We thank all of the reviewers for their helpful reviews. We performed all of the experiments that were suggested: *(1)* we changed the objective function so that a specified level of fairness is guaranteed; *(2)* we performed experiments where there are multiple protected attributes simultaneously (e.g. race and gender on COMPAS); *(3)* we added a large image dataset (CelebA) to our experiments, which shows that adversarial fine-tuning is the most powerful method for complex tasks; *(4)* we implemented the algorithm from a similar paper, Zhang et al. "Mitigating Unwanted Biases with Adversarial Learning", and we show that it does not outperform our methods. We hope that these experiments help to address the reviewer concerns. Please see the details below.

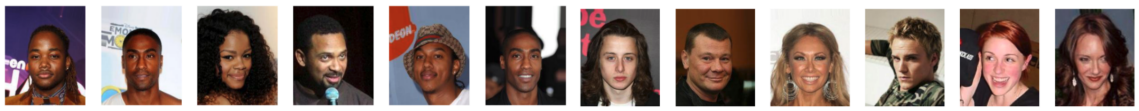
R1 We agree that our original objective function has no guarantee that the outcome is fair. Therefore, we perform an additional experiment where the objective function is $1 - \text{acc} \cdot \left(1 + e^{500(|\text{bias}| - 0.03)}\right)^{-1}$, which enforces the constraint $|\text{bias}| < 0.03$ since it behaves like a (smoothed) indicator function. See the table below (row 1). Next, we show that our algorithms perform well when there are multiple protected attributes. We ran an experiment on COMPAS protecting gender and race, as the reviewer suggested. See the table below (row 2). Next, we ran our experiments on a more complex dataset (the image dataset CelebA), and this one clearly shows the effectiveness of adversarial fine-tuning over the simpler methods (which was a concern of the reviewer). We used a ResNet trained to predict whether a person is smiling, and we debias with respect to race. See the table (row 3) and the figure below. We agree that we should compare our approach to other in-processing approaches (see our response to R2 for one of them). We also note that comparing to in-processing algorithms is sometimes impossible. For example, the in-processing algorithms require the full training dataset, while post-hoc methods only use the validation dataset. Furthermore, it is common to start with a large, expensive model such as GPT-3 or EfficientNet, where retraining from scratch is infeasible. Finally, we agree with all the smaller comments/clarifications (such as adding a table with accuracy/bias before and after debiasing) and we will correct these in the final version of our paper.

R2 As suggested, we compared our approaches to Zhang et al. "Mitigating Unwanted Biases with Adversarial Learning". Note there is a key difference: in that paper, the critic model learns to predict the protected attribute, while in our paper, the critic directly predicts bias. See the table below (row 4 and column 8) for the results, which shows that our methods outperform Zhang et al. Finally, we agree with all the smaller comments and will incorporate them into the paper: adding the three papers to related work, including all details of the critic model, and adding a full table of our results for clarity.

R3 We will make sure to include the size of the datasets in the final version of the paper, which are as follows (and we use a train/val/test split of 60/20/20): ACI: 48842, BM: 45211, COMPAS: 10331, and CelebA (see response to R1): 60000. We see generally that random search performs better on smaller datasets, and adversarial fine-tuning performs better on large datasets. Finally, we will improve the visibility of Figure 1 in our paper.

R4 We will make sure to give much more intuition behind our methods in the final version of the paper. We agree that we should use more datasets, so we ran experiments on the CelebA image dataset (see our response to R1). It is a great idea to include qualitative examples. We give some examples below for the CelebA dataset.

		Default	ROC	EqOdds	CalibEqOdds	Random	Adversarial	LayerwiseOpt	Zhang et al.
(1) Bias Guarantee (on ACI)	objective (See R1)	1	1	1	1	0.18	0.2	1	0.23
	|bias|	0.082	0.055	0.15	0.31	0.02	0.011	0.14	0.006
	performance	0.85	0.79	0.94	0.84	0.83	0.8	0.56	0.77
(2) COMPAS (race & gender)	objective (Eq 1)	0.23	0.14	0.15	0.33	0.15	0.21	0.14	0.16
	|bias|	0.18	0.0016	0.17	0.18	0.025	0.14	0	0.052
	performance	0.66	0.54	0.91	0.34	0.56	0.61	0.54	0.58
(3) CelebA (race)	objective (Eq 1)	0.059	0.072	0.25	0.086	0.051	0.046	0.25	—
	|bias|	0.025	0.052	0.48	0.05	0.0077	0.0054	0	—
	performance	0.91	0.91	0.98	0.88	0.91	0.91	0.5	—
(4) Comp. w. Zhang et al. (on ACI)	objective (Eq 1)	0.098	0.094	0.13	0.27	0.053	0.063	0.062	0.067
	|bias|	0.082	0.055	0.15	0.31	0.0011	0.0047	0	0.025
	performance	0.85	0.79	0.94	0.84	0.79	0.76	0.75	0.81



Before	74%	28%	23%	51%	77%	5%	3%	93%	99%	2%	99%	83%
After	92%	35%	29%	61%	84%	11%	2%	92%	99%	2%	99%	84%

Figure 1: Probability of smiling on the CelebA dataset, before and after debiasing w.r.t. race.