A SUPPLEMENTRAY MATERIALS

A.1 TRAINING DETAILS

Adult Dataset. The parameter setting of Adult Dataset is shown in table 1. 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 1: 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	-	-	-	-	[0.01,0.02,0.03,0.04,0.05]
Range of λ for EO	-	-	[0.5,0.8,1.0,2.0]	[0.5,0.8,1.0,2.0]	[0.5,0.8,1.0,2.0]
β for EO	-	-	-	-	[0.05,0.08,0.1,0.2]

CelebA Dataset. The parameter setting of CelebA Dataset is shown in table 2. 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 2: 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	15	15
Training Epochs for EO	30	30	30	30	30
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001
Range of λ for DP	-	-	[0.2,0.3,0.4,0.5,0.6]	[0.2,0.3,0.4,0.5,]	[0.2,0.3,0.4,0.5,0.6]
β for DP	-	-	-	-	0.01
Range of λ for EO	-	-	[0.1,0.4,0.7,1.0]	[0.1,0.4,0.7,1.0]	[0.1,0.4,0.7,1.0]
β for EO	-	-	-	-	0.01

Credit Dataset. The parameter setting of Credit Dataset is shown in table 3. We follow the settings in Zhang et al. (2020) 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 3: 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]	[0.2,0.4,0.8,2.0]	[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]	[0.8,1.0,2.0]
β for EO	-	-	-	-	0.01

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.

Algorithm 1: Gradient-guided Parity Alignment

Data: Network F with parameters $\mathcal{W} = \{w_0, \ldots, w_K\}$, epoch index set $\mathcal{E}$, training data $\mathcal{D}$, batch size B, network layers L, neurons in the lth layer $\mathcal{K}_l$,hyper-parameters λ and β, learning rate η.

// Training process for EO

for $e \in E$ **do**

  for $i \in I$ **do**

    // Sample data subgroups from D

    $[\mathbf{x_{00}}, \mathbf{y_{00}}] \leftarrow$ Sample(D, a=0, y=0, bs);

    $[\mathbf{x_{01}}, \mathbf{y_{01}}] \leftarrow$ Sample(D, a=0, y=1, bs);

    $[\mathbf{x_{10}}, \mathbf{y_{10}}] \leftarrow$ Sample(D, a=1, y=0, bs);

    $[\mathbf{x_{11}}, \mathbf{y_{11}}] \leftarrow$ Sample(D, a=1, y=1, bs);

    // Update the model

    $\mathcal{L}_c \leftarrow \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{00}}), \mathbf{y_{00}})) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{01}}), \mathbf{y_{01}}) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{10}}), \mathbf{y_{10}})) + \mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{11}}), \mathbf{y_{11}})$;

    $\mathcal{L}_{fair} \leftarrow |\text{mean}(F(\mathbf{x_{00}})) - \text{mean}(F(\mathbf{x_{10}}))| + |\text{mean}(F(\mathbf{x_{01}})) - \text{mean}(F(\mathbf{x_{11}}))|$;

    for $l \in L$ **do**

      for $k \in \mathcal{K}_l$ **do**

        $g_k^{a=0,y=0} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{00}}), \mathbf{y_{00}}))}{\partial w_k}$; $g_k^{a=1,y=0} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{10}}), \mathbf{y_{10}}))}{\partial w_k}$;

        $g_k^{a=0,y=1} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{01}}), \mathbf{y_{01}}))}{\partial w_k}$; $g_k^{a=1,y=1} = \frac{\partial(\mathcal{L}_{\text{cls}}(\text{F}(\mathbf{x_{11}}), \mathbf{y_{11}}))}{\partial w_k}$;

        $\hat{c}_k^{a=0,y=0} \leftarrow (g_k^{a=0,y=0} \cdot w_k)^2$; $\hat{c}_k^{a=1,y=0} \leftarrow (g_k^{a=1,y=0} \cdot w_k)^2$;

        $\hat{c}_k^{a=0,y=1} \leftarrow (g_k^{a=0,y=1} \cdot w_k)^2$; $\hat{c}_k^{a=1,y=1} \leftarrow (g_k^{a=1,y=1} \cdot w_k)^2$;

      $\vec{\mathbf{c}}_l^{a=0,y=0} = [\hat{c}_0^{a=0,y=0}, \hat{c}_1^{a=0,y=0}, ..., \hat{c}_{\mathcal{K}_l}^{a=0,y=0}]$;

      $\vec{\mathbf{c}}_l^{a=1,y=0} = [\hat{c}_0^{a=1,y=0}, \hat{c}_1^{a=1,y=0}, ..., \hat{c}_{\mathcal{K}_l}^{a=1,y=0}]$;

      $\vec{\mathbf{c}}_l^{a=0,y=1} = [\hat{c}_0^{a=0,y=1}, \hat{c}_1^{a=0,y=1}, ..., \hat{c}_{\mathcal{K}_l}^{a=0,y=1}]$;

      $\vec{\mathbf{c}}_l^{a=1,y=1} = [\hat{c}_0^{a=1,y=1}, \hat{c}_1^{a=1,y=1}, ..., \hat{c}_{\mathcal{K}_l}^{a=1,y=1}]$;

    $\mathcal{L}_{d_F} \leftarrow \sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0,y=0}, \vec{\mathbf{c}}_l^{a=1,y=0}) + \sum_{l=0}^{L} \cos(\vec{\mathbf{c}}_l^{a=0,y=1}, \vec{\mathbf{c}}_l^{a=1,y=1})$;

    $\mathcal{L} \leftarrow \mathcal{L}_c + \lambda \mathcal{L}_{fair} - \beta \mathcal{L}_{d_F}$;

    $\theta \leftarrow \theta - \eta \nabla_\Theta \mathcal{L}$

A.2 Algorithm of DRAlign when training with EO metric

The training algorithm for EO metric is shown in Algorithm 1.

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 4. We find that our method outperforms FairReg(noAug) both in AP and in the fairness metric.

Table 4: Comparison between DRAlign(ours) and FairReg(*, noAug) when classifying *attractive* attribute.

	−DP			−EO		
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.1$	$\lambda = 0.4$	$\lambda = 1.0$
$\text{AP}_{DRAlign}$	**0.8956**	**0.8901**	**0.8783**	**0.8742**	**0.8735**	**0.8724**
$\text{Fairness}_{DRAlign}$	**-0.320**	**-0.281**	**-0.213**	**-0.0526**	**-0.0337**	−0.027
$\text{AP}_{FairReg(*,noAug)}$	0.8942	0.8733	0.8807	0.8733	0.8707	0.8690
$\text{Fairness}_{FairReg(*,noAug)}$	−0.331	−0.282	−0.238	−0.053	−0.037	**-0.024**

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:

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

The experimental results are shown in table 5.

Table 5: Comparison between DRAlign(ours) and FairReg(*, noAug) when classifying *Wavy hair* attribute using ResNet-18.

	−DP			−EO		
	$\lambda = 0.1$	$\lambda = 5.0$	$\lambda = 10.0$	$\lambda = 20.0$	$\lambda = 5.0$	$\lambda = 10.0$
$\text{AP}_{DRAlign}$	**0.8578**	**0.8385**	**0.8179**	**0.8212**	**0.7965**	**0.7703**
$\text{Fairness}_{DRAlign}$	**-0.301**	**-0.272**	**-0.248**	**-0.129**	**-0.0495**	**-0.0446**
$\text{AP}_{FairReg(*,noAug)}$	0.8506	0.8355	0.8123	0.8063	0.7857	0.7560
$\text{Fairness}_{FairReg(*,noAug)}$	−0.306	−0.279	−0.255	−0.183	−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.

Table 6: 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	6.0	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

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 7, table 8 and table 9 respectively.

Table 7: 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 Different Model Architectures.

Tables 10 show the AP values of different model architectures. The model is chosen according to the performance on the validation dataset.

Table 8: 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 9: 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

A.7 VARIOUS FAIRNESS METRICS.

In our paper, we mainly focus on the metric demographic parity (DP) and the equalized odds (EO), both of which are introduced detailedly in section 3 (main paper). Our method is also applicable to other fairness metrics that quantify the expected difference between groups. For example, predictive parity focuses on whether the positive predictive value (PPV) is the same for both groups (Garg et al., 2020). We should align the decision rationales for the data in both groups predicted as positive. However, counterfactual fairness (Kusner et al., 2017) quantifies fairness from the perspective of an individual (Garg et al., 2020), which is beyond our current framework. We will further explore it in the future.

A.8 COMBINATION WITH DATA AUGMENTATION.

The data augmentation and our decision rationale alignment are two independent ways to enhance fairness. From Figure 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).

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\text{cls}}(\mathrm{F}(\mathbf{x}), y)) + \lambda \mathcal{L}_{\text{aug}}(\mathrm{F}) + \beta \sum_{k=0}^{K} d_k, \tag{2}$$

We denote the above method for DP regularization as DRAlign(ΔDP, 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.

Table 10: 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

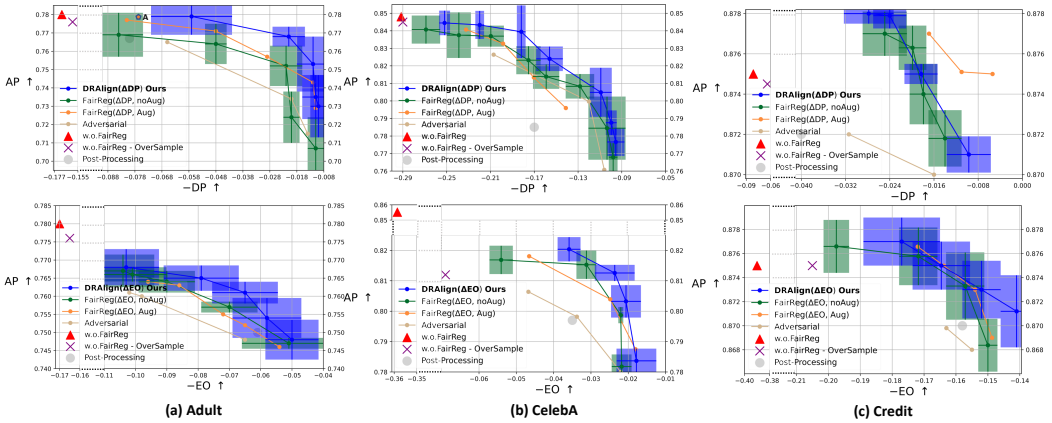

Figure 1: Accuracy and fairness comparison with error bar.

A.9 WITHOUT THE FAIRNESS REGULARIZATION.

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 Figure 2 (main paper), 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.

A.10 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.11 ERROR BAR.

Here, we only show the error bar of our experimental results in Fig. 1 on FairReg(ΔDP, noAug) and DRAlign for better observation. It should be noted that the x-coordinate -DP and y-coordinate AP are both changing with the random seed. We here plot a rectangular region for the error bar of each data point. Moreover, we mark a point "A" in the figure of the Adult dataset under -DP metric. Although we plot a rectangular region for the error bar, it does not mean that point A (-0.073, 0.777) can be reached by FairReg(ΔDP, noAug). It just means the -DP value and the AP value of FairReg(ΔDP, noAug) could arrive at -0.073 and 0.777 separately with different random seeds.

REFERENCES

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

P. Garg, J. Villasenor, and V. Foggo. 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.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

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

Tom R. Tyler. 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.

Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 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, 2020. Association for Computing Machinery. ISBN 9781450371216. doi: 10.1145/3377811.3380331. URL https://doi.org/10.1145/3377811.3380331.