Trustworthy Social Bias Measurement

Rishi Bommasani
Stanford University
nlprishi@stanford.edu

Percy Liang
Stanford University
pliang@cs.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 NLP, 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.

1 Introduction

Language technologies are increasingly critical to our lives and to 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 a central consideration (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), and psychological (e.g. stereotype threat).

Measurement functions as the primary lens for understanding social bias in NLP. And measurement is seen as an essential to successfully 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 between 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 are subject to scrutiny: measures have been shown to be brittle (Ethayarajh et al., 2019; Nissim et al., 2020; Antoniak and Mimno, 2021; Delobelle et al., 2022), contradictory (Bommasani et al., 2020), unreliable (Aribandi et al., 2021; Seshadri et al., 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. Therefore, measurement modeling is well-suited to social bias measurement: the theory has a longstanding tradition for similar social constructs, 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 we are not restricted to binary bias measurement. 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).

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 bias measures, spanning measures for text, word embeddings, and contextualized representations. We put these measures to the test, alongside several prior measures. Measurement modeling specifies 8 well-studied desiderata: in a sense, measurement modeling provides a well-established checklist of criteria to trust 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 *exacerbate*, social bias according to our measure, which calls into question their meaningfulness (Gonen and Goldberg, 2019).

2 Principles for Social Bias

Notation. Following social science research, 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 gender biases in science with $G_1 = \text{female}, G_2 = \text{male}$ and $T = \text{scientist}$.

Reducing bias to associations. Given social groups and a target concept, some social science theories define bias as the target concept's differential *association* with each group. For example, in the Implicit Association Test (Greenwald et al., 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", and Friedman and Nissenbaum (1996) emphasize that social bias pertains to broader social groups, rather than particular individuals (cf. Bommasani et al., 2022).

2.1 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

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

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 et al. (2020) and Hovy and Prabhumoye (2021) attribute bias in NLP to several sources (e.g. data selection, data annotation, model training): effective bias measurement could hope to quantify the relative contribution of each of these sources.

2.2 Defining Social Bias

Having introduced groups, targets, associations, and references, we define social bias.

Definition 2.1 (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.

3 `DivDist` Measurement Framework

Having defined social bias, we propose our two-stage measurement framework `DivDist`. First, given *parameters*, which specify the associations of interest, `DivDist` yields a bias measure `bias`. Second, given *inputs* (i.e. the target concept, social groups, and reference mentioned in our distribution), `bias` yields a bias measurement $\texttt{bias}(T, G_1, \ldots, G_k; \mathbf{p}_0)$ (i.e. a numerical value of how much bias is present).

$$\mathbf{s} \triangleq [\texttt{SoA}(T, G_1), \ldots, \texttt{SoA}(T, G_k)] \quad (1)$$

$$\mathbf{p} \triangleq \texttt{normalize}(\mathbf{s}) \quad (2)$$

$$\texttt{bias}(T, G_1, \ldots, G_k; \mathbf{p}_0) = D(\mathbf{p}, \mathbf{p}_0) \quad (3)$$

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 further 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 $(T, G_1, \ldots, G_k; \mathbf{p}_0)$. In many cases, we will 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. Further, for the reference $\mathbf{p}_0$, we will specify it as a categorical probability distribution distributed over the k social groups, which encodes 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 (e.g. Weitzman et al., 1972; Voigt et al., 2017) and NLP (e.g. Caliskan et al., 2017; Garg et al., 2018) are special cases of `DivDist`. In all of these works, bias is measured in the binary setting as the difference in associations (i.e. how much is the *male* gender associated with *scientist* more than the *female* gender is associated with *scientist*). We show 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]

$$\underset{\text{prev}}{\texttt{bias}} = \texttt{SoA}(T, G_1) \; - \; \texttt{SoA}(T, G_2) \; = \; x - y$$

$$\begin{aligned} \underset{\text{ours}}{\texttt{bias}} &= D\left(\texttt{normalize}\left([x, y]\right), \left[\frac{1}{2}, \frac{1}{2}\right]\right) \\ &= \left\| \left[\frac{x}{x+y}, \frac{y}{x+y}\right] - \left[\frac{1}{2}, \frac{1}{2}\right] \right\|_1 \\ &= \frac{x-y}{x+y} \end{aligned}$$

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

4 Measures

To further demonstrate the generality of `DivDist`, we instantiate several bias measures using it. In NLP, we want to measure 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.

4.1 Text

Since social bias is a systematic phenomenon, bias in text manifests in distributional statistics. To implement $\mathsf{SoA}_{\text{text}}$, 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. $\mathsf{SoA}_{\text{text}}(T, G_j) = \sum_{i=1}^{N} a_{ij}$.

Contexts, mentions, and associations. The aforementioned procedure partially implements $\mathsf{SoA}_{\text{text}}$, 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]

[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.

4.2 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

$$\mathsf{SoA}_{\text{WE}}(T, G_j) = \cos\left(\mathbb{E}_{t \in W(T)}\mathbf{t}, \mathbb{E}_{g \in W(G_j)}\mathbf{g}\right).$$

In words, SoA_{WE} is the cosine similarity between the average target word embedding and average group word embedding, closely resembling how Caliskan et al. (2017) and Garg et al. (2018) quantify association.

4.3 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 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 SoA_{WE}. Our first approach reduces the contextualized case to the static case, following Bommasani et al. (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 et al. (2020) show this produces high-quality static embeddings from contextualized representations: once we have these static embeddings, we then apply the aforementioned SoA_{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 et al., 2020). Therefore, we consider more direct techniques for interpreting contextualized representations (see Rogers et al., 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) have humans 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 contexts c^{test} that mention T:

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

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 et al., 2021; Hewitt et al., 2021).

4.4 Normalization and Divergence Parameters

Beyond `SoA`, `DivDist` requires normalization `normalize` : $\mathbb{R}^k \rightarrow \Delta^{k-1}$ and divergence $D : \Delta^{k-1} \times \Delta^{k-1} \rightarrow \mathbb{R}_{\geq 0}$ 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 (§7).

5 Testing Protocol

We have put forth yet another bias measure(s); why should we trust it? 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 (see Table 2; we closely follow Jacobs and 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.

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 (see Appendix A). For target concepts, we follow Garg et al. (2018), drawing upon 104 professions tracked in the US Census (Levanon et al., 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 et al. (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 $p \leq 0.05$.

6 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). In Table 3, we quantify associations in English Wikipedia, Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embeddings, and BERT (final layer of BERT-base; Devlin et al., 2019) contextualized representations 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

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 in §4 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.

Wikipedia. While currently female-stereotyped in the US, the male-leaning stereotype in relation to Wikipedia is justifiable, as most librarians discussed in Wikipedia refer to high-ranking posts (e.g. Librarian of Congress, University Librarians) historically filled mostly by men.

	TEXT		EMB		CR	
	Human	Aut.	W2V	GLOVE	Red.	Probe
carpenter	-0.5	-0.368	-0.128	-0.05	-0.02	-0.384
dancer	0.167	0.039	0.078	0.086	0.035	0.09
librarian	-0.105	-0.275	0.177	0.124	-0.003	-0.333
nurse	0.373	0.097	0.119	0.114	0.066	0.111
pilot	-0.417	-0.265	-0.099	-0.072	-0.022	-0.33
soldier	-0.473	-0.358	-0.041	-0.065	-0.025	-0.389
businessman	-0.5	-0.341	-0.173	-0.145	-0.056	-0.232
businesswoman	0.5	0.453	0.174	0.385	0.058	0.5

Table 3: **Face validity experiment.** Female-directed gender bias for gender-stereotyped professions (**top**) and explicitly gendered professions (**bottom**) aligns with prevalent US stereotypes.

Content validity requires that the measure reflects theoretical understanding of the underlying construct; the measure's structure should match the construct's structure (*structural fidelity*; Loevinger, 1957). Given the clear and high-fidelity correspondence between our social bias definition (2.1), derived from stated principles in §2, and our framework `DivDist`, we argue 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) credible. Since no prior measure has been subject to rigorous and extensive testing, and measures have been shown to produce drastically different outcomes (Bommasani et al., 2020), we find this criterion does not apply directly.

Instead, we reinterpret convergent validity in the context of our measures for bias in text. Specifically, while there is no known credible bias measure for text, 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 in Table 4. Specifically, we measure binary gender bias for eight professions (those in Table 3) in the context of English Wikipedia with the uniform distribution as the reference. Additionally, because we hypothesized human annotators may make more holistic judgments based on context, whereas automated cooccurrence statistics may be more brittle, we also consider the impact of context length in Table 4. We observe strong correlations for all context lengths and report subsequent results using 3-sentence contexts, given that 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

Context Length (sent.)	ρ	R^2
1	**0.731**	**0.848**
2	**0.814**	**0.858**
3	**0.898**	**0.879**
4	**0.898**	**0.840**
5	**0.898**	**0.829**

Table 4: **Convergent validity experiment.** High correlations between human and automated text bias measures for all context lengths.

linguistic performance (e.g. writing) should predict biases in decision-making (e.g. employment).[4]

As a first experiment (**Diachronic**), we report the correlation between (i) the average bias for 104 professions considered in the US census in Word2Vec embeddings trained on corpora from each decade of 1900–2000 (Hamilton et al., 2016) and (ii) bias in US labor statistics for the corresponding decades (Levanon et al., 2009). As a second experiment (**Contemporary**), we report the correlation between bias measurements for each of 104 professions based on (i) our measurements for contemporary Word2Vec embeddings and (ii) the 2010 Census labor statistics. In Table 5, we see our measurements consistently track biases in hiring practices, with statisitically significant correlations, whereas several other measures (e.g. Bolukbasi et al., 2016; Ethayarajh et al., 2019) do not. We believe these results hinge on implementation differences for bias in embeddings: the highly-correlated measures all average embeddings/bias scores, whereas the weakly-correlated measures all use PCA. Strikingly, the measure of Manzini et al. (2019), which is the only other measure that generalizes to the multi-class setting, is strongly *anti-correlated* with diachronic/historical trends in employment for race and gender. We return to this measure in §8, showing it lacks content validity (i.e. is structurally unfaithful to to the construct of bias) that likely explains its poor predictive validity.

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 NLP.

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

[4]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.

	Diachronic		Contemporary	
	Gender	Race	Gender	Race
Bolukbasi et al. (2016)	0.261	N/A	0.047	N/A
Caliskan et al. (2017)	**0.709**	N/A	**0.505**	N/A
Garg et al. (2018, cosine)	**0.758**	N/A	**0.633**	N/A
Garg et al. (2018, euclidean)	0.127	N/A	**0.553**	N/A
Manzini et al. (2019)	**-0.648**	**-0.903**	0.193	**-0.396**
Ethayarajh et al. (2019)	0.261	N/A	0.065	N/A
Our Measure	**0.83**	**0.842**	**0.42**	**0.369**

Table 5: **Predictive validity experiments.** Our measures demonstrate high Spearman correlation with **diachronic** changes in labor statistics, as well as **contemporary** labor statistics, whereas some other measures do not.

	$\mathcal{L}_2; \mathcal{L}_1$	$\mathcal{L}_3; \mathcal{L}_2$	$\mathcal{L}_3; \mathcal{L}_1$
carpenter	0.176	-0.166	0.010
dancer	0.063	-0.143	-0.080
librarian	0.127	-0.146	-0.019
nurse	0.017	-0.136	-0.119
pilot	-0.063	0.066	0.003
soldier	-0.041	0.089	0.049
businessman	-0.021	0.066	0.044
businesswoman	0.118	-0.086	0.032

Table 6: **Hypothesis validity (amplification) experiment.** GPT-2's learned representations $\mathcal{L}_2$ amplify bias relative to training data $\mathcal{L}_1$ but much of this bias does not persist to (unconditional) machine-generated samples $\mathcal{L}_3$.

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. Due to the stochasticity involved in sampling, we use a large sample of 250000 unconditional generations from GPT-2.[5] 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 $\mathcal{L}_1$. Table 6 shows that representation learning in GPT-2 amplifies gender biases relative to the training data, but that much of this bias does not manifest during

[5]`https://github.com/openai/gpt-2-output-dataset`

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 in Table 7 how mitigation methods change bias according to both our measure and the measure considered in prior work. 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.

Emb.	Method	Groups	Targeted metric Original	Targeted metric Debiased	Our metric Original	Our metric Debiased
W2V	Hard (B)	*gender*	0.050	0.041	0.011	0.004
GLOVE	GN (Z)	*gender*	0.191	0.083	0.009	0.016
W2V	Soft (M)	*gender*	0.330	0.197	0.008	0.012
W2V	Hard (M)	*gender*	0.330	0.281	0.008	0.024
W2V	Soft (M)	*race*	0.026	-0.055	0.018	0.018
W2V	Hard (M)	*race*	0.026	0.005	0.018	0.023
W2V	Soft (M)	*religion*	0.253	0.126	0.023	0.024
W2V	Hard (M)	*religion*	0.253	0.217	0.023	0.074

Table 7: **Hypothesis validity (debiasing) experiment.** Debiasing methods generally reduce bias (green) for the targeted metric, but generally increase bias (red) for our metric. B indicates Bolukbasi et al. (2016), Z indicates Zhao et al. (2018b), M indicates Manzini et al. (2019); Hard/Soft/GN refer to specific debiasing 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, we have proactively implemented our bias measures as the default metrics in the HELM benchmark (Liang et al., 2022) to help accelerate this adoption process. Already our measures have been used to evaluate 30+ prominent language models to understand model biases across a range of different uses (Liang et al., 2022). We will monitor our measures to revisit this question once further evidence accrues on their impact.

7 Testing Protocol for Reliability

Inter-annotator agreement is required for measures to be reliable, though it generally refers to measures that involve human judgments (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 this work. 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 inputs and hyperparameters, we study how sensitive each measure is to perturbations of each of these. In particular, several works (Ethayarajh et al., 2019; Nissim et al., 2020; Antoniak and Mimno, 2021) shows prior bias measures are highly sensitive to word list perturbations. However, in Table 8, we find that all of our measures are quite stable to variations in the word lists, normalization function, and distance function.

Setting	$\mathcal{L}$	Word Lists $\lvert W(G)\rvert = 3$	Word Lists $\lvert W(G)\rvert = 5$	D ℓ_2	normalize Softmax
TEXT-AUT.	Wiki.	**(0.85, 0.89)**	**(0.88, 0.94)**	**(1.00, 1.00)**	**(0.84, 0.82)**
EMB.	W2V	**(0.92, 0.94)**	**(0.90, 0.93)**	**(1.00, 1.00)**	**(0.90, 0.95)**
EMB.	GloVe	**(0.88, 0.94)**	**(0.90, 0.95)**	**(1.00, 1.00)**	**(0.90, 0.90)**
CR-RED.	BERT	**(0.90, 0.98)**	**(1.00, 0.99)**	**(1.00, 1.00)**	**(1.00, 1.00)**
CR-PROBE	BERT	N/A	N/A	**(1.00, 1.00)**	**(0.79, 0.71)**

Table 8: **Sensitivity experiment.** Perturbing any single parameter/input yields stable results for our measures, when compared to the default parameters, based on (ρ, R^2) correlations. For word lists, we subsample lists to the specified size, similar to Ethayarajh et al. (2019).

8 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 et al., 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 et al., 2017; Garg et al., 2018; Ethayarajh et al., 2019; Manzini et al., 2019; Du et al., 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 et al., 2020), either by specifying a singular canonical context (May et al., 2019; Tan and Celis, 2019; Ross et al., 2021) or averaging representations across many contexts (Bommasani et al., 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 in Table 6, 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, in Table 5, 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, in Table 7, 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, consider 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 et al., 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.[6] 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 genera-

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

tion (which is precisely what we saw in our experiments with GPT-2 in Table 6), 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 et al., 2019; Escudé Font and Costa-jussà, 2019; Prates et al., 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).[7] 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).

9 Discussion of Measurement Modeling

In §5, 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 et al., 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 et al., 2020a; Du et al., 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., 2020b; Sellam et al., 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 et al. (2016) and Rogers et al. (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 et al. (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).

10 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

[7] See Czarnowska et al. (2021) for a summary.

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 a general measurement framework `DivDist` to measure bias, based on principles in social science, along with a testing protocol based on measurement modeling. Together, this makes the case for our social bias measures being trustworthy. However, as Messick (1987, 1988) explains, the task of validating a measure is an ongoing process: from the consequentialist perspective, it will be the use of our measures that determines their value.

11 Reproducibility

All code is made available at `https://github.com/rishibommasani/BiasMeasures` with further details on data/sources in Appendix A. We aim to release further tooling to facilitate adoption of our measures in future work along with documentation of the impact of our measures over time, consistent with the discussion of consequential validity.

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.

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations – Workshop Track*.

Gordon W. Allport. 1954. *The Nature of Prejudice*. Addison-Wesley Publishing Company.

Maria Antoniak and David Mimno. 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)*, pages 1889–1904, Online. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785, Online. Association for Computational Linguistics.

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

Jack Bandy and Nicholas Vincent. 2021. Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *ArXiv*, abs/2105.05241.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*.

Emily M. Bender and Batya Friedman. 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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, page 610–623, New York, NY, USA. Association for Computing Machinery.

Camiel J. Beukeboom and Christian Burgers. 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.

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

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 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*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 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)*, pages 1004–1015, Online. Association for Computational Linguistics.

Rae Lesser Blumberg. 2007. *Gender Bias in Textbooks: A Hidden Obstacle on the Road to Gender Equality in Education*. UNESCO.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Rishi Bommasani, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, D. Card, Rodrigo Castellon, Niladri S. Chatterji, Annie Chen, Kathleen Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jackson K. Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258.

Shikha Bordia and Samuel R. Bowman. 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*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

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

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

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios Gonzales, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Muller, Andr'e Muller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, M. Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine cCabuk Balli, Stella Rose Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi N. Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *ArXiv*, abs/2103.12028.

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

Billy Chiu, Anna Korhonen, and Sampo Pyysalo. 2016. Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 1–6, Berlin, Germany. Association for Computational Linguistics.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Kimberlé Crenshaw. 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.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *ArXiv*, abs/2106.14574.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 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*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 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, page 862–872, New York, NY, USA. Association for Computing Machinery.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 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*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yupei Du, Qixiang Fang, and Dong Nguyen. 2021. Assessing the reliability of word embedding gender bias measures. *ArXiv*, abs/2109.04732.

Yupei Du, Yuanbin Wu, and Man Lan. 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)*, pages 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 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*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh. 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)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing gender bias in Wikipedia: Inequalities in article titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85, Online. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 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)*, pages 1926–1940, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 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)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L.K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464.

Wei Guo and Aylin Caliskan. 2020. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. *ArXiv*, abs/2006.03955.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 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)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Joseph Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. Conditional probing: measuring usable information beyond a baseline.

John Hewitt and Percy Liang. 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)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 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)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

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

Anna A. Ivanova, John Hewitt, and Noga Zaslavsky. 2021. Probing artificial neural networks: insights from neuroscience. *ArXiv*, abs/2104.08197.

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

Abigail Z. Jacobs and Hanna Wallach. 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.

Alon Jacovi and Yoav Goldberg. 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*, pages 4198–4205, Online. Association for Computational Linguistics.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2206–2222, New York, NY, USA. Association for Computing Machinery.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 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*, pages 3770–3783, Online. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Vaibhav Kumar, Tenzin Singhay Bhotia, Vaibhav Kumar, and Tanmoy Chakraborty. 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.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *ArXiv*, abs/2211.09110.

Jane Loevinger. 1957. Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3):635–694.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 714–722, New York, NY, USA. Association for Computing Machinery.

Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2022. Measuring data.

Moin Nadeem, Anna Bethke, and Siva Reddy. 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 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)*, pages 1953–1967, Online. Association for Computational Linguistics.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Sam Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *ArXiv*, abs/2110.08193.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Tiago Pimentel and Ryan Cotterell. 2021. A bayesian framework for information-theoretic probing.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and L. Lamb. 2019. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 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*, pages 3363–3377, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Shashwath Hosur Ananthakrishna, and Anna Rumshisky. 2018. What's in your embedding, and how it predicts task performance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2690–2703, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Candace Ross, Boris Katz, and Andrei Barbu. 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*, pages 998–1008, Online. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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), pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. *ArXiv*, abs/2210.04337.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 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*, pages 5248–5264, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 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)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Ryan Steed and Aylin Caliskan. 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, page 701–713, New York, NY, USA. Association for Computing Machinery.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 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)*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

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

Jiao Sun and Nanyun Peng. 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)*, pages 350–360, Online. Association for Computational Linguistics.

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

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Lenore J. Weitzman, Deborah Eifler, Elizabeth Hokada, and Catherine Ross. 1972. Sex-role socialization in picture books for preschool children. *American Journal of Sociology*, 77(6):1125–1150.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haiyang Zhang, Alison Sneyd, and Mark Stevenson. 2020a. 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*, pages 759–769, Suzhou, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 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)*, pages 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

A Reproducibility

A.1 Word Lists

We use social groups from Garg et al. (2018) and word lists from Bommasani et al. (2020), which we explicitly enumerate consistent with the recommendations of Antoniak and Mimno (2021).

`Female` word list = {she', 'daughter', 'hers', 'her', 'mother', 'woman', 'girl', 'herself', 'female', 'sister', 'daughters', 'mothers', 'women', 'girls', 'femen', 'sisters', 'aunt', 'aunts', 'niece', 'nieces' }

`Male` word list = {'he', 'son', 'his', 'him', 'father', 'man', 'boy', 'himself', 'male', 'brother', 'sons', 'fathers', 'men', 'boys', 'males', 'brothers', 'uncle', 'uncles', 'nephew', 'nephews' }

`Asian` word list = {'cho', 'wong', 'tang', 'huang', 'chu', 'chung', 'ng', 'wu', 'liu', 'chen', 'lin', 'yang', 'kim', 'chang', 'shah', 'wang', 'li', 'khan', 'singh', 'hong'}

`Hispanic` word list = {'castillo', 'gomez', 'soto', 'gonzalez', 'sanchez', 'rivera', 'martinez', 'torres', 'rodriguez', 'perez', 'lopez', 'medina', 'diaz', 'garcia', 'castro', 'cruz' }

`White` word list = {'harris', 'nelson', 'robinson', 'thompson', 'moore', 'wright', 'anderson', 'clark', 'jackson', 'taylor', 'scott', 'davis', 'allen', 'adams', 'lewis', 'williams', 'jones', 'wilson', 'martin', 'johnson' }

`Professions` word list = {'accountant', 'acquaintance', 'actor', 'actress', 'administrator', 'adventurer', 'advocate', 'aide', 'alderman', 'ambassador', 'analyst', 'anthropologist', 'archaeologist', 'archbishop', 'architect', 'artist', 'artiste', 'assassin', 'astronaut', 'astronomer', 'athlete', 'attorney', 'author', 'baker', 'ballerina', 'ballplayer', 'banker', 'barber', 'baron', 'barrister', 'bartender', 'biologist', 'bishop', 'bodyguard', 'bookkeeper', 'boss', 'boxer', 'broadcaster', 'broker', 'bureaucrat', 'businessman', 'businesswoman', 'butcher', 'cabbie', 'cameraman', 'campaigner', 'captain', 'cardiologist', 'caretaker', 'carpenter', 'cartoonist', 'cellist', 'chancellor', 'chaplain', 'character', 'chef', 'chemist', 'choreographer', 'cinematographer', 'citizen', 'cleric', 'clerk', 'coach', 'collector', 'colonel', 'columnist', 'comedian', 'comic', 'commander', 'commentator', 'commissioner', 'composer', 'conductor', 'confesses', 'congressman', 'constable', 'consultant', 'cop', 'correspondent', 'councilman', 'councilor', 'counselor', 'critic', 'crooner', 'crusader', 'curator', 'custodian', 'dad', 'dancer', 'dean', 'dentist', 'deputy', 'dermatologist', 'detective', 'diplomat', 'director', 'doctor', 'drummer', 'economist', 'editor', 'educator', 'electrician', 'employee', 'entertainer', 'entrepreneur', 'environmentalist', 'envoy', 'epidemiologist', 'evangelist', 'farmer', 'filmmaker', 'financier', 'firebrand', 'firefighter', 'fireman', 'fisherman', 'footballer', 'foreman', 'gangster', 'gardener', 'geologist', 'goalkeeper', 'guitarist', 'hairdresser', 'handyman', 'headmaster', 'historian', 'hitman', 'homemaker', 'hooker', 'housekeeper', 'housewife', 'illustrator', 'industrialist', 'infielder', 'inspector', 'instructor', 'inventor', 'investigator', 'janitor', 'jeweler', 'journalist', 'judge', 'jurist', 'laborer', 'landlord', 'lawmaker', 'lawyer', 'lecturer', 'legislator', 'librarian', 'lieutenant', 'lifeguard', 'lyricist', 'maestro', 'magician', 'magistrate', 'manager', 'marksman', 'marshal', 'mathematician', 'mechanic', 'mediator', 'medic', 'midfielder', 'minister', 'missionary', 'mobster', 'monk', 'musician', 'nanny', 'narrator', 'naturalist', 'negotiator', 'neurologist', 'neurosurgeon', 'novelist', 'nun', 'nurse', 'observer', 'officer', 'organist', 'painter', 'paralegal', 'parishioner', 'parliamentarian', 'pastor', 'pathologist', 'patrolman', 'pediatrician', 'performer', 'pharmacist', 'philanthropist', 'philosopher', 'photographer', 'photojournalist', 'physician', 'physicist', 'pianist', 'planner', 'playwright', 'plumber', 'poet', 'policeman', 'politician', 'pollster', 'preacher', 'president', 'priest', 'principal', 'prisoner', 'professor', 'programmer', 'promoter', 'proprietor', 'prosecutor', 'protagonist', 'protege', 'protester', 'provost', 'psychiatrist', 'psychologist', 'publicist', 'pundit', 'rabbi', 'radiologist', 'ranger', 'realtor', 'receptionist', 'researcher', 'restaurateur', 'sailor', 'saint', 'salesman', 'saxophonist', 'scholar', 'scientist', 'screenwriter', 'sculptor', 'secretary', 'senator', 'sergeant', 'servant', 'serviceman', 'shopkeeper', 'singer', 'skipper', 'socialite', 'sociologist', 'soldier', 'solicitor', 'soloist', 'sportsman', 'sportswriter', 'statesman', 'steward', 'stockbroker', 'strategist', 'student', 'stylist', 'substitute', 'superintendent', 'surgeon', 'surveyor', 'teacher', 'technician', 'teenager', 'therapist', 'trader', 'treasurer', 'trooper', 'trucker', 'trumpeter', 'tutor', 'tycoon', 'undersecretary', 'understudy', 'valedictorian', 'violinist', 'vocalist', 'waiter', 'waitress', 'warden', 'warrior', 'welder', 'worker', 'wrestler', 'writer' }

A.2 Data sources

English Wikipedia. We use the same subset of English Wikipedia that was used by Bommasani et al. (2020), which was chosen because it filtered for bot-generated content, and was sourced from https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/.

Contemporary Word Embeddings. We use standard static word embeddings: GloVe embeddings (Wikipedia 2014 + Gigaword 5, 300 dimensional) sourced from https://nlp.stanford.edu/projects/glove/ and Word2Vec embeddings (GoogleNews, 300 dimensional) sourced from https://code.google.com/archive/p/word2vec/.

Contextualized Representations. We use BERT and GPT-2 representations from the checkpoints made available through HuggingFace Transformers (Wolf et al., 2020).

GPT-2 related text. We use the sample of both GPT-2's training corpus (actually the identically distributed test) and its (unconditional) samples made available at https://github.com/openai/gpt-2-output-dataset.

US Census data. We use the US Census data of Levanon et al. (2009) that was also used by Garg et al. (2018), made available at https://github.com/nikhgarg/EmbeddingDynamicStereotypes/tree/master/data.

Historic Word Embeddings. We use the word embeddings trained on different decades in the 1900s from Hamilton et al. (2016) that were also used by Garg et al. (2018), made available at https://nlp.stanford.edu/projects/histwords/.