On the Independence of Association Bias and Empirical Fairness in Language Models

Laura Cabello  
lcp@di.ku.dk  
University of Copenhagen  
Copenhagen, Denmark

Anna Katrine Jørgensen  
akj@di.ku.dk  
University of Copenhagen  
Copenhagen, Denmark

Anders Søgaard  
soegaard@di.ku.dk  
University of Copenhagen  
Copenhagen, Denmark

ABSTRACT

The societal impact of pre-trained 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 [11] and empirical fairness [56] and show the two can be independent. Our main contribution, however, is showing why this should not come as a surprise. To this end, we first provide a thought experiment, showing how association bias and empirical fairness can be completely orthogonal. Next, we provide empirical evidence that there is no correlation between bias metrics and fairness metrics across the most widely used language models. Finally, we survey the sociological and psychological literature and show how this literature provides ample support for expecting these metrics to be uncorrelated.

CCS CONCEPTS

- Computing methodologies → Natural language processing.

KEYWORDS

Representational Bias, Fairness, Natural Language Processing

PERMISSION TO MAKE DIGITAL OR HARD COPIES OF ALL OR PART OF THIS WORK FOR PERSONAL OR CLASSROOM USE IS GRANTED WITHOUT FEE PROVIDED THAT COPIES ARE NOT MADE OR DISTRIBUTED FOR PROFIT OR COMMERCIAL ADVANTAGE AND THAT COPIES BEAR THIS NOTICE AND THE FULL CITATION ON THE FIRST PAGE. COPYRIGHTS FOR COMPONENTS OF THIS WORK OWNED BY OTHERS THAN ACM MUST BE HONORED. ABSTRACTING WITH CREDIT IS PERMITTED. TO COPY OTHERWISE, OR REPUBLISH, TO POST ON SERVERS OR TO REDISTRIBUTE TO LISTS, REQUIRES PRIOR SPECIFIC PERMISSION AND/OR A FEE. REQUEST PERMISSIONS FROM PERMISSIONS@ACM.ORG.

ACM Reference Format:

1 INTRODUCTION

The prevalence of unintended social biases in pre-trained language models (PLMs) is alarming, since they impact millions, if not billions of people every day. In recent years, more and more NLP researchers have studied such biases, making up an estimated 6.3% of the literature in 2022 [53]. Much of this work has focused on what Crawford [19] 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 [16].

Association bias is often confused with what is sometimes referred to as performance disparity [29] or empirical fairness [56], i.e., performance differences across end user demographics. Or mitigating association bias is assumed to improve empirical fairness [13, 14, 18, 21, 25, 49]. Note that most fairness metric focus on some form of equal performance and differ only in whether they focus on precision, recall or balancing the two [4]. Empirical fairness refers to equal performance as measured by de facto standard metrics and is arguably the most common fairness metric [4, 62].

In this paper, we will show that the two phenomena, association bias and empirical fairness, are often completely independent matters.1 We devise a thought experiment (§3) to illustrate this, but also present a series of experiments (§4) to show that results obtained the way association bias is normally measured, do not correlate with results obtained the way empirical fairness is normally measured.

Our main contribution, however, is to show that this should not come as a surprise. Research on mitigating association bias and empirical fairness is often motivated by fairness concerns, and bias and fairness are often considered near-synonymous terms in the research literature: Researchers have, for example, said that bias causes unfairness [14, 17, 25]. If this was the case, the independence of association bias and empirical fairness should come as a great surprise. However, the assumption that bias causes unfairness, is unwarranted, as we will see below, from a survey of relevant literature from the social sciences (§5). A causal link between association bias and empirical fairness would seem to require some sort of in-group affinity, i.e., that groups use terms relating to their in-group peers more and in different ways than outsiders, like, for instance, Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts [23]. This assumption, which we call the In-Group Affinity Assumption, seems intuitive, but without much support from the social sciences (§5).

Contributions. In §2, we define association bias and empirical fairness and discuss related work. When we talk about association

1Note that the distinction between association bias and empirical fairness—between how expressions referring to demographic groups are encoded, and how these groups are treated as end users—is different from another distinction made in recent work [22, 26, 35] between intrinsic and extrinsic bias: Intrinsic bias, here, is what we call representational bias, whereas extrinsic bias refers to performance differences on sentences containing entities referring to different demographic groups.
bias, we refer to systematic biases in how words and phrases referring to demographic groups are encoded. Figure 1 visualizes how models may exhibit biased associations because of sample biases, and may even amplify these. We define empirical fairness as equal performance across groups, because this is the most balanced and most widely applicable measure of fairness in NLP, except for specialized applications where equal base rates and calibration take priority over performance. We then move to study how association bias and empirical fairness relate. In §3, we show that theoretically, association bias and empirical fairness are completely independent. That is, mitigating association bias can hurt empirical fairness, and ensuring empirical fairness can introduce more bias. §4 shows there is no obvious correlation between results obtained from standard association bias measurements and results obtained from standard empirical fairness measurements of language models. Finally, §5 surveys the social science literature for explanations on why association bias and empirical fairness may be less related (or related in less obvious ways) than multiple works in the NLP literature have assumed up to this point. The finding that association bias and empirical fairness are independent in this three-way investigation, should help push research horizons and provide strong motivation for targeting empirical fairness directly, as well as for seeing association bias mitigation, not necessarily as a way of promoting fairness, but rather as a way of preventing poor inferences and generation of stereotypical text.

2 DEFINITIONS AND RELATED WORK

In the NLP literature, bias and fairness are often conflated, or it is argued that one follows from the other, e.g., that we can ensure fairness by mitigating bias [13, 14, 18, 21, 25, 49]. In contrast, we will show that this is not always the case, and (association) bias and (empirical) fairness often are independent or at odds.

Bias. Mitigating social biases in NLP models has become an important research goal [33, 51, 55], but there is little consensus on how to evaluate such biases [8, 57]. We focus on association bias and show how, contrary to what seems to be popular belief, it is not unequivocally related to fairness; in fact, it is very often completely independent thereof.

Association bias in a model refers to systematic differences in how words and phrases referring to demographic groups are encoded. Classical tests thereof include comparing the (cosine) distance of terms relating to protected attributes, e.g., woman and man, or their vectors $v_w$ and $v_m$, to terms of particular interest, e.g., slur [54], sentiment [1], or job titles such as doctor ($w_d$) [64]. Early papers would quantify bias with respect to, say, gender, as cosine similarities $(\cos(v_w, w_d) - \cos(v_m, w_d))$ [6, 10, 12, 27, 64], and by seeing whether the nearest neighbor of $w_d + v_w - v_m$ would be nurse or another job stereotypically associated with women [9]. In practice, NLP researchers have used tests such as the ones above for quantifying association bias [6, 10, 12, 27, 64]. We will argue that such quantities are theoretically and, often practically, orthogonal to empirical fairness, which we define in terms of differences in performance estimates across demographics, i.e., social groups [4, 56, 62], often defined by the cross-product of a subset of protected attributes such as gender, age, or race.

Fairness. Fairness metrics come in multiple flavors, but are often divided in three: calibration-based, precision-based, and recall-based metrics. Miconi [44], Friedler et al. [24] and Kleinberg et al. [36] show how pairs of fairness metrics can be mathematically incompatible, i.e., one type of fairness can rule out another. In fact, incompatibility holds for all pairs of metrics such that the two metrics are of different flavor, e.g., calibration-based and recall-based, unless the true base rates are identical across groups, or the classifier has perfect performance. Since the vast majority of NLP applications provide repetitive services, the quality of which can be measured against a gold standard, precision- and recall-based metrics are predominantly used in NLP. We follow several authors [15, 28, 29] in using min-max differences in (the standard) performance (metric) as our go-to fairness metric. Relying on min-max difference captures the widely shared intuition that fairness is always in the service of the worst off group [48]. For a discussion of available fairness metrics, and in what contexts they are relevant, see Mehrabi et al. [43] and Barocas et al. [4]. For a comparison of existing metrics used to quantify social biases in NLP, see Czarnowska et al. [20].

Related Work. Maity et al. [41] study the effect of subpopulation shifts on performance disparities and show that these do not always relate in obvious ways. Goldfarb-Tarrant et al. [26] study the correlation between what they refer to as ‘intrinsic and extrinsic measures of representational bias’. Their intrinsic measures of representational bias amount to word association bias, but their extrinsic measures of representational bias are not empirical fairness measures. To see this, consider the coreference task used in Goldfarb-Tarrant et al. [26]. Goldfarb-Tarrant et al. [26] correlate intrinsic gender bias measures (cosine distances in static word embedding spaces) with coreference performance on sentences with female and male referents. We argue that in this case, empirical fairness would be performance on sentences written by female and male authors. Goldfarb-Tarrant et al. [26] establish that there is no correlation between their two measures of representational bias. Their result superficially looks similar to and in agreement with

\[ \cos(v_w, w_d) - \cos(v_m, w_d) \]

\[ w_d + v_w - v_m \]
ours, but is, in fact, unrelated. If anything, it shows that association bias has been assumed to correlate with many measures that it does not, in fact, correlate with. Cao et al. [13] and Kaneko et al. [35] studied the same problem as Goldfarb-Tarrant et al. [26], but used contextualized token embeddings from PLMs rather than static word embeddings. They both found weak correlations between intrinsic and extrinsic evaluation measures. Again, we emphasize that these results do not contradict ours.

Shah et al. [55] carefully avoid to discuss fairness, saying the fairness literature is outside the scope of their paper, but place outcome disparity (performance disparity) as a central motivation for social bias mitigation. They list four potential causes of outcome disparity: label bias, selection bias, bias amplification, and semantic (representation) bias. We show association (representation) bias and outcome disparity are theoretically, and also often practically, independent, questioning their fourth hypothesis. Moreover, we observe that outcome disparity can arise in the absence of all of the above four factors. Say a group exhibits more variance than others, e.g., because of spelling variation in dyslexics. Even if dyslexics are represented proportionally or equally, they may still see worse performance with dyslexics than for non-dyslexics.

Finally, Shen et al. [56] show how a different form of representational fairness, i.e., whether protected author attributes can be detected from model representations, is also uncorrelated with empirical fairness. Together, our work and previous work [13, 26, 35, 56] establish that four common bias-related measures – (i) association bias, (ii) performance on sentences with protected attribute terms (Goldfarb-Tarrant et al. [26]’s extrinsic measure), (iii) decodability of protected attributes from representations, and (iv) empirical fairness are largely uncorrelated. Specifically, (i) is independent of (ii) and (iv), and (iii) is independent of (iv).

Our work is motivated by the large-spread assumption that association bias and empirical fairness are causally related [5, 13, 14, 18, 21, 25, 40, 47, 52, 58]. Bartl et al. [5], for example, aspire “to promoting fairness in NLP by exploring methods to measure and mitigate gender bias.” Ross et al. [52] say they “believe that by revealing biases, by providing tests for biases that are as focused as possible on the smallest units of systems, we can both assist the development of better models and allow the auditing of models to ascertain their fairness.” Sun et al. [58], argue that “biased predictions may discourage minorities from using those systems and having their data collected, thus worsening the disparity in the data sets”, equating biased predictions with unfair predictions.

All three sets of authors see bias as the primary cause of fairness. Showing such causation is not a given, and that in fact, association bias and empirical fairness need not even correlate and are often orthogonal, is an important correction to this literature, with potential consequences for research methodology, applications of NLP in the social sciences, as well as AI ethics and regulation.

3 ASSOCIATION BIAS AND EMPIRICAL FAIRNESS ARE INDEPENDENT (IN THEORY)
In this section, 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. [2], for example. These associations lead to biased likelihood estimates and would, in the context of a linear model, lead to differences in the degree of isomorphism relative to the group-specific subgraphs. We will 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: $w_1, \ldots, w_7$.

<table>
<thead>
<tr>
<th>$g_1$</th>
<th>$w_{g_1}$</th>
<th>$w_{g_2}$</th>
<th>$w_{g_3}$</th>
<th>$w_{g_4}$</th>
<th>$w_5$</th>
<th>$w_6$</th>
<th>$w_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_1$</td>
<td>0.0</td>
<td>0.25</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$g_2$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$g_3$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$g_4$</td>
<td>0.25</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

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.

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, association bias and
empirical fairness are orthogonal. We will refer to this assumption as the In-Group Affinity Assumption.

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 Naïve 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.

4 ASSOCIATION BIAS AND EMPIRICAL FAIRNESS SCORES ARE UNCORRELATED (IN PRACTICE)

In this section, we study whether association bias and empirical fairness are correlated in practice, i.e., when actual models are evaluated on actual data designed to probe bias and fairness. We apply well-established metrics for measuring the two. While bias and fairness can be studied with respect to any prospective attribute, the vast majority of NLP research has focused on (binary) gender 

In this section, we study whether association bias and empirical fairness are correlated in practice, i.e., when actual models are evaluated on actual data designed to probe bias and fairness. We apply well-established metrics for measuring the two. While bias and fairness can be studied with respect to any prospective attribute, the vast majority of NLP research has focused on (binary) gender

Bias. To measure representational bias, we use three popular metrics, i.e., the Log Probability Bias Score (LPBS) proposed by Kurita et al. [37], as well as two variants of the Word Embedding Association Test (WEAT) [12] for assessing bias in contextual word representations: the adaption proposed by Tan and Celis [59] (henceforth, WEAT_t), and the alternative suggested by Lauscher et al. [38] (henceforth, WEAT_k). All these metrics rely on association tests to compute the relationship between a set of related targets \{t_1, t_2, \ldots\}, e.g., gender words, and attributes \{a_1, a_2, \ldots\}, e.g., occupation words, through definitions of template sentences designed to convey no meaning beyond that of the terms inserted into them.

Kurita et al. [37] use template sentences like \( T = \{ "\text{TARGET}\} \) is a \{ATTRIBUTE\}. The target word is masked, and the attribute word is a place-holder for a specific word denoting an occupation, e.g., \( T_m = \{ \text{MASK} \} \) is a chef\. LPBS uses the prior probability of the target word \( p(\text{prior}) \), i.e., the probability of a target \( t_i \) being generated when both \( t_i \) and the attribute \( a_j \) are masked, as a normalizer, and computes the association as the relative increase in log probability:

\[
\frac{\log p([\text{MASK}] = t_i | T_m)}{p(\text{prior})} = l_{pbs} \tag{1}
\]

The difference between the relative increased log probability scores for two targets is the LPBS measure of bias. For linear models, this correlates strongly with the \( \epsilon \)-isometry of the target word subgraph relative to an equidistant space, if we make the centroid of the set of attribute vectors the reference point. For a non-linear language model, we can compute the \( \epsilon \)-isometry of its linear approximation. Table 3 are for the targets “he” and “she”. \( \epsilon \)-test is used to evaluate the statistical significance of the metric, in which the means of \( a_{\text{he}, a_j}^{\text{lpbs}} \) and \( a_{\text{she}, a_j}^{\text{lpbs}} \) are compared. We draw \( 10^5 \) random permutations, meaning that the \( p \)-values observed will not be less than \( 10^{-5} \).

Tan and Celis [59] follow the methodology of May et al. [42], who extended the WEAT metric to sentences (SEAT) inserting the word of interest in context templates such as \( T = \{ \text{"This is _" } \} \) Tan and Celis [59] use the contextual embedding of the token of interest, instead of using the sentence encoding, to compute the cosine similarities (associations). Lauscher et al. [38] follow Vulic et al. [60] and average the pooled embeddings of the first four attention layers for the word of interest \( t_i \) or \( a_j \) in a template without context, e.g. \{\text{CLS} t_i \text{SEP}\}\. Both approaches report the effect size [12], a normalized measure of how separated the association distributions of target and attributes are. The statistical significance of the associations is also computed with a permutation test as in [12]. Both approaches are an instance of computing the \( \epsilon \)-isometry of the template sentence subgraphs in the cosine metric space. See Table 3 for empirical results. We see that results are somewhat mixed, with LPBS and the two variants of WEAT often disagreeing which models are more biased. All the metrics are evaluated on the same list of sixty attributes –equally split into female and male stereotyped professions from the US bureau of labour, provided in [22].

Bias. To measure representational bias, we use three popular metrics, i.e., the Log Probability Bias Score (LPBS) proposed by Kurita et al. [37], as well as two variants of the Word Embedding Association Test (WEAT) [12] for assessing bias in contextual word representations: the adaption proposed by Tan and Celis [59] (henceforth, WEAT_t), and the alternative suggested by Lauscher et al. [38] (henceforth, WEAT_k). All these metrics rely on association tests to compute the relationship between a set of related targets \{t_1, t_2, \ldots\}, e.g., gender words, and attributes \{a_1, a_2, \ldots\}, e.g., occupation words, through definitions of template sentences designed to convey no meaning beyond that of the terms inserted into them.

Kurita et al. [37] use template sentences like \( T = \{ "\text{TARGET}\} \) is a \{ATTRIBUTE\}. The target word is masked, and the attribute word is a place-holder for a specific word denoting an occupation, e.g., \( T_m = \{ \text{MASK} \} \) is a chef\. LPBS uses the prior probability of the target word \( p(\text{prior}) \), i.e., the probability of a target \( t_i \) being generated when both \( t_i \) and the attribute \( a_j \) are masked, as a normalizer, and computes the association as the relative increase in log probability:

\[
\frac{\log p([\text{MASK}] = t_i | T_m)}{p(\text{prior})} = l_{pbs} \tag{1}
\]

The difference between the relative increased log probability scores for two targets is the LPBS measure of bias. For linear models, this correlates strongly with the \( \epsilon \)-isometry of the target word subgraph relative to an equidistant space, if we make the centroid of the set of attribute vectors the reference point. For a non-linear language model, we can compute the \( \epsilon \)-isometry of its linear approximation. Table 3 are for the targets “he” and “she”. \( \epsilon \)-test is used to evaluate the statistical significance of the metric, in which the means of \( a_{\text{he}, a_j}^{\text{lpbs}} \) and \( a_{\text{she}, a_j}^{\text{lpbs}} \) are compared. We draw \( 10^5 \) random permutations, meaning that the \( p \)-values observed will not be less than \( 10^{-5} \).

Tan and Celis [59] follow the methodology of May et al. [42], who extended the WEAT metric to sentences (SEAT) inserting the word of interest in context templates such as \( T = \{ \text{"This is _" } \} \) Tan and Celis [59] use the contextual embedding of the token of interest, instead of using the sentence encoding, to compute the cosine similarities (associations). Lauscher et al. [38] follow Vulic et al. [60] and average the pooled embeddings of the first four attention layers for the word of interest \( t_i \) or \( a_j \) in a template without context, e.g. \{\text{CLS} t_i \text{SEP}\}\. Both approaches report the effect size [12], a normalized measure of how separated the association distributions of target and attributes are. The statistical significance of the associations is also computed with a permutation test as in [12]. Both approaches are an instance of computing the \( \epsilon \)-isometry of the template sentence subgraphs in the cosine metric space. See Table 3 for empirical results. We see that results are somewhat mixed, with LPBS and the two variants of WEAT often disagreeing which models are more biased. All the metrics are evaluated on the same list of sixty attributes –equally split into female and male stereotyped professions from the US bureau of labour, provided in [22].
We sample members of each group (female and male) in a balanced way across subgroups, as defined by the other variables. This is done to evaluate whether representational bias correlates with outcomes such as the tendency to use gendered language (e.g., "he"/"she"). A list of occupations as attributes. For example, we might compare masked word predictions to human cloze tests, quantifying how often a language model agrees with the members of a particular social group on what is the most likely word in contexts such as: "After waiting three hours, Cal whined and started to...

### Table 3: Three metrics of representational bias. Values are the average difference of associations between the target words “he”/“she,” and a list of occupations as attributes. Larger values reflect a more severe bias. A positive value hints a skewed distribution towards males. A negative value hints a skewed distribution towards females. \( \ast \) statistically significant at 0.01.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \Delta P_{@1} )</th>
<th>( \Delta MRR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert-base-uncased</td>
<td>0.86*</td>
<td>1.01*</td>
</tr>
<tr>
<td>bert-base-cased</td>
<td>0.90*</td>
<td>1.00*</td>
</tr>
<tr>
<td>bert-large-uncased</td>
<td>0.20</td>
<td>0.83*</td>
</tr>
<tr>
<td>bert-large-cased</td>
<td>-1.10*</td>
<td>0.60</td>
</tr>
<tr>
<td>bert-base-multilingual-cased</td>
<td>-1.98*</td>
<td>0.36</td>
</tr>
<tr>
<td>distilbert-base-uncased</td>
<td>-0.46</td>
<td>0.79</td>
</tr>
<tr>
<td>roberta-base</td>
<td>-2.32*</td>
<td>0.51</td>
</tr>
<tr>
<td>roberta-large</td>
<td>-2.63*</td>
<td>0.24</td>
</tr>
<tr>
<td>google/electra-small-generator</td>
<td>-0.20</td>
<td>0.71</td>
</tr>
<tr>
<td>google/electra-large-generator</td>
<td>-2.64*</td>
<td>0.73</td>
</tr>
<tr>
<td>DistilBERT-base-uncased</td>
<td>1.63</td>
<td>0.64</td>
</tr>
<tr>
<td>DistilBERT-base-v2</td>
<td>0.74</td>
<td>0.94</td>
</tr>
<tr>
<td>DistilBERT-xlarge-v2</td>
<td>1.45</td>
<td>1.21</td>
</tr>
<tr>
<td>DistilBERT-base-large</td>
<td>0.48</td>
<td>0.41</td>
</tr>
</tbody>
</table>

### Table 4: Macro-averaged precision and mean reciprocal rank differences between male and female subgroups following experiments in [63]. Values close to zero are preferred for a more equitable model.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \Delta P_{@1} )</th>
<th>( \Delta MRR )</th>
</tr>
</thead>
<tbody>
<tr>
<td>bert-base-uncased</td>
<td>0.69</td>
<td>1.57</td>
</tr>
<tr>
<td>bert-base-cased</td>
<td>0.15</td>
<td>0.74</td>
</tr>
<tr>
<td>bert-large-uncased</td>
<td>0.91</td>
<td>1.34</td>
</tr>
<tr>
<td>bert-large-cased</td>
<td>-0.07</td>
<td>0.32</td>
</tr>
<tr>
<td>bert-base-multilingual-cased</td>
<td>0.89</td>
<td>0.54</td>
</tr>
<tr>
<td>roberta-base</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>roberta-large</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>google/electra-small-generator</td>
<td>0.97</td>
<td>0.43</td>
</tr>
<tr>
<td>google/electra-large-generator</td>
<td>1.22</td>
<td>0.97</td>
</tr>
</tbody>
</table>

**Fairness.** Our fairness evaluation is based on Zhang et al. [63]’s work, who study how the predictions of various PLMs align with the linguistic preferences of different social groups. They directly compare masked word predictions to human cloze tests, quantifying how often a language model agrees with the members of a particular social group on what is the most likely word in contexts such as: "After waiting three hours, Cal whined and started to [MASK].

Zhang et al. [63] use, as their fairness metric, the min-max difference in precision \( \Delta P_{@1} \) across groups defined by the cross-product of several protected attributes, including gender, age, race, and level of education. Since we are comparing with binary gender bias probes, we only consider fairness across (binary) gender here. We sample members of each group (female and male) in a balanced way across subgroups, as defined by the other variables. This is equivalent to reporting the macro-average across subgroups for each group. \( \Delta P_{@1} \) is thus the difference in performance between male and female groups, macro-averaged across subgroups in the cloze test data. We follow Zhang et al. [63] in also reporting the difference in mean reciprocal rank as a second performance metric \( \Delta MRR \). See the individual scores in Table 4.

Results show performance gaps between binary gender groups. Consequently, we would expect models exhibiting high degree of bias in Table 3 to be the least fair. However, this is not the case. Figure 2 displays the results for bias and fairness jointly, often highlighting the lack of correlation. Note that, ideally, all data-points should belong to the bottom-right quadrant.

### Metrics are uncorrelated.** Now that we have our evaluation framework defined, let us analyze whether representational bias correlates with outcome disparity. This amounts to studying the correlations between LPBS and WEAT metrics and the min-max \( P_{@1} \) difference across groups. We report the sign of the Pearson correlation coefficient to ease the interpretation of the (ideally) monotonic relationship\(^4\) between each set of metrics in Figure 2.

Results are two-fold:

(i) The discrepancy across sub-graphs in Figure 2 aligns with results in May et al. [42], Delobelle et al. [22] and Cao et al. [13], who all found different representational bias metrics to lead to mutually inconsistent results. WEAT\(_L\) and WEAT\(_T\) are related and show some agreement, but generally, results are wildly different across metrics.

(ii) More importantly, for our purposes, representational bias and fairness-as-equal-performance (quantified as min-max differences across performance scores for different groups) are, in fact, uncorrelated. Models with high bias values are the most fair according to our fairness metric, and vice versa. These cases are highlighted in red in Figure 2. For example, roberta-base (rb) is among the most biased models according to LPBS, but it exhibits the highest degree of fairness wrt. the MRR metric—and second highest wrt. P\(_{@1}\). The bigger PLM, roberta-large (rl) is slightly less biased according to LPBS, but it is generally less fair. Values from the WEAT metrics are, in this case, somewhat mixed.

Result (ii) is evidence against the In-Group Affinity Assumption and for the independence of bias and fairness. Looking at each model family—separated by horizontal lines in Table 3 and 4—model size does not systematically lead to larger or smaller bias scores, and it does not seem strongly correlated with any of the fairness metrics either.

In the following section, we survey research in the social sciences that also suggest the In-Group Affinity Assumption is mostly false, with one important caveat: Slur words have marked in-group dominance. In most applications, this exception would be insufficient to drive a causal link between association bias and empirical fairness.

\(^4\)We deliberately omit the magnitude of the Pearson coefficient to emphasize the sign of the correlation. Ideally, bias and fairness metrics should have a negative linear dependence \( (p < 0) \).
Figure 2: Scatter plots show the relationship between different representational bias metrics and fairness evaluation. The upper row displays results when evaluating fairness through precision at top-1 ($P@1$). The bottom row displays results when considering $MRR$ to evaluate fairness. The division into quadrants is done according to average scores. Each point represents a language model, labelled with its initials. We see no support for a strong negative correlation between bias and fairness. Red points mark the clear counter-examples to such a negative correlation. Global trend for each plot is summarized with the sign of Pearson coefficient ($p$).

because slur words are rare, and performance differences across social groups are pervasive.

5 ASSOCIATION BIAS AND EMPIRICAL FAIRNESS ARE SOMETIMES AT ODDS (IN HUMANS)

The thought experiment in §3 shows that bias and fairness can in fact be completely independent or orthogonal. The experiments in §4 further showed that there is no direct correlation between the association bias in a model $M$ toward social groups $g_1, \ldots, g_n$, and the performance disparity (fairness) of $M$ across data from these groups $g_1, \ldots, g_n$.

In such cases, debiasing a model with respect to the representation of a certain group (e.g., $g_1$) has no impact on the performance of the model for users from the group. The beneficiaries of such a debiasing procedure are, in other words, not necessarily the group the debiasing was intended to increase fairness for. The idea that debiasing word representations that are related to a particular group increases the fairness of the model for that group, relies on the assumption that those words are also used by the in-group more frequently or in other ways than by other groups. This assumption—which we called the In-Group Affinity Assumption—seems problematic, since there are plenty of examples in the literature of the opposite. In the following, we briefly review some examples that originate from the NLP literature; others from the social sciences.

We are often likely to talk more about members of other groups than our in-group peers. Li and Dickinson [39], for example, find that some of the most indicative n-grams for detecting young female users on Chinese social media are the names of male pop stars. Correcting or debiasing the representations of these names would not improve model fairness on texts written by the male pop stars, but rather on texts written by young female users.

Morgan-Lopez et al. [45] show that young (pre-college) children talk more about college on Twitter than adults in their college age. Wei and Santos Jr. [61] analyze data from Twitter and Reddit and find that the most predictive n-grams for Israeli users include “Iraqis” and “Palestinians”, while for Palestinian users “Israeli military detention centres” and “Lieberman settler rabbis” (referring to the Israeli Defence Minister, Avigdor Lieberman) are among the most predictive n-grams.

Generally, political debates are often experienced as negative in both tone and nature. According to a 2019 Pew Research Center study, 85% of Americans say that the political debate has become “more negative”.

One explanation for the increase in negative sentiment in political discourse is increased attention to what members

\[\text{pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/}\]
of other (political) groups do wrong compared to what the in-group peers do right. Supporting this explanation, Jensen et al. [34] show, for example, that one of the most partisan phrases used by US Democrats in congressional texts was “great Republican Party”.

Similarly, Duijnhoven [23] finds that Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts. In analyzing the language of German political parties, Biessmann [7] likewise finds that the left-wing party, Linke, has a high frequency of mentions of large corporations (konzerne) and policies that negatively impact the social welfare.

On Slur. Some slurring terms (e.g. “dyke”, “queer” and “bitch”) have been reclaimed or reappropriated by the target group resulting in a semantic discrepancy dependent on the speaker’s group membership [31, 50]. This results in what we term the In-Group Affinity Assumption, where the in-group’s use of the term will differ significantly from that of the out-groups. Any debiasing of the term will have no significant impact on the performance for the in-groups, since the language model’s representation of the term will reflect the majority use of the term, which will not be that of the in-group. However, since slurs are per definition defamatory terms, debiasing these terms will result in less insulting outputs in downstream tasks, and this may result in a higher perception of fairness for the target group.

6 DISCUSSION AND CONCLUSION

The independence of representational bias and fairness-as-equal-performance shown here, along with the falsification of the In-Group Affinity Assumption, runs counter to the NLP literature. Bias and fairness have been assumed to be intimately connected, and the In-Group Affinity Assumption has been implicit and unquestioned in much recent work. The results we present in this paper are, at the same time, in a sense not surprising. Or they should not be surprising. 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 [3, 32, 46], it is a good example of a biased process intended to increase the level of fairness.

Methods for correctly assessing model biases remains an open research question. Current evaluation benchmarks give inconsistent results [13, 22, 42]. Moreover, as discussed in §2, evaluating model biases with metrics that only consider local geometries, such as cosine-based metrics, can be inadequate. The fairness metric literature is also full of controversies [24, 30, 36, 44], but there is a broad consensus that performance disparity or outcome disparity is a real challenge for responsible NLP research and development. This consensus is not only limited to NLP research, but also found in legal studies, machine ethics, and the social sciences. Our results have shown that regardless of these open problems in bias and fairness research, the assumption that bias and fairness are always negatively correlated, and that one is a cause of the other, is not always true. Despite being closely related, it is important to understand that biases exist everywhere, but might not be unequivocally harmful. And similarly, fairness issues may arise in non-biased scenarios.

Finally, it is worth noting that we should not solely focus on the correlation between protected attributes such as race or gender and the model’s output, but rather ask the question if they are causing the outcome, and, whether the model is unfair to individuals in virtue of their membership in a certain group [30].

Conclusion. We reviewed part of the NLP literature showing how many researchers conflate bias and fairness, i.e., representational bias and fairness-as-equal-performance, or argue that fixing one will solve the other. In an attempt to explain why this does not hold always true, we devised a thought experiment in §3: a synthetic model that illustrates how bias and fairness can be completely independent of one another. 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. In §5, we surveyed the social science literature and found evidence that often the opposite is the case, which substantiates our findings in §3 and §4. Our survey includes examples from the social sciences, as well as from NLP research, where bias and fairness are (locally) negatively correlated. This provides strong reason to be skeptical of the In-Group Affinity Assumption and shows that bias and fairness are often independent or orthogonal to each other.

In sum, we have shown the importance of studying bias and fairness independently of one another and cautioned against the In-Group Affinity Assumption. We think this, potentially, could lead to a valuable reorientation of the NLP literature, enabling researchers to study representational bias in more adequate ways, focusing on robustness and generation (to avoid bias reinforcement). This also highlights the different contributions of representational bias benchmarks and in-the-wild evaluation datasets with demographic information that can be used to evaluate performance disparities across groups. Bias and fairness seem to be separate issues, and we believe research should be done by disentangling the two.

7 LIMITATIONS

Our paper addresses the relationship between the two specific interpretations of bias and fairness, i.e., representational bias and fairness-as-equality. These are, in our view, the most common and most important definitions of bias and fairness in the NLP literature, but they are not the only ones. We hope others will follow up with studies of how other definitions relate. Our experiments in §3 were limited to English benchmark datasets. We agree with Ruder et al. [53] that the prevalence of bias and fairness studies using English data, is most unfortunate, and we are, in parallel, working to create multilingual benchmarks for bias and fairness studies.

REFERENCES


