A. Appendix

A.1. Additional Metric Details

We provide additional details here about $\text{BiasAmp}_{\rightarrow}$, as defined in Sec. 4.

In practice the indicator variable, y_{at}, is computed over the statistics of the training set, whereas everything else is computed over the test set. The reason behind this is that the direction of bias is determined by the existing biases in the training set.

Comparisons of the values outputted by $\text{BiasAmp}_{\rightarrow}$ should only be done relatively. In particular, within one of the directions at a time, either $A \rightarrow T$ or $T \rightarrow A$, on one dataset. Comparing $A \rightarrow T$ to $T \rightarrow A$ directly is not a signal as to which direction of amplification is stronger.

A.2. Details and Experiment from Variance in Estimator Bias

For the models we trained in Sec. 5.2, we performed hyperparameter tuning on the validation set, and ended up using the following: ResNet18 had a learning rate of .0001, AlexNet of .0003, and VGG16 of .00014. All models were trained with stochastic gradient descent, a batch size of 64, and 10 epochs. We use the given train-validation-test split from the CelebA dataset.

Our method for surveying prominent fairness papers is as follows: on Google Scholar we performed a search for papers containing the keywords of "fair", "fairness", or "bias" from the year 2015 onwards, sorted by relevance. We did this for the three conferences of 1) Conference on Neural Information Processing Systems (NeurIPS), 2) International Conference on Machine Learning (ICML), and 3) ACM Conference on Fairness, Accountability, and Transparency (FAccT). We picked these conferences because of their high reputability as machine learning conferences, and thus would serve as a good upper bound for reporting error bars on fairness evaluation metrics. We also looked at the International Conference on Learning Representations (ICLR), but the Google Scholar search turned up very few papers on fairness. From the three conferences we ended up querying, we took the first 25 papers from each, pruning those that were either: 1) not related to fairness, or 2) not containing fairness metrics for which it error bars could be relevant (e.g., theoretical or philosophical papers). Among the 48 papers that were left of the 75, if there was at least one graph or table containing a fairness metric that did not appear to be fully deterministic, and no error bars were included (even if the number reported was a mean across multiple runs), this was marked to be a "non-error-bar" paper, of which 25 of the 48 papers looked into met this criteria.

A.3. Details on Measuring Bias Amplification in FitBERT

Here we provide additional details behind the numbers presented in Tbl. 2 in Sec. 5.3.

As noted, and done, by (Liang et al., 2020), a large and diverse corpus of sentences is needed to sample from the large variety of contexts. However, that is out of scope for this work, where we run FitBERT on 20 sentence templates of the form "[1) he/she/(they)] [2) is/was] a(n) [3) adjective] [4) occupation]". By varying 2) and using the top 10 most frequent adjectives from a list of adjectives (Hugsy, 2017) that appear in the English Wikipedia dataset (one of the datasets BERT was trained on) that would be applicable as a descriptor for an occupation (pruning adjectives like e.g., "which", "left") for 3), we end up with 20 template sentences. We then alternate conditioning on 1) (to calculate A→T) and 4) (to calculate T→A). The 10 adjectives we ended up wtih are: new, known, single, large, small, major, French, old, short, good. We use the output probabilities rather than discrete predictions in calculating $P(\hat{A}_a = 1|T_t = 1)$ and $P(\hat{T}_t = 1|A_a = 1)$ because there is no "right" answer in sentence completion, in contrast to object prediction, and so we want the output distribution.

When calculating the amount of bias amplification when the base rates are equal, we picked the direction of bias based on that provided by the WinoBias dataset. In practice, this can be thought of as setting the base correlation, $P(A_a = 1|T_t = 1)$ for a men-biased job like "cook" to be $.5 + \epsilon$ for "he" and $.5 - \epsilon$ for "she" when there are two pronouns, and $.33 + \epsilon$ for "he" and $.33 - \epsilon$ for "she" and "they", where in practice we used $\epsilon = 1\text{e}{-7}$. This ensures that the indicator variable, y_{at} from Eq. 2, is set in the direction fo the gender bias, but the magnitudes of Δ_{at} are not affected to a significant degree.

To generate a rough approximation of what training correlation rates could look like in this domain, we look to one of the datasets that BERT was trained on, the Wikipedia dataset. We do so by simply counting the cooccurrences of all the occupations along with gendered words such as "man", "he", "him", etc. There are flaws with this approach because in a sentence like "She went to see the doctor.", the pronoun is in fact not referring to the gender of the person with the occupation. However, we leave a more accurate measurement of this to future work, as our aim for showing these results was

more for demonstrative purposes illustrating the manipulation of the correlation rate, rather than in rigorously measuring the training correlation rate.

We use 32 rather than 40 occupations in WinoBias (Zhao et al., 2018), because when we went to the 2016 U.S. Labor Force Statistics data (of Labor Statistics, 2016) to collect the actual numbers of each gender and occupation in order to be able to calculate $P(T_t = 1|A_a = 1)$, since WinoBias only had $P(A_a = 1|T_t = 1)$, we found 8 occupations to be too ambiguous to be able to determine the actual numbers. For example, for "attendant", there were many different attendant jobs listed, such as "flight attendants" and "parking lot attendant", so we opted rather to drop these jobs from the list of 40. The 8 from the original WinoBias dataset that we ignored are: supervisor, manager, mechanician, CEO, teacher, assistant, clerk, and attendant. The first four are biased towards men, and the latter four towards women, so that we did not skew the distribution of jobs biased towards each gender.

A.4. COCO Masking Experiment Broken Down by Object

In Table 1 of Sec. 4 we perform an experiment whereby we measure the bias amplification on COCO object detection based on the amount of masking we apply to the people in the images. We find that $\text{BiasAmp}_{A \to T}$decreases when we apply masking, but $\text{BiasAmp}_{T \to A}$increases when we do so. To better inform mitigation techniques, it is oftentimes helpful to take a more granular look at which objects are actually amplifying the bias. In Table 3 we provide such a granular breakdown. If our goal is to target $\text{BiasAmp}_{A \to T}$, we might note that objects like `tv` show decreasing bias amplification when the person is masked, while `dining table` stays relatively stagnant.

Table 3. A breakdown of BiasAmp$_{A \to T}$ and BiasAmp$_{T \to A}$ by object for the masking experiment done on COCO in Table 1.

Object	Original		Noisy Person Mask		Full Person Mask	
	$A \to T$	$T \to A$	$A \to T$	$T \to A$	$A \to T$	$T \to A$
teddy bear	-0.13 ± 0.04	1.23 ± 0.32	0.13 ± 0.04	2.27 ± 0.34	-0.09 ± 0.04	1.93 ± 0.43
handbag	0.44 ± 0.14	7.88 ± 0.67	0.4 ± 0.17	5.62 ± 2.13	0.24 ± 0.13	2.96 ± 3.03
fork	0.62 ± 0.22	7.67 ± 1.31	0.63 ± 0.18	7.45 ± 1.65	0.76 ± 0.24	6.22 ± 0.62
cake	-0.29 ± 0.06	5.7 ± 0.59	-0.09 ± 0.04	5.2 ± 0.59	-0.16 ± 0.06	3.3 ± 1.7
bed	0.01 ± 0.04	1.33 ± 1.43	0.06 ± 0.08	5.33 ± 1.43	-0.09 ± 0.03	6.67 ± 0.0
umbrella	-0.05 ± 0.07	9.52 ± 3.92	0.08 ± 0.06	13.33 ± 2.64	-0.04 ± 0.11	11.24 ± 2.76
spoon	0.06 ± 0.06	-2.91 ± 2.78	0.03 ± 0.06	-10.91 ± 4.83	0.04 ± 0.03	-15.27 ± 4.11
giraffe	0.21 ± 0.13	5.32 ± 1.51	0.2 ± 0.05	4.05 ± 2.46	0.02 ± 0.04	4.95 ± 1.34
bowl	0.28 ± 0.04	2.2 ± 1.02	0.06 ± 0.07	4.8 ± 2.56	0.18 ± 0.12	7.4 ± 2.26
knife	-0.22 ± 0.12	-11.74 ± 2.39	-0.35 ± 0.12	-10.43 ± 3.15	-0.31 ± 0.05	-8.84 ± 2.18
wine glass	-0.41 ± 0.14	-5.24 ± 0.83	-0.62 ± 0.09	-7.14 ± 3.49	-0.69 ± 0.07	-10.95 ± 3.64
dining table	-0.75 ± 0.14	4.53 ± 1.02	-0.76 ± 0.14	3.18 ± 2.12	-0.74 ± 0.17	4.58 ± 1.38
cat	0.07 ± 0.04	2.41 ± 2.05	0.19 ± 0.05	10.0 ± 2.42	0.17 ± 0.02	20.0 ± 1.81
sink	0.18 ± 0.15	-4.29 ± 2.4	0.11 ± 0.07	-5.03 ± 2.6	-0.12 ± 0.1	-3.19 ± 1.5
cup	-0.3 ± 0.09	-12.36 ± 3.42	-0.15 ± 0.08	-12.0 ± 1.32	-0.12 ± 0.03	-14.42 ± 2.94
potted plant	0.21 ± 0.18	6.36 ± 5.1	0.34 ± 0.07	0.0 ± 2.82	0.32 ± 0.08	-4.55 ± 2.52
refrigerator	-0.06 ± 0.03	9.6 ± 1.36	-0.07 ± 0.06	7.47 ± 1.75	0.01 ± 0.03	12.0 ± 4.25
microwave	-0.01 ± 0.02	-3.5 ± 5.3	-0.01 ± 0.03	13.0 ± 13.32	-0.03 ± 0.03	6.0 ± 9.05
couch	-1.35 ± 0.16	-0.25 ± 1.24	-0.94 ± 0.23	1.62 ± 1.06	-1.21 ± 0.16	0.15 ± 0.8
oven	0.07 ± 0.09	7.67 ± 1.73	-0.33 ± 0.12	10.67 ± 3.0	0.03 ± 0.13	13.12 ± 1.49
sandwich	-0.98 ± 0.15	-8.93 ± 0.92	-2.6 ± 0.18	-15.23 ± 2.78	-2.46 ± 0.4	-15.67 ± 1.74
book	-0.43 ± 0.08	-3.07 ± 0.57	-0.48 ± 0.13	-3.34 ± 1.24	-0.85 ± 0.11	-3.18 ± 2.0
bottle	0.05 ± 0.11	-7.73 ± 1.54	-0.13 ± 0.14	-8.33 ± 3.87	0.06 ± 0.06	-11.52 ± 1.65
cell phone	-0.09 ± 0.13	3.6 ± 1.92	0.05 ± 0.15	13.72 ± 0.89	-0.1 ± 0.12	18.72 ± 2.36
pizza	-0.19 ± 0.1	6.85 ± 1.81	-0.09 ± 0.03	15.17 ± 2.41	-0.38 ± 0.12	9.3 ± 1.56
banana	0.35 ± 0.08	4.42 ± 0.66	0.56 ± 0.09	6.1 ± 1.06	0.1 ± 0.19	5.19 ± 0.72
toothbrush	-0.47 ± 0.11	-2.42 ± 2.63	-0.55 ± 0.13	-4.32 ± 5.14	-0.5 ± 0.12	-11.85 ± 4.43
tennis racket	-0.09 ± 0.08	9.22 ± 3.05	0.0 ± 0.09	14.75 ± 2.74	0.09 ± 0.12	14.75 ± 1.38
chair	-0.31 ± 0.11	1.87 ± 0.22	-0.31 ± 0.06	3.85 ± 0.37	-0.31 ± 0.05	4.69 ± 0.0
dog	0.14 ± 0.04	0.43 ± 0.19	0.3 ± 0.07	1.6 ± 0.29	0.3 ± 0.04	1.7 ± 0.35
donut	-0.3 ± 0.08	-1.4 ± 0.29	-0.39 ± 0.15	-0.0 ± 0.59	-0.41 ± 0.15	0.09 ± 0.37
suitcase	-0.43 ± 0.08	1.96 ± 1.12	-0.09 ± 0.01	7.3 ± 0.65	-0.11 ± 0.14	8.43 ± 1.4
laptop	0.27 ± 0.07	4.58 ± 3.57	0.06 ± 0.04	7.67 ± 6.33	0.1 ± 0.06	17.39 ± 5.52
hot dog	1.48 ± 0.12	7.63 ± 2.72	1.51 ± 0.07	9.37 ± 1.75	1.86 ± 0.14	9.16 ± 3.03
remote	0.33 ± 0.02	9.14 ± 2.73	0.15 ± 0.09	11.03 ± 2.59	0.15 ± 0.07	18.62 ± 3.12
clock	0.77 ± 0.16	4.48 ± 3.05	0.62 ± 0.08	9.61 ± 3.02	0.88 ± 0.11	11.44 ± 3.78
bench	-0.02 ± 0.05	11.49 ± 3.61	-0.06 ± 0.06	14.68 ± 4.73	-0.04 ± 0.06	16.6 ± 2.61
tv	0.35 ± 0.09	5.16 ± 4.59	0.27 ± 0.12	14.19 ± 2.58	0.21 ± 0.09	18.06 ± 3.03
mouse	-0.22 ± 0.06	0.95 ± 5.76	-0.26 ± 0.05	4.57 ± 4.99	-0.18 ± 0.06	5.14 ± 4.11
horse	0.07 ± 0.03	8.13 ± 5.17	0.11 ± 0.07	13.63 ± 1.96	0.16 ± 0.04	16.59 ± 4.92
fire hydrant	-0.21 ± 0.07	4.71 ± 2.98	-0.15 ± 0.07	1.76 ± 4.64	-0.18 ± 0.06	5.0 ± 6.19
keyboard	0.01 ± 0.08	1.64 ± 3.4	-0.08 ± 0.08	17.38 ± 4.51	0.02 ± 0.08	31.15 ± 6.74
bus	0.02 ± 0.04	-11.33 ± 2.15	-0.17 ± 0.07	-9.0 ± 3.53	-0.1 ± 0.04	3.0 ± 6.62
toilet	0.26 ± 0.06	7.65 ± 4.07	0.2 ± 0.09	12.17 ± 4.15	0.19 ± 0.06	17.22 ± 5.67
person	-0.04 ± 0.1	-1.47 ± 4.54	0.1 ± 0.07	-4.12 ± 5.25	0.03 ± 0.08	-3.53 ± 4.65
traffic light	-0.2 ± 0.03	4.44 ± 2.48	-0.27 ± 0.06	6.06 ± 1.86	-0.27 ± 0.11	11.52 ± 2.28
sports ball	-1.44 ± 0.2	3.41 ± 1.56	-0.95 ± 0.27	3.86 ± 1.83	-1.27 ± 0.32	4.95 ± 1.97
bicycle	-0.23 ± 0.07	13.58 ± 2.02	0.01 ± 0.15	13.09 ± 3.98	-0.21 ± 0.08	12.72 ± 1.83
car	0.2 ± 0.2	3.7 ± 1.49	0.57 ± 0.1	-0.74 ± 2.2	0.32 ± 0.1	0.37 ± 2.66
backpack	0.01 ± 0.03	7.57 ± 1.63	0.0 ± 0.06	17.84 ± 3.09	-0.05 ± 0.04	18.38 ± 2.53

Table 3. Continued

Object	Original		Noisy Person Mask		Full Person Mask	
	$A \to T$	$T \to A$	$A \to T$	$T \to A$	$A \to T$	$T \to A$
train	0.15 ± 0.07	8.47 ± 3.33	0.27 ± 0.08	16.82 ± 1.86	0.14 ± 0.16	19.28 ± 3.97
kite	-0.09 ± 0.04	-2.67 ± 3.35	-0.16 ± 0.07	-7.56 ± 3.4	-0.22 ± 0.05	-4.89 ± 3.78
cow	-0.17 ± 0.08	1.6 ± 1.95	0.11 ± 0.08	0.86 ± 3.0	0.15 ± 0.07	-3.58 ± 2.55
skis	0.12 ± 0.09	0.24 ± 2.13	0.25 ± 0.03	-1.31 ± 3.01	0.23 ± 0.1	-7.5 ± 1.76
truck	-0.27 ± 0.04	-13.33 ± 1.68	-0.25 ± 0.12	-10.26 ± 1.42	-0.16 ± 0.03	-17.95 ± 2.46
elephant	-0.58 ± 0.15	9.09 ± 1.25	-0.24 ± 0.08	4.29 ± 2.94	-0.63 ± 0.06	1.69 ± 3.13
boat	0.03 ± 0.05	2.8 ± 1.79	-0.02 ± 0.07	-0.0 ± 2.48	0.05 ± 0.05	3.2 ± 3.06
frisbee	0.09 ± 0.17	-1.36 ± 1.18	-0.14 ± 0.22	0.7 ± 2.45	-0.79 ± 0.17	-6.88 ± 2.28
airplane	0.14 ± 0.07	4.23 ± 3.27	0.15 ± 0.07	5.77 ± 4.4	0.16 ± 0.05	5.77 ± 3.69
motorcycle	-0.06 ± 0.04	-6.35 ± 3.94	0.01 ± 0.06	-9.04 ± 2.42	0.12 ± 0.04	-1.73 ± 6.64
surfboard	-0.02 ± 0.05	3.83 ± 4.79	-0.06 ± 0.07	1.74 ± 1.36	-0.08 ± 0.06	2.43 ± 4.83
tie	0.16 ± 0.06	3.29 ± 2.1	0.08 ± 0.07	3.53 ± 3.45	0.23 ± 0.04	4.71 ± 4.47
snowboard	0.4 ± 0.13	10.05 ± 2.1	0.53 ± 0.16	9.64 ± 1.0	0.41 ± 0.22	7.9 ± 1.7
baseball bat	0.46 ± 0.04	-3.72 ± 2.47	0.45 ± 0.1	-3.36 ± 1.8	0.23 ± 0.12	1.24 ± 2.17
baseball glove	-0.03 ± 0.05	27.06 ± 4.12	-0.16 ± 0.04	34.9 ± 1.29	-0.1 ± 0.03	36.47 ± 0.84
skateboard	-0.28 ± 0.03	11.37 ± 3.51	-0.34 ± 0.08	10.98 ± 3.86	-0.35 ± 0.11	5.88 ± 3.77