Trustworthy Social Bias Measurement

Rishi Bommasani[1], **Percy Liang**[1]

[1]Stanford University
nlprishi@stanford.edu

Abstract

How do we design measures of social bias that we *trust*? While prior work has introduced several measures, no measure has gained widespread trust: instead, mounting evidence argues we should distrust these measures. In this work, we design bias measures that warrant trust based on the cross-disciplinary theory of measurement modeling. To combat the frequently fuzzy treatment of social bias in natural language processing, we explicitly define social bias, grounded in principles drawn from social science research. We operationalize our definition by proposing a general bias measurement framework `DivDist`, which we use to instantiate 5 concrete bias measures. To validate our measures, we propose a rigorous *testing protocol* with 8 testing criteria (e.g. predictive validity: do measures predict biases in US employment?). Through our testing, we demonstrate considerable evidence to trust our measures, showing they overcome conceptual, technical, and empirical deficiencies present in prior measures.

Introduction

Language technologies shape our lives and broader societal function. As NLP researchers, our work has increasingly direct, immediate, and significant impact: we must reckon with this and, especially, any harms that arise from language technology. Social bias is fundamental to this analysis (Hovy and Spruit 2016; Bender et al. 2021; Weidinger et al. 2022, *inter alia*): how we represent people and what we associate them with has material consequences. Biased language technology can cause several types of harm (Dev et al. 2022; Bommasani et al. 2021, §5.1): allocational (e.g. lower hiring rates for marginalized groups due to algorithmic resume screening), representational (e.g. associating Muslims with violence in machine-generated content), and psychological (e.g. stereotype threat that influences individuals to conform to group-level perceptions).

Measurement is the primary lens for understanding social bias. And measurement is essential to reducing bias: to determine if an intervention mitigates bias, the measured bias should decrease due to the intervention. If all paths forward for making progress on bias in NLP pass through measurement, then what is the current state of bias measurement?

Many works have proposed bias measures, spanning different settings like text, vector representations, language models, and task-specific models (see Blodgett et al. 2020; Dev et al. 2022). Most measure bias for two social groups. However, no standard exists for what evidence is required to *trust* these measures: works provide a mixture of intuitive, empirical, and theoretical justifications. Perhaps as a consequence, many works don't hold up to scrutiny: measures have been shown to be brittle (Ethayarajh, Duvenaud, and Hirst 2019; Nissim, van Noord, and van der Goot 2020; Antoniak and Mimno 2021; Delobelle et al. 2022), contradictory (Bommasani, Davis, and Cardie 2020), unreliable (Aribandi, Tay, and Metzler 2021; Seshadri, Pezeshkpour, and Singh 2022), invalid (Blodgett et al. 2021), and the space overall is unclear on what bias means and what metrics purport to measure (Blodgett et al. 2020). Trust is necessary: for metrics to productively guide progress and inform decision-making, we must trust them.

Consequently, we focus on *trustworthy* bias measurement. We apply *measurement modeling* to address this challenge: measurement modeling is an expansive theory used across the social sciences to design and validate measures of (complex) social constructs (Loevinger 1957; Messick 1987; Jackman 2008; Jacobs and Wallach 2021). Therefore, measurement modeling is well-suited to social bias measurement: the theory has a rich history, including even for social bias in humans (e.g. the Implicit Association Test; Chequer 2014).

Under measurement modeling, we must first define the *theoretical construct* of social bias. In contrast, Blodgett et al. (2020) showed many works in NLP failed to (adequately) define social bias. To define social bias, we draw upon principles in social science research; these principles dictate how we operationalize our definition into a *general* measurement framework. Our measurement framework `DivDist`, based on divergences between probability distributions, improves over prior work in two key ways: (i) *compatibility*, meaning bias measures can be instantiated for several settings (e.g. text, vector representations) that yield comparable measurements and (ii) *multi-group*, meaning bias can be measured for not just two social groups. These properties are valuable for NLP: for example, we may want to understand how different processes change biases (e.g. the potential bias amplification between training

data and a learned model, or between a generative model and its samples), which requires compatability. Or, given the unique forms of marginalization experienced by intersectional groups (Crenshaw 1989), we will need to compare biases in relation to more than two groups.

Beyond offering generality, our framework also makes explicit that bias is fundamentally a *relative* phenomenon, which has been neglected in all prior work. To meaningfully measure bias, one must state the *normative* reference frame: what would constitute (no) bias? This is a material consideration: the relevant reference could be a particular ideal (e.g. equal association across groups), social status quo (e.g. the US labor demographics), or technical contrast (e.g. a model's training data), but regardless the choice determines what bias even means. By allowing the reference to be specified, rather than being assumed, our framework enables pluralism: different normative positions can dictate what constitutes bias.

Using our framework, we instantiate 5 new bias measures, spanning measures for text, word embeddings, and contextualized representations. We put these measures to the test, alongside several prior measures. Measurement modeling (following the presentation of Jacobs and Wallach (2021)) specifies 8 well-studied desiderata: in a sense, measurement modeling provides a well-established checklist of criteria to build trust in measures of social constructs. For each desideratum, we design a test, amounting to the first rigorous *testing protocol* for validating bias measures. Executing these tests, we accrue evidence to trust our measures, while surfacing concerns with prior measures.

Beyond our primary contributions (measurement framework, testing protocol), we make several striking findings while testing our measures. First, our bias measure for word embeddings strongly correlates with societal trends in employment, whereas some prior measures are uncorrelated or even *anti-correlated*, suggesting our measure is more appropriate for certain computational social science applications. Second, our measures indicate the representations in GPT-2 (Radford et al. 2019) amplify biases relative to GPT-2's training data, but this amplification remains latent and unobserved when sampling from the model, which poses broader questions regarding how biases acquired in training language models propagate downstream (Goldfarb-Tarrant et al. 2021; Steed et al. 2022). Third, "debiasing" methods generally fail to reduce, and sometimes even *exacerbate*, social bias according to our measure, which calls into question their effectiveness (Gonen and Goldberg 2019).

Principles for Social Bias

Notation. Following conventions in the social sciences, we define social bias in terms of social *groups* $G_1, \ldots, G_k$,[1] which reflect a categorization of individuals (Allport 1954), and a *target concept* T, which bias is measured with respect to. As an example, we may consider the (binary) gender biases in science with $G_1 = \text{female}, G_2 = \text{male}$ and $T = \text{scientist}$.

[1]We acknowledge that many categories (e.g. race, gender) are the subject of abundant disagreement (Crenshaw 1989; Penner and Saperstein 2015).

Reducing bias to associations. Given social groups and a target concept, some theories define bias as the target concept's differential *association* with each group. For example, in the Implicit Association Test (Greenwald, McGhee, and Schwartz 1998), the test uses response time to quantify the association between the target concept and each group. Further, these associations must be *systematic*: Beukeboom and Burgers (2019) write that "bias is a *systematic* asymmetry", meaning that social bias pertains to broader social groups (Friedman and Nissenbaum 1996), not particular individuals (cf. Bommasani et al. 2022).

Bias is Relative

If a machine translation model exactly replicates the properties of its training data, is it biased? It depends. Relative to its training data, no, but relative to a specific societal reference, potentially yes, namely if the training data was biased with respect to this reference. Most bias measures in NLP ignore this fundamental property: bias is, instead, portrayed as absolute by many measures.

This fundamentally misconstrues what bias is: bias is an inherently *relative* construct, which requires that a *reference* be specified. Bias is precisely the extent to which the observed associations diverge from this reference. Since bias emerges through social processes, reference-sensitive measures allow us to understand how different decisions increase/reduce bias (Friedman and Nissenbaum 1996). In this spirit, Shah, Schwartz, and Hovy (2020) and Hovy and Prabhumoye (2021) attribute bias in NLP to several sources (e.g. data selection, data annotation, model training): effective bias measurement should aim to quantify the relative contribution of each of these sources.

Defining Social Bias

Having introduced groups, targets, associations, and references, we define social bias. *Social bias* is the divergence in the observed associations between a target concept and a set of social groups from corresponding reference associations. In particular, most works in NLP and the social sciences construe social bias as an "asymmetry" in the observed associations (e.g. the bias that *scientist* is more associated with the male gender than the female gender), as in Beukeboom and Burgers (2019). This perspective on bias is a special case of our definition, when the reference is the uniform baseline: no bias corresponds to the target concept being equally associated with every social group.

`DivDist` Measurement Framework

Having defined social bias, we propose our two-stage measurement framework `DivDist`, which stands for Divergence between Distributions. First, given *parameters*, `DivDist` yields a bias measure `bias`. Second, given *inputs* (i.e. the target concept, social groups, and reference mentioned in our definition), we define the bias measurement $\texttt{bias}(T, G_1, \ldots, G_k; \mathbf{p}_0)$ (i.e. a numerical value of

how much bias is present).

$$\mathbf{s} = [\texttt{SoA}(T, G_1), \ldots, \texttt{SoA}(T, G_k)] \quad (1)$$
$$\mathbf{p} = \texttt{normalize}(\mathbf{s}) \quad (2)$$
$$\texttt{bias}(T, G_1, \ldots, G_k; \mathbf{p}_0) = D(\mathbf{p}, \mathbf{p}_0) \quad (3)$$

Parameters. To map from the abstract framework DivDist to a concrete bias measure bias, we specify three functions (SoA, normalize, D). First, SoA quantifies the strength of association between the target concept and a social group as a numerical value in $\mathbb{R}_{\geq 0}$. This function handles both setting-specific aspects of measurement (i.e. SoA is considerably different for text vs. vector representations) and the specific associations of interest (e.g. different SoA implementations are needed to measure frequency-related biases vs. more semantic biases). Applying SoA to every (target concept, social group) pair yields the observed association vector $\mathbf{s} \in \mathbb{R}^k_{\geq 0}$. Second, we normalize $\mathbf{s}$ to a categorical distribution $\mathbf{p}$ using normalize. Third, we quantify the divergence using D between the (normalized) observed associations $\mathbf{p}$ and the reference associations $\mathbf{p}_0$, which we also specify as a categorical distribution distributed over the groups.

Observe the clear correspondence between our framework DivDist and our definition: Step 1 extracts the observed associations, Step 2 prepares these associations, and Step 3 measures the divergence from reference associations. This correspondence indicates our measures demonstrate *structural fidelity* (Loevinger 1957), one of 8 desiderata we consider in measurement modeling.

Inputs. To map from the bias measure bias to the bias measurement $\texttt{bias}(T, G_1, \ldots, G_k; \mathbf{p}_0)$, we specify three inputs. We often represent social groups $G_1, \ldots, G_k$ and the target concept T using *word lists*, i.e. representative words that embody the associated concept. We specify the reference $\mathbf{p}_0$ as a distribution over the k social groups that indicates the association between each group and the target concept when there is no bias.

Generality. We prove several prior bias measures, across the social sciences and NLP (Weitzman et al. 1972; Voigt et al. 2017; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018) are special cases of DivDist. In these works, bias is measured in the binary setting as the difference in associations (i.e. how much more associated is the *male* gender than the *female* gender with *scientist*). This interpretation of bias as a "systematic asymmetry" (Beukeboom and Burgers 2019) is recovered by DivDist using the uniform distribution $\mathbf{p}_0 = \left[\frac{1}{2}, \frac{1}{2}\right]$, up to scaling.[2]

$$\begin{aligned}
\underset{\text{prev}}{\texttt{bias}} &= \texttt{SoA}(T, G_1) - \texttt{SoA}(T, G_2) = a_1 - a_2 \\
\underset{\text{ours}}{\texttt{bias}} &= D\left(\texttt{normalize}\left([a_1, a_2]\right), \left[\frac{1}{2}, \frac{1}{2}\right]\right) \\
&= \left\| \left[\frac{a_1}{a_1 + a_2}, \frac{a_2}{a_1 + a_2}\right] - \left[\frac{1}{2}, \frac{1}{2}\right] \right\|_1 \\
&= \frac{a_1 - a_2}{a_1 + a_2}
\end{aligned}$$

Measures

To further demonstrate the generality of DivDist, we instantiate several bias measures (see tab:measures). In NLP, we want to measure bias in a variety of settings: here, we introduce measures for (human-authored or machine-generated) text, (static) word embeddings, and contextualized representations to provide broad coverage. These measures differ in the implementation of the SoA parameter, which encodes the specifics of each setting; the choices for normalize and D are consistent across settings.

Text

Since social bias is a systematic phenomenon, bias in text manifests in distributional statistics. To implement $\texttt{SoA}_{\text{text}}$, we quantify associations in text based on these statistics. The association between concept T and group G_j in a corpus $\mathcal{L}$ is quantified as follows:

1. Select contexts $c_1, \ldots, c_N$ in corpus $\mathcal{L}$ such that each context c_i mentions concept T.
2. For each context c_i, let $a_{ij} \in \{0, 1\}$ indicate if T is associated with G_j in c_i.
3. $\texttt{SoA}_{\text{text}}(T, G_j) = \sum_{i=1}^{N} a_{ij}$.

Contexts, mentions, and associations. The aforementioned procedure partially implements $\texttt{SoA}_{\text{text}}$, but leaves ambiguous: (i) what are contexts c_i, (ii) what does it mean for a concept T to be mentioned in c_i, and (iii) what does it mean for group G_j to be associated with T in c_i? For contexts, we consider three-sentence spans in $\mathcal{L}$ by default, testing the sensitivity of measurements to this choice subsequently. For mentions, these judgments could ideally be made by human domain-experts, but since this is costly for large corpora, we automate this by requiring that we have a word list $W(T)$ for T such that for each context $c_i, \exists w \in W(T)\ s.t.\ w \in c_i$. For associations in a context, we consider two options. In the **human** variant of our text bias measure, humans make these judgments, whereas in the **automated** variant, we require that some word in the group's word list $W(G_j)$ appears *and* that no word in any other group's word list appears.[3]

[2] For brevity, we abbreviate $\texttt{SoA}(G_1, T)$ and $\texttt{SoA}(G_2, T)$ as a_1 and a_2, respectively. WLOG, let $a_1 \geq a_2$.

[3] In pilot experiments measuring bias in English Wikipedia, the second constraint increased the precision of our heuristic (since contexts are more unambiguously associated with the group) with fairly marginal cost in recall.

Word Embeddings

For word embeddings, we quantify associations using cosine similarity, which is the standard similarity metric for word embeddings (e.g. Mikolov et al. 2013) that has garnered some theoretical justification (Zhelezniak et al. 2019). Let $\mathbf{w}$ be the word vector for word w. We quantify the strength of association for word embeddings as

$$\texttt{SoA}_{\text{WE}}(T, G_j) = \cos\left(\frac{\sum_{t \in W(T)} \mathbf{t}}{|W(t)|}, \frac{\sum_{g \in W(G_j)} \mathbf{g}}{|W(g_j)|}\right).$$

In words, $\texttt{SoA}_{\text{WE}}$ is the cosine similarity between the average target word embedding and average group word embedding, closely resembling how Caliskan, Bryson, and Narayanan (2017) and Garg et al. (2018) quantify association.

Contextualized Representations

For contextualized representations, most prior measures (e.g. May et al. 2019; Tan and Celis 2019; Guo and Caliskan 2020) compute a single bias value for these representations. We argue this is a type error: the bias in contextualized representations will depend on the context in which the representations are used (i.e. the text being represented) and, in fact, Ethayarajh (2019) showed these representations are highly context-sensitive. For example, the gender biases in BERT representations (Devlin et al. 2019) will differ when using BERT to embed text from the New York Times vs. a misogynistic subreddit. With that in mind, we present two context-sensitive approaches that quantify strength of association in contextualized representations $\mathbf{w}_i$, which embed word w in context c_i within a text corpus $\mathcal{D}$.

Reduction to $\texttt{SoA}_{\text{WE}}$. Our first approach reduces the contextualized case to the static case, following Bommasani, Davis, and Cardie (2020). For each (group or target) word w of interest, we compute $\mathbf{w} = \mathbb{E}_{c_i \in \mathcal{D} | w \in c_i} \mathbf{w}_i$ as the average of w's contextualized representations across all contexts in which it appears in corpus $\mathcal{D}$. Bommasani, Davis, and Cardie (2020) show this produces high-quality static embeddings from contextualized representations: once we have these static embeddings, we then apply the aforementioned $\texttt{SoA}_{\text{WE}}$ to quantify strength of association for contextualized representations.

Probing. The key downside to the reduction approach is the reduction may distort associations in the original contextualized representations (Bommasani, Davis, and Cardie 2020). Therefore, we consider more direct techniques for interpreting contextualized representations (see Rogers, Kovaleva, and Rumshisky 2020; Belinkov 2021), which closely resemble the *probing* (Alain and Bengio 2017; Hewitt and Liang 2019) methodology studied in the interpretability community. Namely, we learn a classifier f over the representations that simulates the human annotator from the text setting.

Training. f receives a contextual vector $\mathbf{t}_i$ as input and predicts which social group (if any) the target word t is associated with in context c_i. To assemble f's training data, we (i) sample N contexts c_i that mention T in the corpus $\mathcal{D}$, (ii) manually annotate labels y_i indicating which group G_j (if any) that T is associated with in c_i, (iii) embed the contexts c_i, and (iv) extract any contextual representations $\mathbf{t}_i$ for words $t \in W(T)$. (Note that in step (ii), the human annotations can also be re-purposed to measure bias in the text corpus $\mathcal{D}$ itself.) The resulting $\{(\mathbf{t}_i, y_i)\}_{i=1}^N$ examples are used to learn f by minimizing the cross-entropy loss of predicting group labels from the corresponding contextualized representation.

Inference. To quantify strength of association, we sample further disjoint contexts c^{test} that mention T:

$$\texttt{SoA}_{\text{CR-Probe}}(T, G_j) = \sum_{i=1}^{N} 1\left[f(\mathbf{t_i}) = G_j\right].$$

Decisions. Selecting the complexity of the classifier (i.e. probe) have been the subject of intense debate in the probing community (Hewitt and Liang 2019; Pimentel et al. 2020a,b; Belinkov 2021; Hewitt et al. 2021; Pimentel and Cotterell 2021). We choose to learn linear classifiers, which indicates that we prioritize easily (i.e. linearly) decoded associations (Ivanova, Hewitt, and Zaslavsky 2021; Hewitt et al. 2021).

Normalization and Divergence Parameters

Beyond `SoA`, `DivDist` requires normalization $\texttt{normalize} : \mathbb{R}^k \to \Delta^{k-1}$ and divergence $D : \Delta^{k-1} \times \Delta^{k-1} \to \mathbb{R}_{\geq 0}$ to fully instantiate bias measures. For conceptual simplicity, as defaults, we set `normalize` to be dividing a vector by its sum (since the input vectors are generally/always non-negative, so this is a valid means for yielding a probability distribution) and D to be the ℓ_1 distance as a well-known and simple-to-understand divergence. These settings correspond to our proof, where we show our measure generalizes several prior measures, and we show that our measurements are quite robust to these choices empirically in our sensitivity analysis.

Testing Protocol

We have introduced 5 new bias measures, summarized in Table 1: why should we trust them? In NLP, to build trust, we sometimes have a *ground truth*, where a model/measure is trustworthy if it agrees with the ground truth (e.g. human judgments). However, we do not have a ground truth in bias measurement; in fact that's why we are building these measures. We turn to the tradition of *measurement modeling* (Loevinger 1957; Messick 1987; Jackman 2008), which has been used in many social science disciplines to build trust in measures of complex social constructs like bias. Following Messick (1987) and Jackman (2008), we say a measure is trustworthy if it is simultaneously *valid* and *reliable*. Across disciplines and decades, individual criteria have been defined and refined to designate the key criteria for validity and reliability (we closely follow Jacobs and

Setting	Abbreviation	Implementation of SoA
Text	Human	Number of contexts where T and G_j are associated based on human annotator
Text	Aut.	Number of contexts where T and G_j are associated based on cooccurrence
WE	Emb.	Cosine similarity between average embeddings for $W(T)$ and $W(G_j)$
CR	Red.	Cosine similarity between representations averaged across contexts for $W(T)$ and $W(G_j)$
CR	Probe	Number of contexts where T and G_j are associated based on learned probe

Table 1: Summary of the implementations of SoA we introduce for each setting.

Validity	**Face validity**	Measure passes basic sanity checks.
	Content validity	Measure faithfully reflects theoretical understanding of the construct.
	Convergent validity	Measure correlates with other credible measures of the same construct.
	Predictive validity	Measure predicts other credible measures of related constructs.
	Hypothesis validity	Measure enables scientific inquiry related to the construct.
	Consequential validity	Measure's eventual usage amounts to desirable social impact.
Relability	**Inter-annotator agreement**	Measurements are stable up to difference in annotators.
	Sensitivity	Measurements are stable up to difference in (hyper)parameters.

Table 2: Definitions for the 8 measurement modeling criteria we test for in our testing protocol.

Wallach (2021)). For each criteria, we systematically build tests: each test provides incremental evidence for trust, and measures that fare well under all tests accrue considerable evidence to trust them. Similar approaches have been used to verify the celebrated Implicit Association Test (e.g. Greenwald, McGhee, and Schwartz 1998; Greenwald and Nosek 2001; Nosek, Greenwald, and Banaji 2007), among other prominent measures (e.g. Jacobs and Wallach 2021).

Experimental Details. Along with testing our measure, we test measures from prior work, so we reuse the target concepts, social groups, and word lists from prior work. For target concepts, we follow Garg et al. (2018), drawing upon 104 professions tracked in the US Census (Levanon, England, and Allison 2009). For social groups, we follow Garg et al. (2018), considering either binary gender (female, male) or three-class race/ethnicity (White, Hispanic, Asian). For word lists, we follow Bommasani, Davis, and Cardie (2020). In several experiments, we report correlations: Spearman ρ to measure monotonicity, Pearson R^2 to measure linearity, and **bold** to indicate statistic significance for $p \leq 0.05$.

Testing Protocol for Validity

Face validity requires that the measure passes the "sniff test" (Jacobs and Wallach 2021). To validate our measures in this aspect, we measure gender bias for strongly gender-stereotyped professions (based on heavily imbalanced labor statistics in the 2000 US Census). We quantify associations in (i) *text* (English Wikipedia), (ii) *embeddings* (Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014)) and (iii) *contextualized representations* (final layer of BERT-base (Devlin et al. 2019) applied to English Wikipedia). To measure bias, we juxtapose these observed associations with reference associations of the uniform distribution (i.e. professions being equally associated with both the female and male gender). For all professions, across all settings, the measurements align with prevalent US stereotypes, except for *librarian* in settings involving English Wikipedia. While currently female-stereotyped, the male-leaning measurement is justifiable: most librarians discussed in Wikipedia refer to high-ranking posts (e.g. Librarian of Congress) historically filled mostly by men. See the arXiv version of the paper for the results.[4]

Content validity requires the measure to reflect theoretical understanding of bias; the measure's structure should match bias's structure (*structural fidelity*; Loevinger 1957). Given the high-fidelity correspondence between our social bias definition, derived from stated principles, and our framework DivDist, we conclude our measures demonstrate strong content validity.

Convergent validity requires that the proposed measure patterns similarly to other measures of the same construct (Campbell and Fiske 1959). As Jackman (2008) writes, convergence is only valuable if prior measures are known to be trustworthy. Since no prior measure has been subject to rigorous testing, and measures produce drastically different outcomes (Bommasani, Davis, and Cardie 2020), we cannot apply this criterion.

Instead, we reinterpret convergent validity for our text bias measures. Specifically, we introduce two bias measures for text, based on either human or automated judgments (i.e. cooccurrence). Since we consider human judgments to be ideal, we report the correlation between the human and automated measures. See the arXiv version of the paper for the results.[5] Specifically, we measure binary gender bias for eight professions in English Wikipedia with the uniform distribution as the reference. Additionally, we hypothesized

[4]https://arxiv.org/abs/2212.11672

[5]https://arxiv.org/abs/2212.11672

human annotators make more holistic judgments based on context, whereas automated cooccurrence would be more brittle, so we consider the impact of context length. We observe strong correlations for all context lengths: we report subsequent results using 3-sentence contexts, since the strongest correlations occur in this setting.

Predictive validity considers whether the measure is predictive of measures of related constructs. Since social bias is attributed to domain-general cognitive processes (Tajfel 1969), we expect that human biases will manifest similarly across different human behaviors. Consequently, biases in linguistic performance (e.g. writing) should predict biases in decision-making (e.g. employment).[6]

As a first experiment (**Diachronic**), we report the correlation between (i) the average bias for 104 Census professions in Word2Vec embeddings trained on corpora from each decade of 1900–2000 (Hamilton, Leskovec, and Jurafsky 2016) and (ii) bias in US labor statistics for the corresponding decades (Levanon, England, and Allison 2009). As a second experiment (**Contemporary**), we report the correlation between bias measurements for each of the 104 professions based on (i) our measurements for contemporary Word2Vec embeddings and (ii) the 2010 Census labor statistics. See the arXiv version of the paper for the results.[7]

We report these correlations when the bias in word embeddings are measured using our bias measure, as well as when the bias in word embeddings are measured using bias measures introduced in prior work. (We refer the reader to the cited works for the precise mathematical definitions of these measures.) All bias measures are computed the same word lists, hence any differences in correlation are strictly attributable to the differences in the mathematical form of the bias measures. Our measures across all settings consistently track biases in hiring practices, with statistically significant correlations in all cases. In other words, using our measure to quantify bias in word embeddings yields measurements that strongly correlate with simultaneous demographic trends in US employment.

In contrast, in many cases we find that other bias measures measures (e.g. Bolukbasi et al. 2016; Ethayarajh, Duvenaud, and Hirst 2019) do not. To better understand why some other measures may poorly predict demographic labor biases, we inspected the mathematical form for the measures. We find a clear conceptual separation: the measures of Bolukbasi et al. (2016) and Ethayarajh, Duvenaud, and Hirst (2019) are based on principal component analysis (PCA), whereas the measures of Garg et al. (2018) and our measure rely on averaging. As a result, we believe the first principal component in particular may be unreliable when considering fine-grained trends like these.

In fact, there is no bias measure we consider, except our own, that can measure bias in all settings (i.e. most of the prior measures cannot handle the multi-group case for race) and that is correlated in all settings. Given that the measure of Manzini et al. (2019) is the only other multi-class measure we considered, we find it especially striking that it yields measurements that are strongly *anti-correlated* with employment practices. We return to this measure subsequently, showing it lacks content validity (i.e. is structurally unfaithful to the construct of bias).

Hypothesis validity requires the measure be useful for addressing scientific hypotheses. We study *bias amplification* and *bias mitigation*, since both are central to the social impact of language technologies.

Bias Amplification. For bias amplification, we test whether training language models, as well as generating text using language models, increases bias. There is a prevalent hypothesis that model training generally increases bias, with some evidence of this in particular settings in NLP (Zhao et al. 2017; Jia et al. 2020). To test this hypothesis, we consider GPT-2 medium (Radford et al. 2019), a publicly available language model, and contrast the associations in GPT-2's training data $\mathcal{L}_1$, GPT-2's contextualized representations $\mathcal{L}_2$ (taken from the final layer), and machine-generated text $\mathcal{L}_3$ sampled from GPT-2.[8] Due to the stochasticity involved in sampling, we use a large sample of 250000 unconditional generations from GPT-2.[9] This experiment highlights the benefits of relative bias measurement, i.e. requiring an explicit reference, as the effects of processes (training, sampling) can be directly measured. We use our automated method to measure associations in the (human-authored) training corpus $\mathcal{L}_1$ and machine-generated text corpus $\mathcal{L}_3$; we use probing to measure associations in the contextualized representations $\mathcal{L}_2$ when applied to $\mathcal{L}_1$. See the arXiv version of the paper for the results.[10] We show that representation learning in GPT-2 amplifies gender biases relative to the training data, but that much of this bias does not manifest during generation. Surprisingly, machine-generated text from GPT-2 is measured to be marginally less gender biased than the data used to train GPT-2, which complicates the prevalent hypothesis that learning reliably amplifies the bias in the training data.

Bias Mitigation. Most "debiasing" methods target word embeddings, generally by directly optimizing a bias measure to provably guarantee bias reduction under that measure (e.g. Bolukbasi et al. 2016; Zhao et al. 2018b; Manzini et al. 2019). This brings to mind Strathern's law: *"When a measure becomes a target, it ceases to be a good measure"* (Strathern 1997; Goodhart 1984). Since we have provided significant evidence to trust our measure, we report how mit-

[6]We clarify that our analyses are strictly correlation-based and not causal. Further, perfect predictability is not expected, since it is reasonable that biases in text and hiring are not perfectly correlated, but we do expect significant correlation.

[7]https://arxiv.org/abs/2212.11672

[8]We use GPT-2 because it is the only language model we are aware of with (i) public training data, (ii) public model weights, and (iii) public canonical samples. Namely, many recent models do not release model weights (e.g. GPT-4, Gemini) or training data (e.g. Llama 2, Mistral); we are aware of no other model with large-scale collection of public samples decoded from the model.

[9]https://github.com/openai/gpt-2-output-dataset; temperature = 1

[10]https://arxiv.org/abs/2212.11672

igation methods change bias according to both our measure and the measure considered in prior work. See the arXiv version of the paper for the results.[11] Each results corresponds to (i) a set of pretrained word embeddings that were used in the work introducing (ii) the debiasing method to reduce bias against the (iii) listed groups. While every method reduces bias for the targeted measure, we find that for seven of the eight methods, bias is not reduced and is instead amplified according to ours. Our findings significantly strengthen existing findings that "debiasing" methods are quite limited (e.g. Gonen and Goldberg 2019): how bias is measured can change, and in many cases invert, judgments about the efficacy of bias mitigation methods.

Consequential validity emphasizes the eventual usage and impact of the measure (Messick 1988). While most of these consequences will be determined in the future, our bias measures have already been adopted as the default bias metrics in the HELM benchmark by Liang et al. (2022) to evaluate 30+ prominent language models. We will monitor our measures to revisit this question once further evidence accrues on their impact.

Testing Protocol for Reliability

Inter-annotator agreement is required for measures to be reliable (Jackman 2008). While the majority of our measures are fully automated, we do introduce a method to measure associations in text based on human judgments. To estimate the inter-annotator agreement, we recruit 5 NLP researchers (unaffiliated with the project) to annotate 40 contexts for binary gender with the targets being the eight professions used throughout. We report a very high inter-annotator agreement of Fleiss' $\kappa = 0.79$ (Landis and Koch 1977) for this task.

Sensitivity is not a standard criteria in measurement modeling, to our knowledge, but since our measures involve several parameters/inputs, we quantify sensitivity by perturbing each one. In particular, several works (Ethayarajh, Duvenaud, and Hirst 2019; Antoniak and Mimno 2021) shows prior bias measures are highly sensitive to word list perturbations. However, we find that all of our measures are quite stable to variations in the word lists, normalization function, and distance function. See the arXiv version of the paper for the results.[12]

Related Work

Text. In the social sciences, work across many disciplines has qualitatively characterized social bias in specific corpora of interest (e.g. Blumberg 2007; Atir and Ferguson 2018). While several quantitative measures have recently been proposed (Rudinger, May, and Van Durme 2017; Bordia and Bowman 2019; Field and Tsvetkov 2020; Falenska and Çetinoğlu 2021; Sun and Peng 2021; Mitchell et al. 2022), to our knowledge, these methods have neither been significantly adopted to facilitate social science research nor to measure bias in NLP datasets. We find this surprising given especially how large text corpora have been instrumental to the rise of language models in the field (Peters et al. 2018; Devlin et al. 2019; Brown et al. 2020, *inter alia*), alongside growing broader interest in dataset documentation and governance (Caswell et al. 2021; Bandy and Vincent 2021; Dodge et al. 2021; Bender and Friedman 2018; Gebru et al. 2021; Jernite et al. 2022). For this reason, we apply our measures to bias measurement on both sides of language modeling: the initial human-authored training corpora as well as the final machine-generated samples, and our measures have been similarly applied in the HELM benchmark for many language models and use cases (Liang et al. 2022). Mechanically, our bias measures for text, as well as other bias measures for text (e.g. Bordia and Bowman 2019), bear strong resemblance to the estimates of mutual information introduced by Church and Hanks (1989).

Representations. Bolukbasi et al. (2016) initiated the study of bias measurement for word embeddings, with a growing collection of such measures (e.g. Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Ethayarajh, Duvenaud, and Hirst 2019; Manzini et al. 2019; Du, Wu, and Lan 2019; Kumar et al. 2020). More recently, these measures have been adapted to measure bias in contextualized representations, generally by reducing measurement to the word embedding setting (Bommasani, Davis, and Cardie 2020), either by specifying a singular canonical context (May et al. 2019; Tan and Celis 2019; Ross, Katz, and Barbu 2021) or averaging representations across many contexts (Bommasani, Davis, and Cardie 2020; Guo and Caliskan 2020; Steed and Caliskan 2021). In comparison to prior measures for representations, we delineate the following differences. First, our measures are the only existing measures that are directly constructed under a unified framework for text and representation bias measurement. As we show, this enable us to study the effects of training (a transformation from text to representations) and generation (a transformation from representations to text). Second, all of our measures permit multiclass bias measurement, which is necessary given the underlying social categories are generally non-binary. To our knowledge, the measure of Manzini et al. (2019) (and measures that directly extend it) is the only prior measure that also extends to the multiclass setting.

Given this, we further examined this measure to understand the difference between it and our measures. Empirically, we found our measure was highly correlated with both diachronic and contemporary trends in employment, whereas the measure of Manzini et al. (2019) was either uncorrelated or anti-correlated, indicating it lacks predictive validity. Further, we found that mitigation methods that successfully optimize for the metric of Manzini et al. (2019) always increase bias under our method, independent of the specific optimization method (hard or soft) and the groups considered (i.e. gender, race, religion). Tracing this to the mathematical definition, we find the measure of Manzini et al. (2019) lacks content validity (which likely explain the above empirical findings). As a minimal example, con-

[11] https://arxiv.org/abs/2212.11672

[12] https://arxiv.org/abs/2212.11672

sider that the binary gender bias according to Manzini et al. (2019)'s measure for the concept *scientist*, using the word lists {*man*} and{*woman*}, is proportional to:

$$\cos(\mathbf{scientist}, \mathbf{man}) + \cos(\mathbf{scientist}, \mathbf{woman})$$

This fails to meet the criteria of content validity and structural fidelity, as it is not faithful to the underlying construct of social bias: social bias is proportional to (as codified in all other measures) the difference in the associations, not the sum.

Other settings. In addition to measuring bias in text and language representations, several recent works investigate biases in language models via the probabilities they assign to specific words or sequences (Kurita et al. 2019; Nangia et al. 2020; Nadeem, Bethke, and Reddy 2021). Since language modeling is currently the premier means for representation learning (Devlin et al. 2019; Bommasani et al. 2021), there is a natural question regarding the relationship between measuring biases of a pretrained language model and of representations induced by a pretrained language model.[13] In our work, since we are motivated by the potential downstream harms of language technologies, we elect to measure biases in representations as 1) it is the representations that are used downstream and 2) some biases may not manifest in sequence probabilities, but are latently present in the representations, and therefore may still manifest in downstream settings. To be more explicit, if some biases in the representations remain "dormant" and do not appear during generation (which is precisely what we saw in our experiments with GPT-2 overamplification), they will be invisible in these behavioral evaluations of language models. Nonetheless, these biases could observably affect model behavior once the language model is fine-tuned for downstream tasks, which is likely where the most concerning harms arise.

Further downstream, fairness evaluations exist for specific tasks such as machine translation (Stanovsky, Smith, and Zettlemoyer 2019; Escudé Font and Costa-jussà 2019; Prates, Avelar, and Lamb 2019), text generation (Sheng et al. 2019; Gehman et al. 2020; Dhamala et al. 2021; Lucy and Bamman 2021), coreference resolution (Rudinger et al. 2018; Zhao et al. 2018a; Cao and Daumé III 2020), sentiment analysis (Kiritchenko and Mohammad 2018), relation extraction (Gaut et al. 2020), and question answering (Parrish et al. 2021).[14] Given the existing paradigm of upstream pretraining and downstream adaptation/fine-tuning, future work should investigate the predictive validity of upstream bias measures at predicting downstream bias measures (Goldfarb-Tarrant et al. 2021; Jin et al. 2021).

Discussion of Measurement Modeling

We stress test our measures using measurement modeling, an interdisciplinary theory with a long history (Loevinger 1957; Messick 1987; Jackman 2008). Our work joins a growing collection of recent works that embrace measurement modeling in computational and AI contexts (Jacobs and Wallach 2021; Milli, Belli, and Hardt 2021; Blodgett 2021). For social bias in NLP, recent works use measurement modeling to identify failures in the validity (Blodgett et al. 2021) and reliability (Zhang, Sneyd, and Stevenson 2020; Du, Fang, and Nguyen 2021) of existing bias measures. In contrast, our work is the first to argue for the trustworthiness of social bias measures based on testing via measurement modeling. With that said, we emphasize that this does not unequivocally cement the trustworthiness of our measures, especially in contexts they have not been tested in: we have shown our measures pass the tests we introduce, but there certainly may be (and likely are) others that would demonstrate their weaknesses.

Beyond social bias, we believe measurement modeling can be a powerful general-purpose method in NLP in contexts where measurement/evaluation may be hard (but trust in evaluation is critical). To briefly demonstrate this, we enumerate several instances where existing work that studies evaluation in a particular context can be reinterpreted as referring to one (or more) of the criteria in measurement modeling. Critically, none of these works leverage either the specific language, or broader theory, of measurement modeling, but they can all be unified under this lens. In natural language generation evaluation, numerous works (e.g. Zhang et al. 2020; Sellam, Das, and Parikh 2020; Hessel et al. 2021; Pillutla et al. 2021) argue for their metrics to be used in place of existing metrics like BLEU (Papineni et al. 2002), because they more faithfully capture the semantics of language compared to the brittle overlap-based BLEU, and/or they are more correlated with human judgments. In essence, these works are arguing for the content validity and/or the convergent validity of their metrics. In the analysis of explainability methods, Jacovi and Goldberg (2020) argue several methods improperly conflate the plausibility and faithfulness of evaluations, which can be understood as a failure in the content validity of these methods. And, in the evaluation of word embeddings, Chiu, Korhonen, and Pyysalo (2016) and Rogers, Hosur Ananthakrishna, and Rumshisky (2018) show intrinsic evaluations (e.g. word analogy tests, word similarity/relatedness) do not reliably correlate with extrinsic evaluations of downstream outcomes (e.g. the performance of models built using these embeddings), indicating they lack predictive validity. Ravichander, Belinkov, and Hovy (2021) provide similar results for the intrinsic evaluations of syntactic understanding versus downstream behavior on entailment tasks.

More generally, measurement modeling provides a battle-tested set of well-studied desiderata, which can be used to standardize how we evaluate measures in NLP. In particular, while the criteria in measurement modeling are unlikely to be truly exhaustive, they do represent a comprehensive taxonomy of what properties are important for a measure to satisfy. In practice, we imagine this would yield an explicit protocol for accruing trust in a measure/evaluation by subjecting the measure/evaluation to a battery of tests (cf. the software engineering tests of Ribeiro et al. 2020).

[13] This mirrors the distinction between behavioral and representational methods in interpretability (Belinkov 2021).

[14] See Czarnowska, Vyas, and Shah (2021) for a summary.

Limitations

In this work, we center the trust of social bias measures. Consequently, we make explicit that while our testing suggests reasons to trust our measures, providing greater rigor than all prior work, nonetheless there likely exist other reasons to distrust our measures. Namely, the extend to which our measures are reliable (e.g. to rare words in word lists, to broader perturbations) and generalizable (e.g. to demographic categories beyond our narrow treatments of race and gender, to texts/representations beyond Wikipedia and similar corpora, to languages beyond English) is still incompletely understood. Consequently, this bears on the applicability of our measures to broader settings (e.g. while they they are predictive of trends in hiring, their utility for computational social science applications and fine-grained analysis is underexplored) as well as to their primary uses in NLP for bias mitigation. For this reason, we deliberately do not state optimization procedures to optimize these measures as, while we do trust their ability to measure bias, we are skeptical that direct optimization will lead to less harmful/-more just language technology. In this vein, we recommend caution in employing these measures to understand consequential systems until they have been even more strenuously tested.

Beyond questions of trust, we also note that measurement itself is limited and could benefit from alternative approaches. While we have not observed such approaches achieving widespread adoption in NLP, we believe more grounded approaches (e.g. user studies that document the harms of specific language technologies in specific contexts) are necessary. Namely, we believe the connection between social bias as it is understood in NLP remains tenuously connected with precise accounts of the harms of language technology. Already, we have seen this in the poor correlations between upstream bias measures and downstream bias measures. More generally, we expect bias measurement in the style we advance in this research, even with its sensitivity to additional factors unconsidered in prior work (e.g. normative positions/references, the corpus in the contextualized representation setting), still is likely to diverge for more user-centric analyses of specific technology (e.g. in the style of certain traditions in sociology, anthropology, human-computer interaction). To this end, we hope our work on adopting measurement modeling from the social sciences can be a step on a broader trajectory towards more mixed-methods approaches to social bias in NLP.

Conclusion

In this work, we foreground trust in social bias measurement: how do we accrue the evidence necessary to warrant trusting bias measures? Trustworthy bias measures are integral for making progress on broader goals (e.g. harm reduction through bias mitigation), which are of increasing consequence as the footprint of language technology and NLP grows. Our work contributes (i) a general measurement framework `DivDist` to measure bias, based on principles in social science, along with (ii) a testing protocol based on measurement modeling. Together, this makes the case for our social bias measures being trustworthy. However, as Messick (1987, 1988) reminds us, the task of validating a measure is an ongoing process.

Reproducibility

All code required to use our measurement framework and instantiated measures, as well as to replicate our experiments/tests, will be made publicly available on GitHub under the (open-source, permissive) MIT license. All data, including word lists (Antoniak and Mimno 2021), that we use is publicly available.

Acknowledgements

Thanks to Kawin Ethayarajh, Sidd Karamcheti, Nelson Liu, Claire Cardie, Su Lin Blodgett, Shyamal Buch, Lisa Li, Tianyi Zhang, Xikun Zhang, Shiori Sagawa, Michael Xie, Ananya Kumar, Deep Ganguli, Tatsu Hashimoto, Dan Ho, and members of p-lambda, Stanford NLP and Cornell NLP for feedback on this work. RB was supported by an NSF Graduate Research Fellowship, under grant number DGE-1656518. Other funding was provided by a PECASE award to PL.

Ethical Concerns

By its nature, our work on bias measurement runs the risk of ethics washing, especially if our measures are used inappropriately. Namely, while this work presents reasons to trust these measures, they likely should not be directly optimized (see Reich, Sahami, and Weinstein 2021) as a means for directly "guaranteeing" so-called "unbiased" NLP models. We see this as the central ethical risk of this work. Beyond this, we believe the themes of trust/carefulness in the study of bias in NLP should be adopted more broadly, and we believe measurement modeling is an effective means for operationalizing this in many contexts.

Research Positionality

The researchers involved in this work are situated at academic institutions: in large part, they are motivated to conduct this work by deficiencies they see in the scientific literature on social bias in NLP, but also the inadequate emphasis placed on social bias in the development/deployment of widespread language technology.

Adverse Impact

We do not foresee any significant adverse impacts of this work. While we prophylactically raise concerns of bias/ethics-washing in the above statement on ethical concerns, we believe the ultimate likelihood and magnitude of such adverse impacts are unlikely to be of significant concern if they materialize. These are the only adverse impacts we can currently envision.

References

Alain, G.; and Bengio, Y. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations – Workshop Track*.

Allport, G. W. 1954. *The Nature of Prejudice*. Addison-Wesley Publishing Company. ISBN 9780201001754.

Antoniak, M.; and Mimno, D. 2021. Bad Seeds: Evaluating Lexical Methods for Bias Measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1889–1904. Online: Association for Computational Linguistics.

Aribandi, V.; Tay, Y.; and Metzler, D. 2021. How Reliable are Model Diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1778–1785. Online: Association for Computational Linguistics.

Atir, S.; and Ferguson, M. J. 2018. How gender determines the way we speak about professionals. *Proceedings of the National Academy of Sciences*, 115(28): 7278–7283.

Bandy, J.; and Vincent, N. 2021. Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus. *ArXiv*, abs/2105.05241.

Belinkov, Y. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*.

Bender, E. M.; and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Beukeboom, C. J.; and Burgers, C. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7: 1–37.

Blodgett, S. L. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015. Online: Association for Computational Linguistics.

Blumberg, R. L. 2007. *Gender Bias in Textbooks: A Hidden Obstacle on the Road to Gender Equality in Education*. UNESCO.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Lee, D. D.; Sugiyama, M.; Luxburg, U. V.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29*, 4349–4357. Curran Associates, Inc.

Bommasani, R.; Creel, K.; Kumar, A.; Jurafsky, D.; and Liang, P. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Bommasani, R.; Davis, K.; and Cardie, C. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. Online: Association for Computational Linguistics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N. S.; Chen, A.; Creel, K.; Davis, J.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L. E.; Goel, K.; Goodman, N. D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T. F.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M. S.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S. P.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y. H.; Ruiz, C.; Ryan, J. K.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K. P.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M. A.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; and Liang, P. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv*, abs/2108.07258.

Bordia, S.; and Bowman, S. R. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 7–15. Minneapolis, Minnesota: Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Campbell, D. T.; and Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2): 81.

Cao, Y. T.; and Daumé III, H. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4568–4595. Online: Association for Computational Linguistics.

Caswell, I.; Kreutzer, J.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A. A.; Subramani, N.; Sokolov, A.; Sikasote, C.; Setyawan, M.; Sarin, S.; Samb, S.; Sagot, B.; Rivera, C.; Gonzales, A. R.; Papadimitriou, I.; Osei, S.; Suarez, P. O.; Orife, I.; Ogueji, K.; Niyongabo, R. A.; Nguyen, T. Q.; Muller, M.; Muller, A.; Muhammad, S. H.; Muhammad, N.; Mnyakeni, A.; Mirzakhalov, J.; Matangira, T.; Leong, C.; Lawson, N.; Kudugunta, S.; Jernite, Y.; Jenny, M.; Firat, O.; Dossou, B. F. P.; Dlamini, S.; de Silva, N.; cCabuk Balli, S.; Biderman, S. R.; Battisti, A.; Baruwa, A.; Bapna, A.; Baljekar, P. N.; Azime, I. A.; Awokoya, A.; Ataman, D.; Ahia, O.; Ahia, O.; Agrawal, S.; and Adeyemi, M. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *ArXiv*, abs/2103.12028.

Chequer, S. 2014. *Evaluating the Construct Validity of Implicit Association Tests using Confirmatory Factor Analytic Models*. Ph.D. thesis, University of Tasmania, Australia.

Chiu, B.; Korhonen, A.; and Pyysalo, S. 2016. Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 1–6. Berlin, Germany: Association for Computational Linguistics.

Church, K. W.; and Hanks, P. 1989. Word Association Norms, Mutual Information, and Lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, 76–83. Vancouver, British Columbia, Canada: Association for Computational Linguistics.

Crenshaw, K. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, Vol.1989, Article 8.

Czarnowska, P.; Vyas, Y.; and Shah, K. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *ArXiv*, abs/2106.14574.

Delobelle, P.; Tokpo, E.; Calders, T.; and Berendt, B. 2022. Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1693–1706. Seattle, United States: Association for Computational Linguistics.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; and Chang, K.-W. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, 246–267. Online only: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 862–872. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; and Gardner, M. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Du, Y.; Fang, Q.; and Nguyen, D. 2021. Assessing the Reliability of Word Embedding Gender Bias Measures. *ArXiv*, abs/2109.04732.

Du, Y.; Wu, Y.; and Lan, M. 2019. Exploring Human Gender Stereotypes with Word Association Test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6133–6143. Hong Kong, China: Association for Computational Linguistics.

Escudé Font, J.; and Costa-jussà, M. R. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 147–154. Florence, Italy: Association for Computational Linguistics.

Ethayarajh, K. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65. Hong Kong, China: Association for Computational Linguistics.

Ethayarajh, K.; Duvenaud, D.; and Hirst, G. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705. Florence, Italy: Association for Computational Linguistics.

Falenska, A.; and Çetinoğlu, Ö. 2021. Assessing Gender Bias in Wikipedia: Inequalities in Article Titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural*

Language Processing, 75–85. Online: Association for Computational Linguistics.

Field, A.; and Tsvetkov, Y. 2020. Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 596–608. Online: Association for Computational Linguistics.

Friedman, B.; and Nissenbaum, H. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3): 330–347.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Gaut, A.; Sun, T.; Tang, S.; Huang, Y.; Qian, J.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E.; Chang, K.-W.; and Wang, W. Y. 2020. Towards Understanding Gender Bias in Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2943–2953. Online: Association for Computational Linguistics.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; III, H. D.; and Crawford, K. 2021. Datasheets for Datasets. *Commun. ACM*, 64(12): 86–92.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Online: Association for Computational Linguistics.

Goldfarb-Tarrant, S.; Marchant, R.; Muñoz Sánchez, R.; Pandya, M.; and Lopez, A. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1926–1940. Online: Association for Computational Linguistics.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. Minneapolis, Minnesota: Association for Computational Linguistics.

Goodhart, C. A. 1984. Problems of monetary management: the UK experience. In *Monetary Theory and Practice*, 91–121. Springer.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6): 1464.

Greenwald, A. G.; and Nosek, B. A. 2001. Health of the Implicit Association Test at Age 3. *Zeitschrift fur experimentelle Psychologie: Organ der Deutschen Gesellschaft fur Psychologie*.

Guo, W.; and Caliskan, A. 2020. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. *ArXiv*, abs/2006.03955.

Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1501. Berlin, Germany: Association for Computational Linguistics.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. J. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *ArXiv*, abs/2104.08718.

Hewitt, J.; Ethayarajh, K.; Liang, P.; and Manning, C. D. 2021. Conditional probing: measuring usable information beyond a baseline.

Hewitt, J.; and Liang, P. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743. Hong Kong, China: Association for Computational Linguistics.

Hovy, D.; and Spruit, S. L. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. Berlin, Germany: Association for Computational Linguistics.

Hovy, E. H.; and Prabhumoye, S. 2021. Five sources of bias in natural language processing. *Lang. Linguistics Compass*, 15.

Ivanova, A. A.; Hewitt, J.; and Zaslavsky, N. 2021. Probing artificial neural networks: insights from neuroscience. *ArXiv*, abs/2104.08197.

Jackman, S. 2008. *Measurement*. The Oxford Handbook of Political Methodology. Oxford Handbooks.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, FAccT '21. New York, NY, USA: Association for Computing Machinery.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. Online: Association for Computational Linguistics.

Jernite, Y.; Nguyen, H.; Biderman, S.; Rogers, A.; Masoud, M.; Danchev, V.; Tan, S.; Luccioni, A. S.; Subramani, N.; Johnson, I.; Dupont, G.; Dodge, J.; Lo, K.; Talat, Z.; Radev, D.; Gokaslan, A.; Nikpoor, S.; Henderson, P.; Bommasani, R.; and Mitchell, M. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 2206–2222. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Jia, S.; Meng, T.; Zhao, J.; and Chang, K.-W. 2020. Mitigating Gender Bias Amplification in Distribution by Posterior

Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2936–2942. Online: Association for Computational Linguistics.

Jin, X.; Barbieri, F.; Kennedy, B.; Mostafazadeh Davani, A.; Neves, L.; and Ren, X. 2021. On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3770–3783. Online: Association for Computational Linguistics.

Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53. New Orleans, Louisiana: Association for Computational Linguistics.

Kumar, V.; Bhotia, T. S.; Kumar, V.; and Chakraborty, T. 2020. Nurse is Closer to Woman than Surgeon? Mitigating Gender-Biased Proximities in Word Embeddings. *Transactions of the Association for Computational Linguistics*, 8: 486–503.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics.

Landis, J. R.; and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Levanon, A.; England, P.; and Allison, P. 2009. Occupational Feminization and Pay: Assessing Causal Dynamics Using 1950–2000 U.S. Census Data. *Social Forces*, 88(2): 865–891.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; R'e, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L. J.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N. S.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T. F.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models. *ArXiv*, abs/2211.09110.

Loevinger, J. 1957. Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, 3(3): 635–694.

Lucy, L.; and Bamman, D. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. Virtual: Association for Computational Linguistics.

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. Minneapolis, Minnesota: Association for Computational Linguistics.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Messick, S. 1987. Validity. *ETS Research Report Series*, 1987(2): i–208.

Messick, S. 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. *ETS Research Report Series*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 26, 3111–3119. Curran Associates, Inc.

Milli, S.; Belli, L.; and Hardt, M. 2021. From Optimizing Engagement to Measuring Value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 714–722. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Mitchell, M.; Luccioni, A. S.; Lambert, N.; Gerchick, M.; McMillan-Major, A.; Ozoani, E.; Rajani, N.; Thrush, T.; Jernite, Y.; and Kiela, D. 2022. Measuring Data.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. Online: Association for Computational Linguistics.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.

Nissim, M.; van Noord, R.; and van der Goot, R. 2020. Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. *Computational Linguistics*, 46(2): 487–497.

Nosek, B. A.; Greenwald, A. G.; and Banaji, M. R. 2007. The Implicit Association Test at age 7: A methodological and conceptual review. *Automatic processes in social thinking and behavior*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2021. BBQ: A Hand-Built Bias Benchmark for Question Answering. *ArXiv*, abs/2110.08193.

Penner, A. M.; and Saperstein, A. 2015. Disentangling the effects of racial self-identification and classification by others: The case of arrest. *Demography*, 52(3): 1017–1024.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics.

Pillutla, K.; Swayamdipta, S.; Zellers, R.; Thickstun, J.; Welleck, S.; Choi, Y.; and Harchaoui, Z. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Pimentel, T.; and Cotterell, R. 2021. A Bayesian Framework for Information-Theoretic Probing.

Pimentel, T.; Saphra, N.; Williams, A.; and Cotterell, R. 2020a. Pareto Probing: Trading Off Accuracy for Complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3138–3153. Online: Association for Computational Linguistics.

Pimentel, T.; Valvoda, J.; Hall Maudslay, R.; Zmigrod, R.; Williams, A.; and Cotterell, R. 2020b. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4609–4622. Online: Association for Computational Linguistics.

Prates, M. O. R.; Avelar, P. H. C.; and Lamb, L. 2019. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32: 6363–6381.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Ravichander, A.; Belinkov, Y.; and Hovy, E. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3363–3377. Online: Association for Computational Linguistics.

Reich, R.; Sahami, M.; and Weinstein, J. M. 2021. *System Error: Where Big Tech Went Wrong and How We Can Reboot*. HarperCollins. ISBN 9780063066205.

Ribeiro, M. T.; Wu, T.; Guestrin, C.; and Singh, S. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4902–4912. Online: Association for Computational Linguistics.

Rogers, A.; Hosur Ananthakrishna, S.; and Rumshisky, A. 2018. What's in Your Embedding, And How It Predicts Task Performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2690–2703. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2020. A Primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*.

Ross, C.; Katz, B.; and Barbu, A. 2021. Measuring Social Biases in Grounded Vision and Language Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 998–1008. Online: Association for Computational Linguistics.

Rudinger, R.; May, C.; and Van Durme, B. 2017. Social Bias in Elicited Natural Language Inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 74–79. Valencia, Spain: Association for Computational Linguistics.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 8–14. New Orleans, Louisiana: Association for Computational Linguistics.

Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881–7892. Online: Association for Computational Linguistics.

Seshadri, P.; Pezeshkpour, P.; and Singh, S. 2022. Quantifying Social Biases Using Templates is Unreliable. *ArXiv*, abs/2210.04337.

Shah, D. S.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264. Online: Association for Computational Linguistics.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–3412. Hong Kong, China: Association for Computational Linguistics.

Stanovsky, G.; Smith, N. A.; and Zettlemoyer, L. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. Florence, Italy: Association for Computational Linguistics.

Steed, R.; and Caliskan, A. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21,

701–713. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Steed, R.; Panda, S.; Kobren, A.; and Wick, M. 2022. Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3524–3542. Dublin, Ireland: Association for Computational Linguistics.

Strathern, M. 1997. 'Improving ratings': audit in the British University system. *European Review*, 5(3): 305–321.

Sun, J.; and Peng, N. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 350–360. Online: Association for Computational Linguistics.

Tajfel, H. 1969. Cognitive aspects of prejudice. *Journal of Biosocial Science*, 1(S1): 173–191.

Tan, Y. C.; and Celis, L. E. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 13230–13241. Curran Associates, Inc.

Voigt, R.; Camp, N. P.; Prabhakaran, V.; Hamilton, W. L.; Hetey, R. C.; Griffiths, C. M.; Jurgens, D.; Jurafsky, D.; and Eberhardt, J. L. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25): 6521–6526.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks Posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Weitzman, L. J.; Eifler, D.; Hokada, E.; and Ross, C. 1972. Sex-Role Socialization in Picture Books for Preschool Children. *American Journal of Sociology*, 77(6): 1125–1150.

Zhang, H.; Sneyd, A.; and Stevenson, M. 2020. Robustness and Reliability of Gender Bias Assessment in Word Embeddings: The Role of Base Pairs. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 759–769. Suzhou, China: Association for Computational Linguistics.

Zhang, T.; Kishore, V.; Wu*, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. Copenhagen, Denmark: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018b. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. Brussels, Belgium: Association for Computational Linguistics.

Zhelezniak, V.; Savkov, A.; Shen, A.; and Hammerla, N. 2019. Correlation Coefficients and Semantic Textual Similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 951–962. Minneapolis, Minnesota: Association for Computational Linguistics.