Sociodemographic Bias in Language Models: A Survey and Forward Path

Vipul Gupta[1] **Pranav Narayanan Venkit**[2] **Shomir Wilson**[2] **Rebecca J. Passonneau**[1]
[1] Depart. of Computer Science & Engineering, College of Engineering
[2] College of Information Sciences and Technology
Pennsylvania State University
{vkg5164, pranav.venkit, shomir, rjp49}@psu.edu

Abstract

Sociodemographic bias in language models (LMs) has the potential for harm when deployed in real-world settings. This paper presents a comprehensive survey of the past decade of research on sociodemographic bias in LMs, organized into a typology that facilitates examining the different aims: types of bias, quantifying bias, and debiasing techniques. We track the evolution of the latter two questions, then identify current trends and their limitations, as well as emerging techniques. To guide future research towards more effective and reliable solutions, and to help authors situate their work within this broad landscape, we conclude with a checklist of open questions.

1 Introduction

Language models (LMs) have demonstrated impressive performance across diverse tasks (Raffel et al., 2020; Zhong et al., 2020; Yang et al., 2019). However, much work reveals that LMs adopt biases present in training data (Wen et al., 2022; España-Bonet and Barrón-Cedeño, 2022; Gupta et al., 2022b; Hutchinson and Mitchell, 2019). Sociodemographic bias has been defined to occur when a model performs differently across social groups (Czarnowska et al., 2021; Chouldechova and Roth, 2020). When LMs are used in real-world applications, sociodemographic bias has the potential for negative societal impacts (Field et al., 2023; Rudin, 2019; Blodgett et al., 2020). With increasingly widespread usage, the urgency to understand and mitigate bias has grown. Fig. 1 shows an increasing rate of publications on LM bias over the past decade, sourced from SCOPUS. Our survey synthesizes results from this rapidly growing area into a roadmap for future investigations.

Other surveys on bias in NLP have thoroughly examined a particular aspect of bias, such as methods for measuring bias (Czarnowska et al., 2021; Bansal, 2022), or identification of gender bias (Stanczak and Augenstein, 2021; Devinney et al., 2022). Unlike previous surveys, we provide **a typology** of works on bias over the past decade. Further, we build upon foundational issues identified by Blodgett et al. (2020) by delving more deeply into methodological limitations, such as reliability issues. We also follow the recommendations of Blodgett et al. (2020) in consulting interdisciplinary approaches to improve the understanding of social bias. Thus we begin the survey with a discussion of psychosocial perspectives on benefits versus harms of bias. Our survey offers an **up-to-date** understanding of a topic that has been garnering increasing interest. Early in this literature, the development of bias mitigation or debiasing methods had questionable success; we argue that recent work using expert models during training shows particular promise. We conclude with a **checklist of key questions** that have continued to be challenging, to help steer future studies toward more effective and reliable methods, well-situated within the landscape of work on bias.

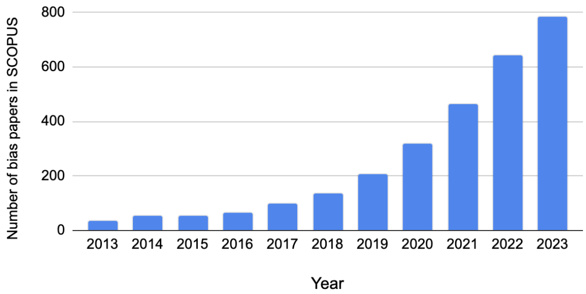

Figure 1: This graph shows number of papers/articles published each year (from 2013 to 2023) in SCOPUS that contain the term 'bias' and ('nlp' or 'language models') in the title, abstract, or keywords.

We present a synthesis of works on bias through three perspectives: **1) a taxonomic categorization, 2) an evolutionary timeline,** and **3) a roadmap for future work**. We categorized the surveyed works into three major strands of investigation, as shown in Fig. 2: types of bias, quantifying bias, and debiasing techniques. We then organize the findings

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 295–322
August 16, 2024 ©2024 Association for Computational Linguistics

within each category and subcategory of our taxonomy. In addition, we examined the evolution over the past decade of techniques for bias measurement and bias mitigation, as shown in Fig. 3. This perspective separates trends that had a brief life from those that continue to have promise.

While LM bias measurement and mitigation are critical for progress towards equitable LMs, understanding the potential for harm is deeply intertwined with social factors outside the scope of NLP proper. Thus we precede the presentation of the major types of bias research with a discussion of psychosocial perspectives (cf. Omrani et al., 2023; Mei et al., 2023). This is followed by a section describing our process for identifying candidate works, and our resulting typology where we place most of the surveyed works. Sections 4-6 present limitations, the checklist and future directions.

2 Understanding Bias

Interdisciplinary approaches to understanding bias as a psychosocial phenomenon have been argued to be important for clarifying how social harms might arise. Research into human cognition and social behavior can provide valuable insights on sociodemographic bias in LMs, as well as assessment of their potential for harm. For instance, research in psychology has long addressed the origins and expressions of social bias (Osborne et al., 2023) Recent studies have begun to integrate ideas from psychology with NLP to better understand bias (Spliethöver et al., 2022; Omrani et al., 2023; Mei et al., 2023; Omrani Sabbaghi et al., 2023), showcasing the usefulness of interdisciplinary approaches. For example, research in psychology proposes that reduction of social bias can be achieved by engaging with individuals from diverse groups (Pettigrew and Tropp, 2006; Reimer and Sengupta, 2023). A similar idea is reflected in Blodgett et al. (2020), which advocates for LM engineers to reduce bias through engagement with people who might be affected by applications that use LMs. One of the early works on quantifying bias - WEAT (Caliskan et al., 2017) used the Implicit Association Test from psychology (Greenwald et al., 1998) to develop a foundation bias metric for LMs.

The Stereotype Content Model (SCM), a framework from social psychology, categorizes stereotypes into interpersonal and intergroup interactions, providing insights into bias dynamics (Cuddy et al., 2008). It proposes that human stereotypes are captured by two social dimensions: warmth (e.g., trustworthiness, friendliness) and competence (e.g., capability, assertiveness). A recent work by Omrani et al. (2023) used the SCM framework to develop a bias mitigation method that generalizes across multiple social attributes, rather than one at a time.

The Nobel Prize-winning psychologist and behavioral economist, Daniel Kahneman, argues that mental shortcuts (biases) are advantageous in situations requiring quick judgments (Kahneman, 2011). For example, due to bias based on strong knowledge priors, the sentence "a large mouse climbed over a small elephant" will immediately call to mind the appropriate relative sizes; to counter this assumption would require extra information (Grice, 1975). Extrapolating Kahneman's argument to NLP, bias based on common-sense knowledge could be advantageous in enhancing an LM's understanding of relations among real-world entities. This argues for the potential benefit of certain kinds of bias.

Kahneman (2011) defines disadvantageous bias as "the tendency to make systematic errors in judgment or decisions based on factors that are irrelevant or immaterial to the task at hand" and cautions that human judgment is susceptible to bias from irrelevant factors. Turning to LM bias, we find previous NLP work aligned with Kahneman's perspective in definitions of *representational harm* (Crawford, 2017) and *alloted harm* (Barocas et al., 2017). Representational harm arises when an NLP system represents some social groups in a less favorable light than others. Allotted harm arises when a system allocates resources or opportunities unfairly to a social group (Shahbazi et al., 2023).

In conclusion, ideas from psychology and behavioral economics provide a more informed understanding of bias. While some biases might contribute positively to model performance, others can have detrimental societal effects. An interdisciplinary approach would not only enrich our theoretical understanding of bias but could also guide the development of more effective methods to identify undesirable LM bias and lessen social harm.

3 Categories of Work on Bias in LMs

We used two strategies to identify candidate papers for our survey: 1) using the keywords "bias", "fair" and "fairness", we searched for papers in the ACL Anthology, NeurIPS proceedings, FAccT, and AIES conferences; 2) we included papers from citation graphs for retrieved papers. We examined pa-

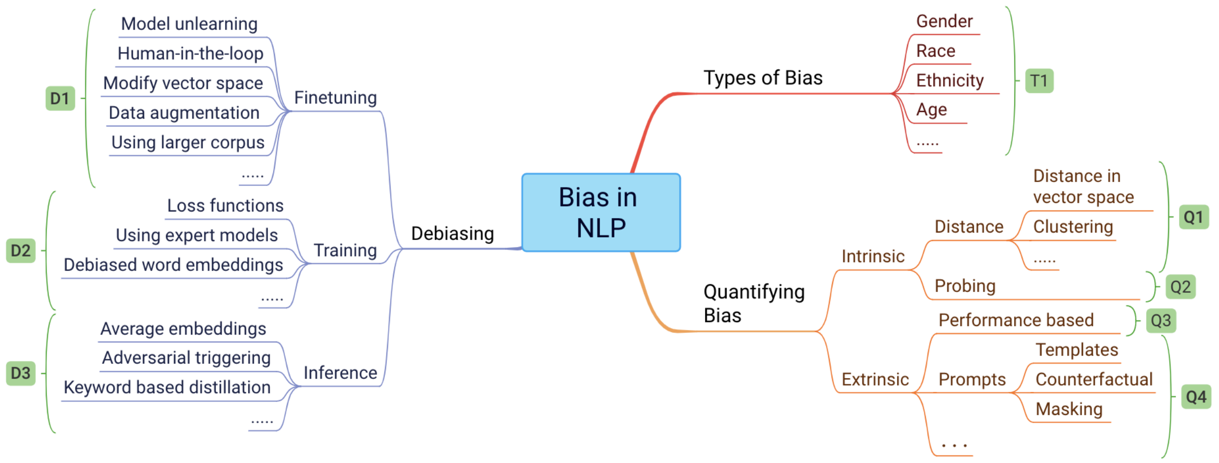

Figure 2: Three broad categories of bias research, and the upper hierarchy of each category (T, Q, D).

pers released before January 1, 2024, and included them only if they addressed language modeling, thus omitting papers on speech, where different issues arise. These criteria narrowed down an initial large set of 308 papers to 273.

We categorized the literature into three key areas. Fig. 2 illustrates our taxonomy. We came up with this organization while iteratively reviewing all papers, and we believe it effectively captures the main trends in the field. Due to the rapidly evolving field of LMs, some upcoming studies may not fit neatly into these categories. To address this, we plan to release our literature repository publicly and update it regularly with the latest research. Our work summarizes all 273 surveyed papers to provide a comprehensive understanding. Due to space constraints, we couldn't cite all 273 works in the main body. To maximize coverage within the page limits, we selected at least two papers from each line of research depicted in Figure 2 to be part of the main paper. In some cases, we wanted to cite more works but had to remove them due to space limitations. We apologize for any relevant works missed in the main body and have included a comprehensive list of all 273 papers in the Appendix.

3.1 Types of Bias - *T1*

In the realm of NLP, sociodemographic bias is particularly concerning as it can lead to differential model performance across various social groups (Deas et al., 2023; Smith et al., 2022). Sociodemographic bias includes gender bias, when models are biased against a particular gender (Hada et al., 2023; De-Arteaga et al., 2019; Park et al., 2018; Du et al., 2021; Bartl et al., 2020; Webster et al., 2021); racial bias, when models are biased against certain races (Nadeem et al., 2021; Garimella et al., 2021; Nangia et al., 2020; Tan and Celis, 2019); ethnic bias, when models are partial towards certain ethnicity (Ahn and Oh, 2021; Garg et al., 2018; Li et al., 2020; Abid et al., 2021; Manzini et al., 2019; Narayanan Venkit et al., 2023); age bias (Nangia et al., 2020; Diaz et al., 2018), sexual-orientation bias (Nangia et al., 2020; Cao and Daumé III, 2020) and many others as shown in Table 1.

Sociodemographic bias can emerge from language patterns that imply assumptions about demographic differences (Lauscher et al., 2020). These biases are often ingrained in the cultural or societal nuances of training data. For example, LMs can perpetuate biases by associating certain lexical items more strongly with particular social groups. Beyond the influence of training data, Zhou et al. (2023b) found that the size of the model, its training objectives, and tokenization strategies are important factors that affect the social bias in LMs.

Our review indicates a disproportionate focus on gender bias: it is the subject of nearly half of the surveyed papers, as Table 1 illustrates. Additionally, we observed that bias evaluation and mitigation efforts are often specific to certain biases and may not generalize well. Furthermore, over 90% of the papers we reviewed focus on English, with lim-

Types of Bias	No. of papers	Percentage
Gender	114	48%
Race	36	15%
Ethnicity	24	10%
Nationality	18	7%
Sexual Orientation	12	5%
Ableism	11	5%
Age	9	4%
Political	6	2%
Physical Appearance	5	2%
Socioeconomic status	4	2%

Table 1: Distribution of papers on bias shows a predominant focus on gender bias.

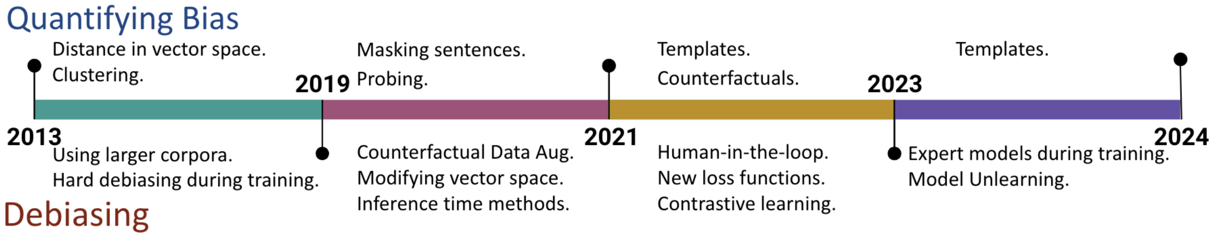

Figure 3: Evolution of changes in methods to quantify LM bias and debiasing LMs over the past decade.

ited coverage of other languages such as German, Spanish, Korean, Turkish, Chinese, and Hindi.

3.2 Quantifying Bias

Measurement of bias is challenging because it is often hidden within complex LMs. However, quantifying bias is a precondition to addressing or mitigating bias that might be harmful. Here we review different methods of measuring bias in LMs and how they differ from each other. We present an overview of evaluation datasets in the appendix.

Methods in *Q1* and *Q2* are often known as intrinsic methods as they focus on a model's internal representations to quantify bias.

3.2.1 Distance-based metrics - *Q1*

Distance in vector space. Early efforts to quantify bias in NLP (from 2013-2019, as seen in Fig. 3) primarily utilized distance metrics within embedding spaces. These approaches define certain words as 'target words' (like professions 'engineer' and 'nurse'), along with certain words as 'attributes' (often related to social categories like 'male' and 'female') (Bolukbasi et al., 2016; Brunet et al., 2019a; Dev et al., 2021). The aim was to measure the conceptual distance between these targets and attributes. The pioneering work is the Word Embedding Association Test (WEAT) score (Caliskan et al., 2017). They calculate bias as the differential association of target words with attribute sets based on cosine similarity. Subsequent to WEAT, Dev and Phillips (2019) proposed ECT score, which simplifies an attribute category, like 'female', into a single vector by averaging the embeddings of related attribute words such as 'she', 'women', and 'girl'. Ethayarajh et al. (2019) introduced RIPA, for which they used the inner product instead of cosine similarity to account for vector magnitude and directionality in measuring bias.

Some works expanded WEAT to contextual embeddings (Guo and Caliskan, 2021; Tan and Celis, 2019) and sentence-level embeddings (May et al., 2019). Other metrics used the clustering of word embeddings (Chaloner and Maldonado, 2019). Some work quantified bias based on co-occurrence of words (Valentini et al., 2023; Bordia and Bowman, 2019). Bordia and Bowman (2019) hypothesized that words occurring in close proximity to a particular gender in the train data are prone to be more biased towards that gender during testing.

3.2.2 Probing metrics - *Q2*

This category evaluates bias by examining how LMs process information, often by adding a classification layer or employing probes to test the inner workings of LMs (Chen et al., 2021; White et al., 2021). Mendelson and Belinkov (2021) used a classifier trained on latent spaces to detect biases and found that debiasing against a particular bias may increase the extent to which that bias is encoded in the inner representations of models. Orgad et al. (2022) trained a classifier to predict gender from the model's representations and shows it correlates with extrinsic bias measures better than metrics in *Q1*. Immer et al. (2022) proposed a Bayesian framework for quantifying inductive bias with probes.

In recent years, there has been less use of intrinsic methods, as they require accessing a model's internal layers to quantify bias. The increasing size of modern LMs complicates identifying the right layer for bias assessment, and the limited open-source availability of LMs raises further obstacles.

Methods in the next two subsections, *Q3* and *Q4* are often known as extrinsic methods as they focus on bias that shows up in a downstream task.

3.2.3 Performance-based metrics - *Q3*

These approaches examine how models perform across different social groups. They typically divide the test dataset into different groups to assess disparities. These works aim to quantify group differences in performance - to document whether models perform the same for all groups. De-Arteaga et al. (2019) measured gender bias by comparing the true positive rates for classification involving male versus female names and pronouns. Dixon et al. (2018) and Zhao et al. (2018a) took similar approaches, using area under the curve and false positive rate (Dixon et al., 2018), and relative accuracy (Zhao et al., 2018a). Zhang et al. (2022) and Huang et al. (2020) generated augmented datasets to measure bias as the difference in accuracy between the original and augmented datasets. Stanovsky et al. (2019) proposed a metric based on differences in accuracy across genders for machine translation.

3.2.4 Prompt-based metrics - *Q4*

Here we review methods that use various prompt-generation techniques. The first two methods in this subsection are specific to autoregressive models, while the latter focuses on Masked LMs.

Template-based methods. In these approaches, models are prompted through a set of pre-defined templates, or patterns, that capture specific types of bias or stereotypes. The templates contain slots that are filled through selection from a set of pre-defined demographic target terms during evaluation. For instance, a template could be "A <PERSON> is walking" where <PERSON> is systematically substituted with names associated with different demographic groups. By analyzing the differences in the model's responses to these substitutions, the presence and degree of bias can be measured.

Prabhakaran et al. (2019) generated templates for toxicity detection, and proposed metrics based on performance differences for target groups. Smith et al. (2022) introduces a holistic dataset, measuring bias across a dozen social demographic axes. Webster et al. (2021) defined fourteen templates to determine gender identity bias. Felkner et al. (2023) created a dataset of 45,540 sentences using 11 templates for measuring anti-LGBTQ+ bias in LMs. Gupta et al. (2023) focused on creating a robust dataset and generated 224 templates from diverse domains across 3 tasks. An et al. (2023); Parrish et al. (2022a); Li et al. (2020) proposed question-answering datasets to measure demographic bias. In contrast to performance-based metrics (*Q3*), these approaches are primarily concerned with representational harms, which occur when certain groups are depicted stereotypically or inaccurately.

Counterfactual-based methods. Several works aim to make template-based approaches more rigorous by examining how changing irrelevant attributes, known as protected attributes, affects model predictions. Specifically, "a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belongs to a different social group."

Counterfactual methods alter these protected attributes in test examples to identify attributes that significantly affect model decisions (Garg et al., 2019; Kusner et al., 2017). Huang et al. (2020) created counterfactuals for a test dataset and found that generative LLMs like GPT-2 (Radford et al., 2019) tend to generate continuations with more positive sentiment for "baker", and more negative sentiment for "accountant" as the occupation. Gardner et al. (2020) created contrast sets by generating counterfactuals for ten NLP datasets and showed that model performance drops significantly on counterfactuals. Liang et al. (2022) substituted terms linked to specific demographic groups in the test set, examining the impact on model accuracy.

Masking Sentences. Another approach to bias measurement is to mask certain words in sentences, and then analyze the model's predictions for these blanks to assess bias. Kurita et al. (2019) used this technique with occupation-related sentences, like "[MASK] is a programmer," comparing the probabilities given to male and female pronouns to identify gender biases in job associations. Similarly, Ahn and Oh (2021) quantified bias as the variance of normalized probabilities across various demographic groups. Other works using this approach include (Ousidhoum et al., 2021; Bartl et al., 2020).

Extrinsic approaches, particularly template-based ones, have gained traction in recent years (Nagireddy et al., 2023; Touileb et al., 2023; Akyürek et al., 2022), as seen in Fig. 3. The advantage of *Q4* metrics is their ability to reflect potential real-world impacts of bias by focusing on model outputs rather than solely analyzing internal parameters as in *Q1*. Extrinsic methods apply broadly to open-source or proprietary models of any size.

3.3 Debiasing

Debiasing methods aim to make models more fair and accurate in their predictions and recommendations (Subramanian et al., 2021). Turning to Daniel Kahneman again, he argues that reducing social stereotyping and bias has costs, but that the costs are worthwhile to achieve a better society (Kahneman, 2011). Extending the same principle to NLP, the effort and cost required for reducing biases are essential for creating fair NLP systems.

3.3.1 Debiasing during Finetuning - *D1*

These debiasing methods are applied during the finetuning phase of pre-trained LMs.

Data augmentation. Zmigrod et al. (2019) and Lu et al. (2020) introduced Counterfactual Data Augmentation (CDA), to reduce gender bias by generating counterfactual instances to balance gender representation. This involves substituting gender-specific words, such as *he* and *she* to construct novel sentences. Maudslay et al. (2019) enhanced this approach with Counterfactual Data Substitution (CDS), which assigns probabilities to these changes, aiming for more realistic modifications. Building upon these insights, various swapping mechanisms have been proposed to re-balance data distributions (Zhou et al., 2023a; Panda et al., 2022; Liang et al., 2020; Lauscher et al., 2021; Wen et al., 2022). Some of these augmentation approaches are also being adapted for use during model training.

Modifying vector space. Limisiewicz and Mareček (2022); Dev et al. (2020, 2021) proposed a subspace correction method for modifying embedding space. They aimed to disentangle associations between concepts that are bias-prone. Yifei et al. (2023); Manzini et al. (2019) used principal component analysis to identify and address the bias in embedding spaces. Gaci et al. (2022) redistributed attention scores to assign an equal weight for words related to bias. Ravfogel et al. (2020) learned a linear projection over representations after training, to remove the bias components in embeddings.

Fine-tuning with large corpora. Park et al. (2018) demonstrated that debiasing models benefit from fine-tuning with extensive datasets, avoiding the pitfalls of small, biased datasets. Ahn and Oh (2021) proposed that training BERT (Devlin et al., 2019) on multiple languages helps to reduce ethnic biases in each language.

Human-in-the-loop. These methods involve humans to detect and mitigate biases. Yao et al. (2021) used human-provided explanations to identify and reduce bias. Felkner et al. (2023) showed bias against marginalized communities can be mitigated using data written by that community. Chopra et al. (2020) used humans to find words linking a social group to a positive or negative trait.

Model Unlearning Recently, there has been more focus on model unlearning methods (Fig. 3). Here, the main idea is to identify and alter specific model weights that are responsible for bias. Meissner et al. (2022) identified a subset of model weights responsible for bias and masked them during testing. The advantage of their approach is it does not require finetuning. Lauscher et al. (2021); Kumar et al. (2023) captured bias mitigation functionalities using "adapters" attached to transformer blocks. Adapters offer a unique advantage in that they can be added to the model for bias correction in a plug-and-play fashion. Agarwal et al. (2023) improved on adapters by adjusting weights with data augmentation, then finetuning for specific tasks with fixed weights to prevent relearning.

Works in *D1* offer greater ease of implementation, with customizable solutions for each model. However, as the prevalence of large language models grows, they are being trained on enormous amounts of data. In such cases, bias becomes more difficult to mitigate after models have been trained.

3.3.2 Debiasing during Training - *D2*

These works apply debiasing at the pre-training time or to word embeddings used at initialization.

Debiased word embeddings Bolukbasi et al. (2016) proposed a hard debiasing technique aimed at reducing gender bias in embeddings by adjusting the vector deviations between gendered and gender-neutral terms, offering these adjusted embeddings as an alternative to standard Word2Vec embeddings. Park et al. (2018); Zhao et al. (2018b) further illustrate the effectiveness of debiased embeddings in reducing gender bias in LMs.

Loss function Several methods employ specialized loss functions to minimize bias during model pre-training. Garimella et al. (2021) used declustering loss to reduce bias. Bordia and Bowman (2019) proposed a loss regularization method. Huang et al. (2020) proposed a three-step curriculum training using the distance between the embeddings as a fairness loss to reduce sentiment bias. Liu et al. (2021) and He et al. (2022a) used adversarial training and contrastive loss respectively to reduce bias in LMs. Li et al. (2023) shows that using contrastive learning during training helps in debiasing.

Expert Models for Bias Reduction Recently methods using an auxiliary model, or so-called expert model, to reduce bias have gained prominence (cf. Fig. 3). Orgad and Belinkov (2023) predicted biased samples using an auxiliary model and performed sample reweighting to downweight these samples during pre-training. Jeon et al. (2023) used binary classifiers, referred to as bias experts, to identify biased examples within a specific class. Zhang et al. (2023) used gradient-based explanations to focus on sensitive attributes and downstream tasks, adjusting the training process to balance fairness and performance effectively.

3.3.3 Debiasing at Inference Time- *D3*

These methods apply debiasing methods at test time. In general, these methods are quite diverse. Venkit et al. (2023b) and Abid et al. (2021) applied adversarial machine learning to trigger positive associations in text generative models to reduce anti-Muslim bias and nationality bias, respectively, through prompt modifications. Majumder et al. (2023) used humans to provide feedback to balance between task performance and bias mitigation. Qian et al. (2021) performed keyword-based distillation to remove bias during inference, and to block bias acquired during training. Zhao et al. (2019) addressed gender bias through averaging representations for different gender vocabulary. Schick et al. (2021a) also presents the concept of self-debiasing, in which a model can identify and eliminate biases after generating text.

Work on debiasing during inference time faces the same issues as those in *D1*. They are easy to implement but give a false impression of debiasing and do not completely remove the model bias.

4 Limitations of Current Approaches

The works surveyed here offer valuable insights towards understanding bias in LMs, and demonstrating many innovative approaches and methodologies that have advanced the field. Alongside the commendable progress, however, a thorough analysis of the body of work on bias reveals limitations.

Reliability issues with bias metrics. The robustness of existing bias metrics is questionable. Metrics introduced in works within *Q1* and *Q3* change significantly, given minor changes in datasets or evaluation settings (Antoniak and Mimno, 2021; Spliethöver et al., 2022; Du et al., 2021; Valentini et al., 2022). Similarly, template-based methods are highly sensitive to small modifications to words used in the templates (Selvam et al., 2023; Seshadri et al., 2022; Alnegheimish et al., 2022).

Use of identical templates across bias categories. Many of the works using template-based approaches (An et al., 2023; Smith et al., 2022) use the same templates to assess diverse social biases, without considering whether certain template features should be specific to distinct types of bias. This approach risks conflating bias scores across categories, suggesting a need for more tailored templates to measure specific social biases accurately. Alternatively, investigation of ways to generalize across templates to a more abstract approach, as in Omrani Sabbaghi et al. (2023), holds promise.

Limited scope of template-based bias measurement. Template-based methods often use a restricted range of templates and target words, for example, focusing on US-based names for targets. Additionally, these approaches suffer from author bias, as templates are manually designed by the authors (Seshadri et al., 2022; Pikuliak et al., 2023).

Gap in translating bias metrics to real-world effects. There is a notable disconnect between bias metrics and their implications for real-world applications. Bias metrics in *Q1* have been claimed to correlate poorly with real-world biases (Goldfarb-Tarrant et al., 2021; Cao et al., 2022). Orgad et al. (2022) argued that intrinsic and extrinsic metrics do not correlate with each other. Such observations underscore the need for improvements in metric robustness and interpretability.

Weaknesses in finetuning approaches for debiasing. The majority of recent works on LM debiasing focus on finetuning, valued mainly for its simplicity and adaptability. However, its effectiveness is often questionable (DiCiccio et al., 2023). The complexity and size of modern LLMs, which require extensive data, time, and resources to train, make it particularly challenging to eliminate bias through finetuning. Further, these methods treat symptoms rather than root causes of bias, adjusting model outputs to appear less biased without actually removing bias from models (Gonen and Goldberg, 2019; Tokpo et al., 2023). Remarkably, some debiasing techniques can potentially increase bias (Mendelson and Belinkov, 2021). The absence of reliable bias metrics complicates the evaluation of the effectiveness of debiasing methods. We recommend that future works utilize a variety of metrics to thoroughly assess debiasing results.

Overemphasis on gender bias. Table 1 shows that about half of the literature focuses solely on gender bias. Although gender bias is a significant concern, other types of sociodemographic bias also deserve attention. Expanding research to cover a wider range of bias categories could provide a more comprehensive understanding of bias.

Lack of sociotechnical understanding of bias. The NLP literature exhibits little attention to the sociotechnical impacts of bias (Venkit et al., 2023a). Similarly, there can be incomplete consideration of the complexity of sociodemographic bias (Blodgett et al., 2020). Interdisciplinary collaborations could offer more nuanced insights and improved methodologies to measure, mitigate, prevent, and assess harms from bias.

Difficulty of comparing approaches. A better understanding is needed of strengths and weaknesses across approaches, given that works often focus on different domains and tasks. Kaneko et al. (2023) compared different bias evaluation approaches without requiring the expense of human labels. We need more work in the direction of reliable and cost-effective comparison among different measurement and mitigation methods.

5 Checklist

A checklist can assist future work to avoid the pitfalls of previous work and build more effective and reliable measurement and debiasing methods across more types of sociodemographic bias. We present 13 questions divided into three categories. Questions 1-6 focus on bias measurement (**QB**); questions 7-8 focus on bias mitigation (**BM**); questions 9-13 are applicable to all works on LM bias. We hope that future work guided by these questions can help authors situate their results within the broader literature on sociodemographic bias.

[Q1:QB] *Robustness:* Is your bias measurement stable against small modifications to templates/descriptors?

[Q2:QB] *Country-focused data:* Does your method rely on country-specific data, such as the U.S.? If so, how can it be adapted to others?

[Q3:QB] *Real-World Relevance:* How do your bias measurements reflect real-world biases, and affect end-users?

[Q4:QB] *Future Usability:* Have you taken measures to make sure your approach is easily extendable to ensure that it is useable after 5 years?

[Q5:QB] *Data Diversity:* Have you used diverse data sources to diminish biases present in the data?

[Q6:QB] *Verification of Bias Type*: What measures have you taken to ensure your bias measurement on a given type of bias doesn't overlap or confuse with other biases?

[Q7: BM] *Scalability and Efficiency:* Can your debiasing method efficiently scale to large models and datasets while maintaining effectiveness?

[Q8: BM] *Monitoring and Evaluation:* Is there a way for you to continuously assess and adjust the effectiveness of your approach?

[Q9: GQ] *Extensibility to other Social Groups:* Can your method be extended to additional sociodemographic groups?

[Q10: GQ] *Risk of Misinterpretation:* Can there be a situation when your approach falsely indicates reduced bias in models?

[Q11: GQ] *Cultural Sensitivity:* Does your approach take into account the contextual and cultural variations in language use?

[Q12: GQ] *Interdisciplinary Insights:* Does your method integrate knowledge from multiple disciplines to understand bias?

[Q13: GQ] *Transparency and Reproducibility:* Can others reproduce your method and results?

6 Future Directions

Looking ahead, we anticipate greater emphasis on bias mitigation at training time. Post-training bias mitigation adds to the costliness of very large LMs, and serves as a filter rather than a corrective. Subsequent to the first drafts of this survey, we have already seen progress in this direction (Jeon et al., 2023). The recent progress in the usage of contrastive learning during training (Li et al., 2023) and using expert models during training (Orgad and Belinkov, 2023), has shown to generate less biased models and we expect more research in these directions.

Despite their growing popularity, template-based methods for measuring bias face challenges (Selvam et al., 2023; Seshadri et al., 2022). We believe that these challenges can be tackled with careful consideration of the limitations, such as lack of robustness, leading to more effective and reliable bias measurement. We anticipate that prompt-based methods will gain prominence. Additionally, integrating interdisciplinary insights with algorithmic analysis will likely gain traction for quantifying and mitigating sociodemographic bias.

Finally, as robust methodologies emerge, we anticipate increased hope for and emphasis on intersectional bias, the overlap of multiple types of bias.

7 Conclusion

We have presented a comprehensive literature survey based on the iterative consideration of 273 works on sociodemographic bias in LMs. Our proposed typology provides an overview of the current research landscape. We identified promising directions for future research and introduced a 13-question checklist designed to guide future research towards more effective and reliable approaches and to avoid pitfalls of previous works. We encourage increased reliance on interdisciplinary methods to better capture and address the nuances of sociodemographic bias in LMs.

8 Limitations

In our survey, we focused on works from ACL Anthology, NeurIPS proceedings, FAccT, and AIES. We might have missed some relevant works in our survey, that appeared in other venues. While we have systematically organized the bias literature into categories as shown in Fig. 2, which came from an extensive survey of current literature, our framework might not encompass all existing or future research. We would like to emphasize that most of the works covered in this survey focus on the English language and some approaches discussed might not transfer to other languages. Additionally, our emphasis on sociodemographic bias means that valuable insights from works addressing other forms of bias in language models were not covered in our analysis.

9 Bias Statement

In this work, we provide a comprehensive survey of works on sociodemographic bias in language models. We defined sociodemographic bias as the difference in model performance across social groups. Such bias has the potential for harm in a real-world setting. Our definition applies to prominent demographic distinctions such as gender identity (male, female, non-binary), or income-based groupings (e.g., low, middle, and high income), or other broad-coverage distinctions that are learnable by LMs. For example, associating “Caucasian man” with “handsome”, and "African-American man" with “angry” is a clear indication of bias in models (Garimella et al., 2021). In occupation-related tasks, associating “receptionist” with “she”, and “philosopher” with “he” can have harmful effects in real-world settings (Bolukbasi et al., 2016).

10 Ethics Statement

Our work addresses the ethical impact of sociodemographic bias in NLP, offering a comprehensive survey of 273 peer-reviewed articles to highlight the presence and implications of bias within language models. By systematically organizing research findings and tracking bias approaches over the past decade, our work promotes transparency, awareness, and accountability within and beyond the NLP community. The survey provides a meticulously designed checklist, based on the weaknesses and limitations of the field, to guide future research toward more effective solutions for mitigating bias.

We also emphasize the social and ethical implications of bias underscoring the significance of addressing these issues to prevent potential negative consequences. We hope that our analysis aids in shaping more inclusive and equitable NLP technologies by fostering dialogue, awareness, and proactive measures to address sociodemographic bias, incorporating ideas beyond the field of NLP.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Oshin Agarwal, Funda Durupınar, Norman I. Badler, and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 205–211, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumit Agarwal, Aditya Veerubhotla, and Srijan Bansal. 2023. PEFTDebias : Capturing debiasing information using PEFTs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1992–2000, Singapore. Association for Computational Linguistics.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning. In *Findings of the Associ-*

ation for Computational Linguistics: NAACL 2022, pages 551–564, Seattle, United States. Association for Computational Linguistics.

Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using natural sentence prompts for understanding biases in language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2830, Seattle, United States. Association for Computational Linguistics.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1573–1596, Dubrovnik, Croatia. Association for Computational Linguistics.

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

K. Anoop, Manjary P. Gangan, P Deepak, and VL Lajish. 2022. Towards an enhanced understanding of bias in pre-trained neural language models: A survey with special emphasis on affective bias. In *Responsible Data Science: Select Proceedings of ICDSE 2021*, pages 13–45. Springer.

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. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Giuseppe Attanasio, Flor Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv. Association for Computational Linguistics.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rajas Bansal. 2022. A Survey on Bias and Fairness in Natural Language Processing. ArXiv:2204.09591 [cs].

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Lisa Bauer, Hanna Tischer, and Mohit Bansal. 2023. Social commonsense for explanation and cultural bias discovery. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3745–3760, Dubrovnik, Croatia. Association for Computational Linguistics.

Samuel James Bell and Levent Sagun. 2023. Simplicity bias leads to amplified performance disparities. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 355–369, New York, NY, USA. Association for Computing Machinery.

Emily M Bender. 2019. A typology of ethical risks in language technology with an eye towards where transparent documentation can help. the future of artificial intelligence: Language. *Ethics, Technology*.

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.

Cynthia L. Bennett and Os Keyes. 2020. What is the point of fairness? disability, AI and the complexity of justice. *SIGACCESS Access. Comput.*

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

Jayadev Bhaskaran and Isha Bhallamudi. 2019. Good secretaries, bad truck drivers? Occupational gender stereotypes in sentiment analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 62–68, Florence, Italy. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. 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.

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 *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019a. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019b. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

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

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

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.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing {bert} in hyperbolic spaces. In *International Conference on Learning Representations*.

Lingwei Cheng, Isabel O Gallegos, Derek Ouyang, Jacob Goldin, and Dan Ho. 2023. How redundant are redundant encodings? blindness in the wild and racial disparity when race is unobserved. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 667–686, New York, NY, USA. Association for Computing Machinery.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):386–393.

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89.

Kate Crawford. 2017. The trouble with bias. Keynote, *Neural Information Processing Systems (NeurIPS)*.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128. ACM.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

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, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2021. OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5034–5050. Association for Computational Linguistics.

Sunipa Dev and Jeff Phillips. 2019. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887. PMLR.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA. Association for Computing Machinery.

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.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Cyrus DiCiccio, Brian Hsu, Yinyin Yu, Preetam Nandy, and Kinjal Basu. 2023. Detection and mitigation of algorithmic bias via predictive parity. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1801–1816, New York, NY, USA. Association for Computing Machinery.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Li Du, Xiao Ding, Zhouhao Sun, Ting Liu, Bing Qin, and Jingshuo Liu. 2023. Towards stable natural language understanding via information entropy guided debiasing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2868–2882, Toronto, Canada. Association for Computational Linguistics.

Yupei Du, Qixiang Fang, and Dong Nguyen. 2021. Assessing the Reliability of Word Embedding Gender Bias Measures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10012–10034. Association for Computational Linguistics.

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding gender bias in knowledge base embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1395, Dublin, Ireland. Association for Computational Linguistics.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

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.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. 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. Association for Computational Linguistics.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Anjalie Field, Amanda Coston, Nupoor Gandhi, Alexandra Chouldechova, Emily Putnam-Hornstein, David Steier, and Yulia Tsvetkov. 2023. Examining risks of racial biases in nlp tools for child protective services. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1479–1492, New York, NY, USA. Association for Computing Machinery.

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.

Joseph Fisher, Dave Palfrey, Christos Christodoulopoulos, and Arpit Mittal. 2020. Measuring social bias in knowledge graph embeddings. In *AKBC 2020 Workshop on Bias in Automatic Knowledge Graph Construction*.

Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. 2022. An ontology for fairness metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 265–275, New York, NY, USA. Association for Computing Machinery.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 18–24, Florence, Italy. Association for Computational Linguistics.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

SongYang Gao, Shihan Dou, Qi Zhang, and Xuanjing Huang. 2022. Kernel-whitening: Overcome dataset bias with isotropic sentence embedding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4112–4122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

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.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. ACM.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, Anandhavelu N, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? On Mitigating Social Biases in Language Modelling and Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L. Alfonso Ureña-López. 2021. A Survey on Bias in Deep NLP. *Applied Sciences*, 11(7):3184.

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. Association for Computational Linguistics.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware adversarial training for name regularity bias in named entity recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

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. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Björn Ross, and Adam Lopez. 2023. Cross-lingual transfer can worsen bias in sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5691–5704, Singapore. 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.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Ben Green. 2019. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, volume 16.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, pages 41–58. Academic Press, New York.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 122–133, New York, NY, USA. Association for Computing Machinery.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022a. Mitigating gender bias in distilled language models via counterfactual role reversal. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 658–678, Dublin, Ireland. Association for Computational Linguistics.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022b. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5078–5088.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J Passonneau. 2023. Calm: A multi-task benchmark for comprehensive assessment of language model bias. *arXiv preprint arXiv:2308.12539*.

Enoch Opanin Gyamfi, Yunbo Rao, Miao Gou, and Yanhua Shao. 2020. deb2viz: Debiasing gender in word embedding data using subspace visualization. In *Eleventh International Conference on Graphics and Image Processing (ICGIP 2019)*, volume 11373, pages 671–678. SPIE.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "fifty shades of bias": Normative ratings of gender bias in GPT generated English text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, Singapore. Association for Computational Linguistics.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR.

Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm. 2021. Unpacking the interdependent systems of discrimination: Ableist bias in NLP systems through an intersectional lens. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3116–3123, Punta Cana, Dominican Republic. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022a. MABEL: Attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zexue He, Yu Wang, Julian McAuley, and Bodhisattwa Prasad Majumder. 2022b. Controlling bias exposure for fair interpretable predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5854–5866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, page 793–811, Berlin, Heidelberg. Springer-Verlag.

Samhita Honnavalli, Aesha Parekh, Lily Ou, Sophie Groenwold, Sharon Levy, Vicente Ordonez, and William Yang Wang. 2022. Towards understanding gender-seniority compound bias in natural language generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1665–1670, Marseille, France. European Language Resources Association.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual

Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83. Association for Computational Linguistics.

Christoph Hube, Maximilian Idahl, and Besnik Fetahu. 2020. Debiasing word embeddings from sentiment associations in names. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 259–267, New York, NY, USA. Association for Computing Machinery.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, New York, NY, USA. Association for Computing Machinery.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501. Association for Computational Linguistics.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1839–1851, Dublin, Ireland. Association for Computational Linguistics.

Abigail Z. Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 706, New York, NY, USA. Association for Computing Machinery.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Hailey James and David Alvarez-Melis. 2019. Probabilistic bias mitigation in word embeddings. *arXiv preprint arXiv:1910.14497*.

Eojin Jeon, Mingyu Lee, Juhyeong Park, Yeachan Kim, Wing-Lam Mok, and SangKeun Lee. 2023. Improving bias mitigation through bias experts in natural language understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11053–11066, Singapore. 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.

Przemyslaw Joniak and Akiko Aizawa. 2022. Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 67–73, Seattle, Washington. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022a. Gender bias in meta-embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3118–3133, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023. Comparing intrinsic gender bias evaluation measures without using human annotated examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2857–2863, Dubrovnik, Croatia. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022b. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716. Association for Computational Linguistics.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 45–55. ACM.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised Mary because _he_? implicit causality bias and its interaction with explicit cues in LMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871, Online. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, 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.

Satyapriya Krishna, Rahul Gupta, Apurv Verma, Jwala Dhamala, Yada Pruksachatkun, and Kai-Wei Chang. 2022. Measuring fairness of text classifiers via prediction sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5830–5842, Dublin, Ireland. Association for Computational Linguistics.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. 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.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bum Chul Kwon and Nandana Mihindukulasooriya. 2022. An empirical study on pseudo-log-likelihood bias measures for masked language models using paraphrased sentences. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 74–79, Seattle, U.S.A. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8131–8138.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Anton Ragni, Shi Wang, and Jie Fu. 2022. HERB: Measuring hierarchical regional bias in pre-trained language models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 334–346, Online only. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

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, Omar 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. *Annals of the New York Academy of Sciences*.

Tomasz Limisiewicz and David Mareček. 2021. Introducing orthogonal constraint in structural probes. 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 428–442, Online. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2022. Don't forget about pronouns: Removing gender bias in language models without losing factual gender information. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 17–29, Seattle, Washington. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020a. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The Authors Matter: Understanding and Mitigating Implicit Bias in Deep Text Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 74–85. Association for Computational Linguistics.

Tianyu Liu, Zheng Xin, Baobao Chang, and Zhifang Sui. 2020b. HypoNLI: Exploring the artificial patterns of hypothesis-only bias in natural language inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6852–6860, Marseille, France. European Language Resources Association.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202.

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.

Hongyin Luo and James Glass. 2023. Logic against bias: Textual entailment mitigates stereotypical sentence reasoning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1243–1254, Dubrovnik, Croatia. Association for Computational Linguistics.

Queenie Luo, Michael J Puett, and Michael D Smith. 2023. A perspectival mirror of the elephant: Investigating language bias on google, chatgpt, wikipedia, and youtube. *arXiv preprint arXiv:2303.16281*.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Bodhisattwa Majumder, Zexue He, and Julian McAuley. 2023. InterFair: Debiasing with natural language feedback for fair interpretable predictions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9466–9471, Singapore. Association for Computational Linguistics.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. 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.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. 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 5267–5275, Hong Kong, China. 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.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity Beyond Bias in Language Technologies for Education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460. Association for Computational Linguistics.

Katherine McCurdy and Oguz Serbetci. 2020. Grammatical gender associations outweigh topical gender bias in crosslinguistic word embeddings. *arXiv preprint arXiv:2005.08864*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1699–1710, New York, NY, USA. Association for Computing Machinery.

Johannes Mario Meissner, Saku Sugawara, and Akiko Aizawa. 2022. Debiasing masks: A new framework for shortcut mitigation in NLU. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7607–7613, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mara Mills and Meredith Whittaker. 2019. *Disability, Bias, and AI*. AI Now Institute Report. AI Now Institute Report.

Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

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.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. 2023. Socialstigmaqa: A benchmark to uncover stigma amplification in generative language models. *arXiv preprint arXiv:2312.07492*.

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. Association for Computational Linguistics.

Pranav Narayanan Venkit. 2023. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1004–1005.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Unmasking nationality bias: A study of human perception of nationalities in ai-generated articles. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 554–565, New York, NY, USA. Association for Computing Machinery.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 149–155, New York, NY, USA. Association for Computing Machinery.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 542–553, New York, NY, USA. Association for Computing Machinery.

Hadas Orgad and Yonatan Belinkov. 2023. BLIND: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821, Toronto, Canada. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Merrick Osborne, Ali Omrani, and Morteza Dehghani. 2023. The sins of the parents are to be laid upon the children: Biased humans, biased data, biased models. *Perspectives on psychological science : a journal of the Association for Psychological Science*, page 17456916231180099.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained 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 4262–4274, Online. Association for Computational Linguistics.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Swetasudha Panda, Ari Kobren, Michael Wick, and Qinlan Shen. 2022. Don't just clean it, proxy clean it: Mitigating bias by proxy in pre-trained models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5073–5085, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA. Association for Computing Machinery.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022a. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022b. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Thomas F Pettigrew and Linda R Tropp. 2006. A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, 90(5):751.

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation Sensitivity Analysis to Detect Unintended Model Biases. 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 5740–5745. Association for Computational Linguistics.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Ivan Provilkov and Andrey Malinin. 2021. Multi-sentence resampling: A simple approach to alleviate dataset length bias and beam-search degradation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8612–8621, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2020. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 450–456, New York, NY, USA. Association for Computing Machinery.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual Inference for Text Classification Debiasing. 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 5434–5445. Association for Computational Linguistics.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Nils Karl Reimer and Nikhil Kumar Sengupta. 2023. Meta-analysis of the "ironic" effects of intergroup contact. *Journal of Personality and Social Psychology*, 124(2):362.

Brianna Richardson, Prasanna Sattigeri, Dennis Wei, Karthikeyan Natesan Ramamurthy, Kush Varshney, Amit Dhurandhar, and Juan E. Gilbert. 2023. Add-remove-or-relabel: Practitioner-friendly bias mitigation via influential fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 736–752, New York, NY, USA. Association for Computing Machinery.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. 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 4187–4195, Minneapolis, Minnesota. 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.

David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PloS one*, 15(4):e0231189.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

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. Association for Computational Linguistics.

Magnus Sahlgren and Fredrik Olsson. 2019. Gender bias in pretrained Swedish embeddings. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 35–43, Turku, Finland. Linköping University Electronic Press.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021a. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021b. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Reva Schwartz, Leann Down, Adam Jonas, and Elham Tabassi. 2021. A proposal for identifying and managing bias in artificial intelligence. *Draft NIST Special Publication*, 1270.

João Sedoc and Lyle Ungar. 2019. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada. Association for Computational Linguistics.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726, Seattle, United States. Association for Computational Linguistics.

Procheta Sen and Debasis Ganguly. 2020. Towards socially responsible ai: Cognitive bias-aware multi-objective learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2685–2692.

Preethi Seshadri, Pouya Pezeshkpour, and Sameer Singh. 2022. Quantifying social biases using templates is unreliable. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

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.

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation bias in data: A survey on identification and resolution techniques. *ACM Comput. Surv.* Just Accepted.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. 2018. Darling or babygirl? investigating stylistic bias in sentiment analysis. *Proc. of FATML*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Seungjae Shin, Kyungwoo Song, JoonHo Jang, Hyemi Kim, Weonyoung Joo, and Il-Chul Moon. 2020. Neutralizing gender bias in word embeddings with latent disentanglement and counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3126–3140, Online. Association for Computational Linguistics.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211. Association for Computational Linguistics.

Maximilian Spliethöver, Maximilian Keiff, and Henning Wachsmuth. 2022. No word embedding model is perfect: Evaluating the representation accuracy for social bias in the media. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2081–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. 2022. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 39–51, Abu Dhabi, UAE. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing.

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. Association for Computational Linguistics.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640. Association for Computational Linguistics.

Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini. 2018. Biased embeddings from wild data: Measuring, understanding and removing. In *Advances in Intelligent Data Analysis XVII: 17th International Symposium, IDA 2018,'s-Hertogenbosch, The Netherlands, October 24–26, 2018, Proceedings 17*, pages 328–339. Springer.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Chris Sweeney and Maryam Najafian. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 359–368, New York, NY, USA. Association for Computing Machinery.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Jacob Thebault-Spieker, Sukrit Venkatagiri, Naomi Mine, and Kurt Luther. 2023. Diverse perspectives can mitigate political bias in crowdsourced content moderation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1280–1291, New York, NY, USA. Association for Computing Machinery.

Ewoenam Kwaku Tokpo, Pieter Delobelle, Bettina Berendt, and Toon Calders. 2023. How far can it go? on intrinsic gender bias mitigation for text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3418–3433, Dubrovnik, Croatia. Association for Computational Linguistics.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are fairy tales fair? analyzing gender bias in temporal narrative event chains of children's fairy tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2023. Measuring normative and descriptive biases in language models using census data. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2242–2248, Dubrovnik, Croatia. Association for Computational Linguistics.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Francisco Valentini, Germán Rosati, Damián Blasi, Diego Fernandez Slezak, and Edgar Altszyler. 2023. On the interpretability and significance of bias metrics in texts: a PMI-based approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 509–520, Toronto, Canada. Association for Computational Linguistics.

Francisco Valentini, Germán Rosati, Diego Fernandez Slezak, and Edgar Altszyler. 2022. The undesirable dependence on frequency of gender bias metrics based on word embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5086–5092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J. Passonneau, and Shomir Wilson. 2023a. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Shomir Wilson, et al. 2023b. Nationality bias in text generation. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2022. A study of implicit bias in pretrained language models against people with disabilities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332.

Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. 2023c. Automated ableism: An exploration of explicit disability biases in sentiment and toxicity analysis models. In *The Third Workshop on Trustworthy Natural Language Processing*, page 26.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models. *arXiv preprint arXiv:2111.13259*.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, and Yi Chang. 2021. Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3002–3012, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and Reducing Gendered Correlations in Pre-trained Models. ArXiv:2010.06032 [cs].

Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. AutoCAD: Automatically generate counterfactuals for mitigating shortcut learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. 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 454–463, Online. Association for Computational Linguistics.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-gap: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3743–3755.

Zekun Yang and Juan Feng. 2020. A causal inference method for reducing gender bias in word embedding relations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9434–9441.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining Language Models with Compositional Explanations. In *Advances in Neural Information Processing Systems*, volume 34, pages 8954–8967. Curran Associates, Inc.

Li Yifei, Lyle Ungar, and João Sedoc. 2023. Conceptor-aided debiasing of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10703–10727, Singapore. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020a. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020b. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

Jindi Zhang, Luning Wang, Dan Su, Yongxiang Huang, Caleb Chen Cao, and Lei Chen. 2023. Model debiasing via gradient-based explanation on representation. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 193–204, New York, NY, USA. Association for Computing Machinery.

Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. 2022. Interpreting the robustness of neural NLP models to textual perturbations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3993–4007, Dublin, Ireland. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized 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 629–634. Association for Computational Linguistics.

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. 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.

Alina Zhiltsova, Simon Caton, and Catherine Mulway. 2019. Mitigation of unintended biases against non-native english texts in sentiment analysis. In *AICS*, pages 317–328.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023a. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. 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 5276–5284. Association for Computational Linguistics.

Yi Zhou, Jose Camacho-Collados, and Danushka Bollegala. 2023b. A predictive factor analysis of social biases and task-performance in pretrained masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11082–11100, Singapore. Association for Computational Linguistics.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

Ran Zmigrod, Sebastian J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661. Association for Computational Linguistics.

A Appendix

A.1 Evaluation Datasets

Bias benchmark datasets provide valuable resources for NLP fairness research. These datasets commonly contain illustrative examples of biased language, often templated sentences filled with contrastive social group terms. Datasets allow standardized bias evaluation on diverse tasks using controlled examples. Many of them focus on a particular type of language context, such as coreference, sentiment, or question answering, while others probe for stereotype bias through word associations. Table present in the *Appendix* summarizes these datasets.

In the case of *coreference resolution*, Zhao et al. (2018a) proposed a method for identifying gender bias using Winograd-schema sentences for occupation terms. Webster et al. (2018) introduced GAP, a gender-balanced, labeled corpus of 8,908 ambiguous pronoun–name pairs designed to detect gender bias in coreference resolution. In the *word association* domain, Nangia et al. (2020) presented CrowS-Pairs, a sentence pair corpus that measures a model's bias by assessing if it favors sentences with stereotypes. Nadeem et al. (2021) released StereoSet, a large-scale natural dataset in English designed to measure stereotypical bias using inter- and intra-sentence association of words to stereotypical contexts. Li et al. (2020) proposed UNQOVER, a general framework for probing bias in *question answering* models using questions to probe whether a model associates a sociodemographic group to a stereotype. Smith et al. (2022) published HolisticBias, consisting of 450,000 unique sentence prompts for measuring 13 types of sociodemographic bias in generative LMs.

In the domain of *sentiment evaluation*, Kiritchenko and Mohammad (2018) released EEC, an 8,640 English sentence collection curated to test bias toward certain races and genders in sentiment analysis models. BITS (Venkit and Wilson, 2021; Venkit et al., 2023c) is a similar corpus of 1,126 sentences curated to measure disability, race, and gender bias in sentiment and toxicity analysis models.

Table 2 provides list of datasets for quantifying bias in NLP models.

A.2 List of papers surveyed

Below is the list of papers surveyed in this work, sorted based on our taxonomy.

Explicit Bias(T1) :

(Mei et al., 2023; Deas et al., 2023; Liu et al., 2021; De-Arteaga et al., 2019; Bell and Sagun, 2023; Silva et al., 2021; Park et al., 2018; Sap et al., 2020; B et al., 2021; Lauscher and Glavaš, 2019; Rozado, 2020; Rudinger et al., 2017; Shah et al., 2020; Du et al., 2022; Nozza et al., 2022; Honnavalli et al., 2022; Lucy and Bamman, 2021; Mendelson and Belinkov, 2021; Matthews et al., 2021; Cao et al., 2022; Papakyriakopoulos et al., 2020; Kementchedjhieva et al., 2021; Garrido-Muñoz et al., 2021; Strengers et al., 2020; Delobelle et al., 2022; Fisher et al., 2020; Sheng et al., 2020; Zhang et al., 2020a; Hendricks et al.,

Dataset name	Task	Bias Type	Dataset Size
WinoBias (Zhao et al., 2018a)	Coreference Resolution	Gender	1,580
WinoGender (Rudinger et al., 2018)	Coreference Resolution	Gender	720
GAP (Webster et al., 2018)	Coreference Resolution	Gender	8,908
Counter-GAP (Xie et al., 2023)	Coreference Resolution	Gender	4,008
CrowS-Pairs (Nangia et al., 2020)	Word Association	Gender, race, religion, age, sexual orientation, nationality, disability, physical appearance, and socioeco. status.	1,508
StereoSet (Nadeem et al., 2021)	Word Association	Race, gender, religion, and profession	16,995
WikiGenderBias (Gaut et al., 2020)	Word Association	Gender	45,000
UnQOVER (Li et al., 2020)	Word Association	Gender, Nationality, Ethnicity,Religion	8,908
WinoGrande (Sakaguchi et al., 2021)	Word Association	Dataset Bias	1,767
BBQ (Parrish et al., 2022b)	Word Association	9 Sociodemographic Group	58,492
EEC (Kiritchenko and Mohammad, 2018)	Sentiment Evaluation	Gender, Race	8,640
BITS (Venkit and Wilson, 2021)	Sentiment Evaluation	Gender, Race, Disability	1,126
HolisticBias (Smith et al., 2022)	Text Generation	13 Sociodemographic Group	450,000

Table 2: List of Evaluation datasets used to measure bias in NLP models

2018; Mehrabi et al., 2021; Mayfield et al., 2019; Schwartz et al., 2021; Nozza et al., 2019; Vaidya et al., 2020; He et al., 2019; Hovy and Søgaard, 2015; Wolfe and Caliskan, 2021; Sakaguchi et al., 2021; Agarwal et al., 2019; White and Cotterell, 2021; Luo and Glass, 2023)

Gender Bias : (Sharma et al., 2022; Kaneko et al., 2022a; Stahl et al., 2022; Kaneko et al., 2023; Toro Isaza et al., 2023; Hada et al., 2023; Attanasio et al., 2023; Goldfarb-Tarrant et al., 2023; Lee et al., 2023; Gaut et al., 2020; Sun et al., 2019; Hamidi et al., 2018; Zhou et al., 2019; Savoldi et al., 2021; Sahlgren and Olsson, 2019; Ahn et al., 2022; Tal et al., 2022; Kaneko et al., 2022b; Field and Tsvetkov, 2020; Garimella et al., 2019; Escudé Font and Costa-jussà, 2019; Bhaskaran and Bhallamudi, 2019; McCurdy and Serbetci, 2020; Kaneko and Bollegala, 2019; Larson, 2017; Du et al., 2021; Bartl et al., 2020; Webster et al., 2021; Tan and Celis, 2019; Bolukbasi et al., 2016; Maudslay et al., 2019; Zhao et al., 2019; Rudinger et al., 2018; Lu et al., 2020)

Racial Bias : (Goldfarb-Tarrant et al., 2023; Levy et al., 2023; Field et al., 2023; Cheng et al., 2023; Sap et al., 2019; Hanna et al., 2020; Blodgett et al., 2016; Davidson et al., 2019; Friedman et al., 2019; Shen et al., 2018; Karve et al., 2019; Nadeem et al., 2021; Garimella et al., 2021; Nangia et al., 2020; Tan and Celis, 2019; Guo and Caliskan, 2021; Brown et al., 2020)

Disability bias : (Venkit and Wilson, 2021; Venkit et al., 2022; Hutchinson et al., 2020; Bennett and Keyes, 2020; Mills and Whittaker, 2019; Hassan et al., 2021; Narayanan Venkit, 2023)

Ethnicity bias : (Bauer et al., 2023; Levy et al., 2023; Malik et al., 2022; Li et al., 2022; Ahn and Oh, 2021; Garg et al., 2018; Li et al., 2020; Abid et al., 2021; Manzini et al., 2019; Venkit et al., 2023b; Bhatt et al., 2022), Nationality bias - (Ladhak et al., 2023; Levy et al., 2023; Narayanan Venkit et al., 2023), Political bias - (Thebault-Spieker et al., 2023; Shen et al., 2018; Rozado, 2020), Age bias (Nangia et al., 2020; Diaz et al., 2018) and sexual-orientation bias (Ovalle et al., 2023; Nangia et al., 2020; Cao and Daumé III, 2020)

Distance based metrics(Q1) : (Caliskan et al., 2017; Dev and Phillips, 2019; Zhao et al., 2017; Basta et al., 2019; Shen et al., 2018; Brunet et al., 2019b; May et al., 2019; Dev et al., 2021; Zhou et al., 2019; Pujari et al., 2020; Sutton et al., 2018; Lauscher et al., 2020; Guo and Caliskan, 2021; Bolukbasi et al., 2016; Ross et al., 2021; Tan and Celis, 2019; Ethayarajh et al., 2019; Chaloner and Maldonado, 2019; Bordia and Bowman, 2019; Valentini et al., 2023)

Probing based metrics(Q2) : (Orgad et al., 2022; Immer et al., 2022; Chen et al., 2021; Limisiewicz and Mareček, 2021; Kennedy et al., 2020; Sweeney and Najafian, 2019; Tan et al., 2020; Mendelson and Belinkov, 2021; White et al., 2021)

Performance metrics(Q3) : (De-Arteaga et al., 2019; Kwon and Mihindukulasooriya, 2022; Zhang et al., 2022; Huang et al., 2020; Dixon et al., 2018; Zhao et al., 2018a; Cho et al., 2019; Stanovsky et al., 2019; Gonen and Webster, 2020; Borkan et al., 2019; Dev et al., 2020)

Prompt based metrics(Q4) : (Nagireddy et al., 2023; Webster et al., 2021; Smith et al., 2022; Kurita et al., 2019; Krishna et al., 2022; Bhaskaran and Bhallamudi, 2019; Gupta et al., 2022b; Prabhakaran et al., 2019; Ahn and Oh, 2021; Bartl et al., 2020; Li et al., 2020; Venkit and Wilson, 2021; Salazar et al., 2020; Dev et al., 2020; Diaz et al., 2018; Zhang et al., 2020b; Garg et al., 2019; Liang et al., 2022; Kusner et al., 2017; Huang et al., 2020; Akyürek et al., 2022; Gardner et al., 2020; Ousidhoum et al., 2021; Parrish et al., 2022a; Kiritchenko and Mohammad, 2018; Touileb et al., 2023; Gupta et al., 2023; Pikuliak et al., 2023; Touileb et al., 2023; An et al., 2023; Felkner et al., 2023; Ousidhoum et al., 2021)

Debiasing during Finetuning(D1) : (Ungless et al., 2022; Du et al., 2023; Omrani et al., 2023; Zhou et al., 2023a; Thakur et al., 2023; Jin et al., 2021; He et al., 2022b; Zmigrod et al., 2019; Jin et al., 2021; Gaci et al., 2022; Gupta et al., 2022a; Ghaddar et al., 2021; Kumar et al., 2020; Han et al., 2021; Attanasio et al., 2022; Joniak and Aizawa, 2022; Chopra et al., 2020; Maudslay et al., 2019; Park et al., 2018; Yao et al., 2021; Liang et al., 2020; Sen et al., 2022; Ma et al., 2020; Limisiewicz and Mareček, 2022; Yang et al., 2021; Wang et al., 2021; Pujari et al., 2020; Sedoc and Ungar, 2019; Tan et al., 2020; Sutton et al., 2018; Ravfogel et al., 2020; Kaneko and Bollegala, 2019; Karve et al., 2019; Gyamfi et al., 2020; Shin et al., 2020; Zhang et al., 2020a; Wen et al., 2022; Chopra et al., 2020; Yang and Feng, 2020; Lu et al., 2020; Lauscher et al., 2021; Garg et al., 2019; Dev et al., 2020, 2021; Manzini et al., 2019; Bolukbasi et al., 2016; Ahn and Oh, 2021; Orgad et al., 2022; Felkner et al., 2023; de Vassimon Manela et al., 2021)

Debiasing during Training (D2) : (An et al., 2022; Bolukbasi et al., 2016; He et al., 2019; Han et al., 2022; Liu et al., 2020b; Escudé Font and Costa-jussà, 2019; Prost et al., 2019; James and Alvarez-Melis, 2019; Park et al., 2018; Zhao et al., 2018b; Gao et al., 2022; Sweeney and Najafian, 2020; Hube et al., 2020; Sen and Ganguly, 2020; Saunders and Byrne, 2020; Dixon et al., 2018; Karimi Mahabadi et al., 2020; He et al., 2022a; Richardson et al., 2023) Loss functions for bias mitigation : (Hashimoto et al., 2018; Qian et al., 2019; Berg et al., 2022; Romanov et al., 2019; Garimella et al., 2021; Bordia and Bowman, 2019; Huang et al., 2020; Provilkov and Malinin, 2021; Liu et al., 2021; Orgad and Belinkov, 2023; Li et al., 2023)

Debiasing during Inference (D3) : (Majumder et al., 2023; Qian et al., 2021; Zhao et al., 2019; Abid et al., 2021; Guo et al., 2022; Schick et al., 2021b; Venkit et al., 2023b)

Works on Bias : These are works that are difficult to categorize in one of the above categories. (Chouldechova and Roth, 2020; Green, 2019; Zhang and Bareinboim, 2018; Mayfield et al., 2019; Katell et al., 2020; Dwork et al., 2012; Jacobs et al., 2020; Anoop et al., 2022; Czarnowska et al., 2021; Blodgett et al., 2021; Zhuo et al., 2023; Mulligan et al., 2019; Jacobs and Wallach, 2021; Schoch et al., 2020; Franklin et al., 2022; Bender, 2019; España-Bonet and Barrón-Cedeño, 2022; Hutchinson and Mitchell, 2019; Bender et al., 2021; Goldfarb-Tarrant et al., 2021; Brown et al., 2020; Li et al., 2020; Bagdasaryan et al., 2019; Liu et al., 2020a; Zhiltsova et al., 2019; Chopra et al., 2020; Luo et al., 2023; Shah et al., 2020; Garrido-Muñoz et al., 2021; Delobelle et al., 2022; Czarnowska et al., 2021)