661.8s	761.8s	865.8s	3640.8s	3278.2s

Table 10: Training time estimation when training with Credit dataset under the DP and EO metric.

	w.o.FairReg	w.o.FairReg - OverSample	FairReg(*,noAug)	FairReg(*,Aug)	DRAlign
DP	6.1s	8.6s	8.7s	12.5s	12.1s
EO	8.5s	10.7s	11.1s	13.5s	13.0s

We find that the alignment itself could still slightly improve fairness when fairness regularization is removed. Specifically, we remove the L_{fair} term in Eq.(6) (main paper) and retain the classification loss and the decision rationale alignment loss and compare the results of the two loss functions $L = L_{cls}$ and $L = L_{cls} + L_{DRA}$. We denote this version as w.o.FairReg-DRAlign. From Fig. 5 we can see that: compared with the model only trained with the classification loss (i.e., w.o.FairReg, w.o.FairReg - Oversample), w.o.FairReg-DRAlign increases the experimental results (AP, -DP) from (0.776, -0.16) to (0.781, -0.14). The results are consistent with our observation that our decision rationale alignment method could further improve fairness and demonstrate that decision rationale alignment is actually a favorable supplement for existing fairness regularization terms.

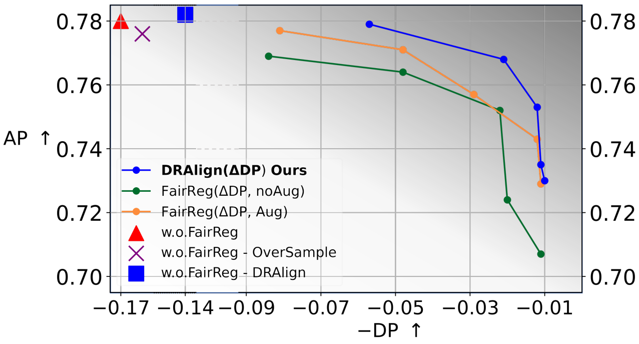


Figure 5: Accuracy and fairness comparison of five different methods on the Adult dataset. The hyperparameter λ increases from 0.2 to 0.6 along the $-$DP axis as it becomes larger.

A.10. The T-Test Results to Show The Significant Fairness Improvement

We showcase our results are statistically significant via t-test (regarding FairReg(ΔDP,noAug) and DRAlign under the DP metric on the Adult dataset and the CelebA dataset in Table 12 and Table 13. We can see that under all parameter settings, the p-values < 0.05. This means that our results are statistically significant. Note that, our method not only improves the model performance in fairness but also improves the model accuracy. Taking both AP values and fairness performance into consideration, our DRAlign outperforms other methods saliently.

A.11. Experiments on More Fairness Measures

we further evaluate and compare our method with FairReg methods on the third popular metric Equality of Opportunity (EOP), and Predictive Parity (PP). Specifically, we adopt the EOP definition in (Wang et al., 2022), and the PP definition in (Garg et al., 2020).

$$EOP = TPR_{a=0}/TPR_{a=1} = P(\hat{y}=1|a=0,y=1)/P(\hat{y}=1|a=1,y=1)$$

$$PP = |p(y=1|a=0,\hat{y}=1) - p(y=1|a=1,\hat{y}=1)|.$$

Under the above definition, EOP close to 1 and PP close to 0 indicate fair classification results. We carefully modify the FairReg method for the EOP and PP metrics.

$$L_{fair,EOP} = \Delta\text{EOP}(F) = E_{\mathbf{x}\sim P_0^1}(F(\mathbf{x})) - E_{\mathbf{x}\sim P_1^1}(F(\mathbf{x})),$$

Table 11: The AP Values of Different Model Architectures.

	$\lambda = 0$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$
c_{10}	0.781	0.780	0.776	0.768	0.758	0.745	0.731	0.729
c_{20}	0.782	0.780	0.777	0.768	0.757	0.743	0.734	0.728
c_{50}	0.783	0.781	0.776	0.769	0.758	0.741	0.737	0.730
c_{200}	0.784	0.781	0.777	0.769	0.760	0.744	0.744	0.738

Table 12: T-test results for FairReg(ΔDP,noAug) and our DRAlign method on the Adult Dataset.

	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$
P Value (Adult)	1.19e-05	3.74e-05	0.011	0.027

$$L_{fair,PP} = \Delta\text{PP}(F) = \frac{E_{\mathbf{x}\sim P_0^0}(F(\mathbf{x})) * N_0^0) + (E_{\mathbf{x}\sim P_0^1}(F(\mathbf{x})) * N_0^1)}{N_0^0 + N_0^1} - \frac{E_{\mathbf{x}\sim P_1^0}(F(\mathbf{x})) * N_1^0) + (E_{\mathbf{x}\sim P_1^1}(F(\mathbf{x})) * N_1^1)}{N_1^0 + N_1^1},$$

N_0^0, N_0^1, N_1^0, N_1^1 are the sample numbers of subgroups D_{00}, D_{01}, D_{10}, and D_{11}, which satisfy the following attribute and category conditions: a=0,y=0, a=0,y=1,a=1,y=0,a=1,y=1 in the batch of data. The sampling methods of FairReg(Δ EOP, noAug) and FairReg(Δ PP, noAug) follow those of FairReg(Δ EO, noAug). For the EOP metric, we align the decision rationales between subgroups a=0,y=1 and a=1,y=1. For the PP metric, we align the decision rationales of (D_{00}, D_{10}) and the decision rationales of (D_{01},D_{11}).

We showcase the experimental results here (the report averages over 10 times). The models evaluated on the EOP metrics are trained for 20 epochs. λ is set as in range {0.03,0.1,0.5} and β is set as $\lambda/10$. The models evaluated on the EOP metrics are trained for 5 epochs. λ is set as in range {0.07,0.09,0.1}. From the Table 14, we can see that our method DRAlign consistently improves the fairness performance under the EOP and PP metrics, that is, our method could be extended to EOP and PP.

A.12. Experiments on the COMPAS dataset

We further evaluate another widely-used dataset, i.e., COMPAS (Office). We train all the models in 5 epochs. The learning rate is set as 0.001. For our method DRAlign, is set as $\beta/10$. From Table 15, we see that: our method consistently outperforms all baseline methods, which is in line with the conclusion of our paper. For instance, when we set λ as 0.1, DRAlign(ΔDP) achieves higher AP (0.637) and -DP (-0.033) scores than the baseline methods. We will include the results in our paper.

A.13. Independent assumption across neurons

The assumption of neuron independence is used to estimate the influence/importance of a group of neurons by calculating the influence/importance of each individual neuron. In our work, the essential objective of decision rationale alignment is to ensure that eliminating any random combination of neurons has the same influence on various subgroups. However, evaluating the importance of all potential neuron combinations is impractical and computationally challenging due to the vast search space for possible combinations of neurons. Therefore, we assume that neurons are independent, which enables us to estimate the influence of any neuron combination by summing the individual impacts of each neuron in the combination. This assumption can be viewed as a "greedy" approximation strategy to assess the influence of arbitrary neuron combinations (Molchanov et al., 2016; 2019). Note that the independence assumption doesn't limit our method to networks with all independent neurons. Instead, it is an approximation to align the influence of any neuron combination in regular networks.

Note that, the neuron independence assumption has been widely used in previous works (Molchanov et al., 2016; 2019) to achieve the approximation to evaluate the influence of a group of neurons.

Table 13: T-test results for FairReg(ΔDP,noAug) and our DRAlign method on the CelebA dataset.

	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$
P Value (CelebA)	0.017	0.018	0.003	0.002

Table 14: Comparison between DRAlign(ours) and FairReg(*, noAug) under the EOP metric and the PP metric.

	Equality of Opportunity				Predictive Parity		
$AP_{DRAlign}$	**0.7787**	**0.7737**	**0.7597**	$AP_{DRAlign}$	**0.7854**	**0.7852**	0.7839
$EOP_{DRAlign}$	**0.9494**	**0.9510**	**0.9692**	$PP_{DRAlign}$	**0.0219**	**0.0186**	**0.0175**
$AP_{FairReg(\Delta EOP,noAug)}$	0.7769	0.7695	0.7552	$AP_{FairReg(\Delta PP,noAug)}$	0.7852	0.7845	**0.7840**
$EOP_{FairReg(\Delta EOP,noAug)}$	0.9411	0.9456	0.9606	$PP_{FairReg(\Delta PP,noAug)}$	0.0276	0.0246	0.0205

A.14. Comparing pre-processing and post-processing techniques

We further add experiments to compare our method with a pre-processing method which adjusts the inputs to be uncorrelated with the sensitive attribute in each iteration. This method achieves a -DP value of -0.086 and an AP value of 0.761. We test the parity score of the trained model and find it to be 0.43, which is smaller than the parity score of 0.624 obtained through standard training. This observation is consistent with our conclusion drawn using the FairReg(*,noAug) method and shows that the pre-processing method could also implicitly lead to lower neuron/parameter parity scores.

However, it should be noted that post-processing techniques only modify the outputs of the trained model and do not alter the model itself. Therefore, the post-processing methods could be regarded as aligning the neuron/parameter of the last layer while neglecting the alignment of the middle neurons/parameters. Thus, with these observations, we can regard the misalignment as a fundamental reason for unfairness.

Table 15: Experimental results on the COMPAS dataset.

Method					Method				
Van	AP	0.649			Van	AP	0.643		
	-DP	-0.245				-EO	-0.442		
Oversampling	AP	0.636			Oversampling	AP	0.65		
	-DP	-0.189				-EO	-0.102		
FairReg(DP,noAug)	**lam**	**0.05**	**0.1**	**0.3**	FariReg(EO,noAug)	**lam**	**0.02**	**0.03**	**0.05**
	AP	0.637	0.635	0.634		AP	0.647	0.646	0.646
	-DP	-0.117	-0.055	-0.015		-EO	-0.045	-0.044	-0.034
FairReg(DP,Aug)	**lam**	**0.02**	**0.04**	**0.05**	FairReg(EO, Aug)	**lam**	**0.02**	**0.03**	**0.05**
	AP	0.638	0.637	0.636		AP	0.648	0.647	0.644
	-EO	-0.145	-0.122	-0.105		-EO	0.07	0.06	0.032
DRAlign	**lam**	**0.05**	**0.1**	**0.3**	DRAlign	**lam**	**0.02**	**0.03**	**0.05**
	AP	0.639	0.637	0.635		AP	0.649	0.647	0.646
	-DP	-0.104	-0.033	-0.008		-EO	-0.044	-0.039	-0.032