On the Independence of Bias and Fairness in Language Models

Anonymous[1]

[1]Affiliation

Abstract

The societal impact of pretrained language models has prompted researchers to probe them for strong associations between protected attributes and value-loaded terms, from slur to prestigious job titles. Such work is said to probe models for *bias or fairness*—or such probes 'into representational biases' are said to be 'motivated by fairness'—suggesting an intimate connection between bias and fairness. We provide conceptual clarity by distinguishing between association biases and empirical fairness and show the two can be independent.

Introduction The prevalence of unintended social biases in pretrained language models is alarming, since they impact millions, if not billions of people every day. Researchers have studied such biases focusing on what Crawford [7] called *representational bias*, which manifests when portrayals of certain demographic groups are discriminatory. In NLP, representational bias often arises when associations between a protected attribute, e.g., gender, and certain concepts, e.g., job titles, are captured in the model space. Thus, to avoid ambiguity, we will refer to this type of bias as *association bias*, following Chaloner and Maldonado [5]; Figure 1 illustrates this concept.

Association bias is often confused with what is sometimes referred to as performance disparity [10] or *empirical fairness* [14], i.e., performance differences across end user demographics[1]. Or it is believed that mitigating association bias is assumed to improve empirical fairness [6, 9, 3, 8, 4, 13].

We show that theoretically, association bias and empirical fairness can be completely independent. That is, mitigating bias can hurt fairness, and ensuring fairness can introduce more bias.

[1]This work is based on the definition of fairness as equal performance across groups.

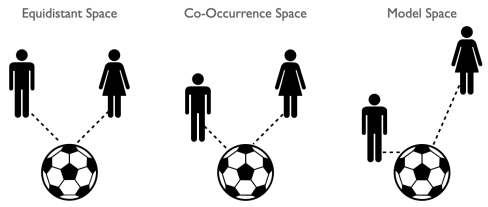

Figure 1: Association bias of group-related terms: *man* may be strongly associated with *soccer* in a *model*, less so *empirically*, and not at all in an *equidistant* space.

Association Bias and Fairness are Independent We produce a thought experiment—a synthetic model—to illustrate how bias and fairness can in fact be completely independent of one another. We construct a synthetic ternary (positive/negative/neutral) sentiment analysis model with a small feature space, including words that refer to demographic subgroups of a population. These words, denoting various groups, will be biased and associated with sentiment, because of biases in our training data. This assumption is also made in Ali et al. [1], for example. These associations lead to biased likelihood estimates. We show, however, that the resulting biases are independent of the group fairness of the model, i.e., to the min-max performance disparities across the same groups. Such a connection, if it exists, could be explained by an *in-group affinity*, which relies on the assumption that those biased terms are used by the in-group more frequently or in other ways than by other groups.

Say a population consists of members of groups $g_1, \ldots, g_4$, e.g., defined according to their address as *north*, *east*, *west* and *south*. Everyone speaks the same language and expresses sentiment with a vocabulary of seven words: $\{w_{g_1}, \ldots, w_{g_4}, w_5, w_6, w_7\}$. Except w_6 (positive)

and w_7 (neutral), all words express negative sentiment, including the words that refer to (or are associated with) other demographic subgroups (w_{g_i}), for instance, *northern*, *eastern*, *western* and *southern*. The subgroups use the terms with the following probabilities (Table 1):

	w_{g_1}	w_{g_2}	w_{g_3}	w_{g_4}	w_5	w_6	w_7
g_1	0.0	0.25	0.0	0.0	0.25	0.25	0.25
g_2	0.0	0.0	0.25	0.0	0.25	0.25	0.25
g_3	0.0	0.0	0.0	0.25	0.25	0.25	0.25
g_4	0.25	0.0	0.0	0.0	0.25	0.25	0.25

Table 1: Probability of a group g_i using the word w_j for expressing sentiment. Only w_6 (positive) and w_7 (neutral) express a non-negative sentiment.

This data exhibits four representational biases, e.g., the association of g_1 with negative sentiment, the association of g_2 with negative sentiment, and so forth. If we have sufficient data, a simple model, e.g., a Naive Bayes classifier trained on simple bag-of-words representations, should induce the maximum likelihood estimates (where '0' denotes negative, '1' positive and '2' neutral sentiment) showcased in Table 2.

	$P(w_{g_1}\|0)$	$P(w_{g_2}\|0)$	$P(w_{g_3}\|0)$	$P(w_{g_4}\|0)$	$P(w_5\|0)$	$P(w_6\|0)$	$P(w_7\|0)$
g_1	0.0	0.25	0.0	0.0	0.25	0.0	0.0
g_2	0.0	0.0	0.25	0.0	0.25	0.0	0.0
g_3	0.0	0.0	0.0	0.25	0.25	0.0	0.0
g_4	0.25	0.0	0.0	0.0	0.25	0.0	0.0

	$P(w_{g_1}\|1)$	$P(w_{g_2}\|1)$	$P(w_{g_3}\|1)$	$P(w_{g_4}\|1)$	$P(w_5\|1)$	$P(w_6\|1)$	$P(w_7\|1)$
g_1	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_2	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_3	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_4	0.0	0.0	0.0	0.0	0.0	0.25	0.0

	$P(w_{g_1}\|2)$	$P(w_{g_2}\|2)$	$P(w_{g_3}\|2)$	$P(w_{g_4}\|2)$	$P(w_5\|2)$	$P(w_6\|2)$	$P(w_7\|2)$
g_1	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_2	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_3	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_4	0.0	0.0	0.0	0.0	0.0	0.0	0.25

Table 2: Maximum likelihood estimates from a linear classifier on our synthetic data modelled in Table 1.

Now, say we employ an existing debiasing approach and manage to debias the model with respect to its representation of group g_1 by setting $P(w_{g_1}|0) = P(w_{g_1}|1) = P(w_{g_1}|2)$, which, in this case, would equal zero. This would hurt performance on data from g_4 (bottom row), increasing the empirical risk on this sub-population, but more surprisingly, note that it would not help us on classifying the data from g_1. That is, an attempt to make the model fairer towards *north* by equalizing the use of the term *northern*, would result in increased unfairness towards members from *south*, who tend to use *northern* more often (and in a negative context). Removing bias in how terms referring to a group are represented, only improves performance on data from members from that group, if these members use such in-group terms in non-standard ways, i.e., differently from everyone else. In the absence of this assumption, what we call *In-Group Affinity Assumption*, association bias and empirical fairness are orthogonal.

Note that while we make use of a linear model and likelihood estimates in our thought experiment, it would be very easy to translate this into a deep neural network and cosine distances instead. To see this, consider, for example, how any Naive Bayes model can be translated into a deep neural network, and how the differences in likelihood can, under such a translation, be translated into differences in cosine instances.

Conclusion The independence of representational bias and fairness as equal performance shown here runs counter to the NLP literature, where bias and fairness have been assumed to be intimately connected. In an attempt to theoretically explain why this does not hold always true, we devise a synthetic experiment. We show that, regardless of the metric used for assessing model biases and fairness, the assumption that bias and fairness are always negatively correlated, or that one is a cause of the other, is not always true. In many aspects of private and public life, we encounter decisions or patterns where bias and fairness exist or fluctuate independently of each other, or in which they are negatively correlated. In affirmative action, for example, we tolerate and encourage a (more) biased decision-making process to achieve (higher) fairness. While positive discrimination is heavily debated [11, 2, 12], it is a good example of a biased process intended to increase the level of fairness. We introduced the *In-Group Affinity Assumption* to highlight the assumption that a particular demographic groups use in-group terms more frequently–or in different ways–than other groups (non-standard). This, we argue, is a necessary assumption to drive a causal connection between bias and fairness, if it exists. We believe research should be done by disentangling the two.

References

[1] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf. Xai for transformers: Better explanations through conservative propagation, 2022. URL https://arxiv.org/abs/2202.07304.

[2] L. Barmes. Equality law and experimentation: The positive action challenge. *The Cambridge Law Journal*, 68(3):623–654, 2009. ISSN 00081973, 14692139. URL http://www.jstor.org/stable/40388838.

[3] Y. Cao, Y. Pruksachatkun, K.-W. Chang, R. Gupta, V. Kumar, J. Dhamala, and A. Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.62. URL https://aclanthology.org/2022.acl-short.62.

[4] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1), March 2022. doi: 10.1038/s41598-022-07939-1. URL https://doi.org/10.1038/s41598-022-07939-1.

[5] K. Chaloner and A. Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3804. URL https://aclanthology.org/W19-3804.

[6] W.-F. Chen, K. Al Khatib, H. Wachsmuth, and B. Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcss-1.16. URL https://aclanthology.org/2020.nlpcss-1.16.

[7] K. Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

[8] E. Dayanik and S. Padó. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.404. URL https://aclanthology.org/2020.acl-main.404.

[9] N. Friedrich, A. Lauscher, S. P. Ponzetto, and G. Glavaš. DebIE: A platform for implicit and explicit debiasing of word embedding spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 91–98, Online, Apr. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.11. URL https://aclanthology.org/2021.eacl-demos.11.

[10] T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1934–1943. PMLR, 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

[11] H. Holzer and D. Neumark. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, September 2000. doi: 10.1257/jel.38.3.483. URL https://www.aeaweb.org/articles?id=10.1257/jel.38.3.483.

[12] M. Noon. The shackled runner: time to rethink positive discrimination? *Work, Employment and Society*, 24(4):728–739, 2010. doi: 10.1177/0950017010380648. URL https://doi.org/10.1177/0950017010380648.

[13] C. Reddy, D. Sharma, S. Mehri, A. Romero Soriano, S. Shabanian, and S. Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf.

[14] A. Shen, X. Han, T. Cohn, T. Baldwin, and L. Frermann. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only, Nov. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.8.