Social-Group-Agnostic Bias Mitigation via the Stereotype Content Model

Ali Omrani[†] Alireza S. Ziabari[†] Charles Yu[§] Preni Golazizian[†]

Brendan Kennedy[†] Mohammad Atari[†] Heng Ji[§] Morteza Dehghani[†]

[†] University of Southern California
[§] University of Illinois at Urbana-Champaign
{aomrani, salkhord, golazizi, btkenned, atari, mdehghan}@usc.edu
{ctyu2, hengji}@illinois.edu

Abstract

Existing bias mitigation methods require social-group-specific word pairs (e.g., "man" – "woman") for each social attribute (e.g., gender), restricting the bias mitigation to only one specified social attribute. Further, this constraint renders such methods impractical and costly for mitigating bias in understudied and/or unmarked social groups. We propose that the Stereotype Content Model (SCM) — a theoretical framework developed in social psychology for understanding the content of stereotyping — can help debiasing efforts to become social-group-agnostic by capturing the underlying connection between bias and stereotypes. SCM proposes that the content of stereotypes map to two psychological dimensions of *warmth* and *competence*. Using only pairs of terms for these two dimensions (e.g., warmth: "genuine" – "fake"; competence: "smart" – "stupid"), we perform debiasing with established methods on both pre-trained word embeddings and large language models. We demonstrate that our social-group-agnostic, SCM-based debiasing technique performs comparably to group-specific debiasing on multiple bias benchmarks, but has theoretical and practical advantages over existing approaches.

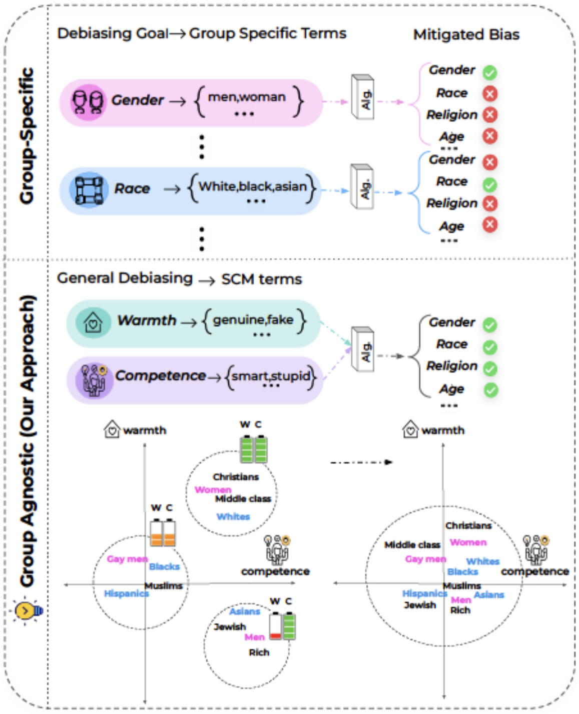

Figure 1: Group-specific debiasing vs. our proposed group-agnostic approach. Rather than iteratively debiasing with respect to each social attribute (e.g., gender or race), embeddings or language models are debiased with respect to *warmth* and *competence*, the two dimensions of the Stereotype Content Model (SCM).

1 Introduction

The societal impacts of Natural Language Processing (NLP) have stimulated research on measuring and mitigating the unintended social-group biases encoded in language models (Hovy and Spruit, 2016). However, the majority of this important line of work is atheoretical in nature and "fails to engage critically with what constitutes 'bias' in the first place" (Blodgett et al., 2020). The bias found in language models is rooted in human biases (Caliskan and Lewis, 2022); thus, to alleviate such biases, we should ground our debiasing approaches in social psychological theories of stereotyping. These theories can help us shed light on the underlying structure of language-embedded biases rather than attending to ad hoc superficial patterns (Osborne et al., 2022).

Although there is a multitude of approaches to bias mitigation (Bolukbasi et al., 2016; Zhao et al., 2018; Dev and Phillips, 2019; Kaneko and Bollegala, 2021; Solaiman and Dennison, 2021), most of these approaches are *group-specific*. Such methods, which debias along subspaces defined by social groups or attributes (e.g., gender or race), are not only atheoretical but also unscalable. Resources developed for bias mitigation on one social group or attribute (e.g., gender) do not axiomatically translate easily into other groups or attributes (e.g., age).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4123–4139
July 9-14, 2023 ©2023 Association for Computational Linguistics

For example, previous works' focus on gender bias has driven the development of resources that are only applicable to gender debiasing (e.g., equality word sets for gender), but biases associated with other social groups and/or attributes remain understudied. Beyond the challenge of creating such resources for a given attribute, to achieve an "unbiased" model with group-specific debiasing, one would have to iterate over all social groups. This approach is practically impossible and arguably would result in significant degrading in the expressiveness of the model. Furthermore, group-specific debiasing is limited in terms of effectiveness: stereotypic relations in distributed representations are deep-rooted, and thus may not be easily removed using explicit sets of group-specific words (Agarwal et al., 2019; Gonen and Goldberg, 2019). In contrast, a social-group-agnostic approach would not have such restrictions.

The Stereotype Content Model (SCM; Fiske et al., 2002) is a theoretical framework developed in social psychology to understand the content and function of stereotypes in interpersonal and intergroup interactions. The SCM proposes that human stereotypes are captured by two primary dimensions of social perception: *warmth* (e.g., trustworthiness, friendliness) and *competence* (e.g., capability, assertiveness). From a socio-functional, pragmatic perspective, people's perception of others' intent (i.e., warmth) and capability to act upon their intentions (i.e., competence) affect their subsequent emotion and behavior (Cuddy et al., 2009). Warmth divides people (or groups of people under a social identity) into "friends" or "foes," while competence contains perceptions of social groups' status. Depending on historical processes, various social groups may be located in different stereotypic quadrants (high vs. low on warmth and competence) based on this two-dimensional model (Charlesworth et al., 2022). While there are alternative theories related to social evaluation and stereotyping (e.g., Abele and Wojciszke, 2007; Ellemers et al., 2013; Koch et al., 2016; Yzerbyt, 2018) Ellemers et al. (2020) demonstrate the possibility of establishing a theoretical alignment between them.

In this work, we propose SCM-based bias mitigation and demonstrate that by relying on a theoretical understanding of social stereotypes to define the bias subspace (rather than group-specific subspaces), bias in pre-trained word embeddings and large language models can be adequately mitigated across multiple social attributes (Figure 1). We chose to focus on SCM not only because of the availability of validated linguistic resources (Nicolas et al., 2021), but also due to SCM's prominence, parsimony, and robustness across languages and cultures (e.g., Bye and Herrebrøden, 2018; Grigoryev et al., 2019; Sharifian et al., 2022; Liang et al., 2022). Specifically, we confirm that by debiasing with respect to the subspace defined by warmth and competence, our SCM-based approach performs comparably with group-specific debiasing for a given group (e.g., SCM-based debiasing is comparable to race-debiasing on race, see §3.5). We also show that SCM-based debiasing simultaneously reduces bias for understudied attributes such as religion and age (§3.5 and §4.4). Finally, we show that for both word embeddings and large language models SCM-based debiasing retains original model's expressiveness (§3.6 and §4.4). Overall, our results confirm the viability of a theory-based, social-group-agnostic approach to bias mitigation[1].

2 Background

2.1 Bias Mitigation in Word Embedding

Part of our work builds on post hoc bias mitigation which aims to remove biases by modifying pre-trained word embeddings and language models. Most efforts we review are restricted to gender-related debiasing (e.g., Bolukbasi et al., 2016; Zhao et al., 2018; Dev and Phillips, 2019; Ravfogel et al., 2020); importantly, we focus our work on other social categories as well, bringing attention to these understudied groups and attributes. Originally, Bolukbasi et al. (2016) proposed Hard Debiasing (HD) for gender bias. HD removes the gender component from inherently non-gendered words and enforces an equidistance property for inherently gendered word pairs (equality sets). Two follow-ups to this work include: Manzini et al. (2019), which formulated a multiclass version of HD for attributes such as race; and Dev and Phillips (2019), which introduced Partial Projection (PP), a method that does not require equality sets and is more effective than HD in reducing bias. Extending these approaches to other social attributes is not trivial because a set of definitional word pairs has to be curated for each social group/attribute; this curation is a non-trivial task as the list of words required is

[1] https://github.com/Ali-Omrani/Social-Group-Agnostic-Bias-Mitigation

dynamic and context-dependent.

Gonen and Goldberg (2019) demonstrated that gender bias in word embeddings is deeper than previously thought, and methods based on projecting words onto a "gender dimension" only hide bias superficially. They showed that after debiasing, most words maintain their relative position in the debiased subspace. Our work is motivated by this important criticism. Specifically, we argue that our theory-driven approach removes the actual psychological 'bias' subspace, rather than subspaces, often superficially, containing bias for specific groups.

2.2 Bias Mitigation in Language Models

Fine-tuning and prompt-tuning are two major paradigms for using pre-trained language models in NLP, and bias mitigation approaches have been proposed based on both. Kaneko and Bollegala (2021) introduced Debiasing Pre-trained Contextualized Embeddings (DPCE), a finetuning method for bias mitigation in language models with a loss term that aims to find a balance between bias mitigation and language modeling ability. Other fine-tuning methods include AutoDebias (Guo et al., 2022), which samples the language model to find examples for finetuning-based debiasing. Promp-tuning (Li and Liang, 2021; Liu et al., 2021; Lester et al., 2021) is mainly done through either discrete prompts, which consist of text (i.e., tokens), or continuous prompts, which consist of a continuous array of numbers prefixed to a language model and trained. It has been shown that bias in language models can be mitigated by providing descriptions of desired and undesired behavior as discrete prompts (Schütz et al., 2021; Askell et al., 2021; Solaiman and Denison, 2021). More recently Yang et al. (2023) proposed "A DEbiasing PrompT" (ADEPT) that outperforms DPCE by using continuous prompt tuning to mitigate biases in a language model. Yu et al. (2023) proposed partitioned contrastive gradient unlearning (PCGU), a gray-box method for debiasing pretrained masked language models by optimizing only the weights that contribute most to a specific domain of bias.

2.3 The SCM and Language

SCM is a well-established theoretical framework of stereotyping, and has begun to be applied in NLP. Recently, Nicolas et al. (2021) developed dictionaries to measure warmth and competence in textual data. Each dictionary was initialized with a set of seed words from the literature which was further expanded using WordNet (Miller, 1995) to increase the coverage of stereotypes collected from a sample of Americans. Cao et al. (2022) employ Agency-Belief-Communion (ABC; Koch et al., 2016), an alternative theory from social psychology for stereotype content, to discover stereotyping in language models. Fraser et al. (2021) demonstrated that, in word embeddings, SCM dictionaries capture the group stereotypes documented in social psychological research. Recently, Davani et al. (2023) applied SCM dictionaries to quantify social group stereotypes embedded in language, demonstrating that patterns of prediction biases can be explained using social groups' warmth and competence embedded in language. Contemporary to our work, Ungless et al. (2022) explore the idea of using SCM for bias mitigation but fall short of evaluating their models on the necessary benchmarks.

3 SCM-Based Bias Mitigation for Static Word Embeddings

Before we discuss our proposed method, we briefly review the algorithms and benchmarks on bias mitigation in word embeddings. There are two components to each post hoc bias mitigation approach for static word embeddings: the **Bias Subspace**, which determines the subspace over which the algorithms operate, and the **Algorithm**, which is how the word embeddings are modified with respect to the bias subspace. In this section, we review the concept of bias subspaces, established algorithms for debiasing, and how bias is quantified in word embeddings. Finally, we introduce our social-group-agnostic framework; SCM-based debiasing.

3.1 Identifying a Bias Subspace

Post hoc word embedding debiasing algorithms operate over a subspace of bias in the embedding space. Given a set $D = \{(d_1^+, d_1^-), ..., (d_n^+, d_n^-)\}$ of word pairs that define the bias concept (e.g. "father"–"mother" for binary gender) the bias subspace v_B is the first k principal components of matrix C, constructed from stacking the difference in embeddings of d_i^+ and d_i^-.

3.2 Debiasing Algorithms

Method definitions below use the following notation: W denotes vocabulary, w and w' denote the embedding of word w before and after debiasing.

Hard Debiasing (HD) An established approach for mitigating bias in word embeddings is Hard

Debiasing (HD; Bolukbasi et al., 2016). For gender, HD removes the gender subspace from words that are not inherently gendered by projecting them orthogonal to gender subspace. For word pairs that are inherently gendered, HD equalizes them, modifying the embeddings such that they are equidistant from the inherently non-gendered words.

Subtraction (Sub) Sub was introduced as a baseline by Dev and Phillips (2019) wherein the bias subpspace v_B is subtracted from all word vectors. Formally, for all $w \in W$,$w' := w - v_B$.

Linear Projection (LP) To mitigate the bias with respect to bias dimension v_B, Linear Projection (LP) projects every word $w \in W$ to be orthogonal to v_B. Formally, $w' := w - \pi_B(w)$ where $\pi_B(w) = \langle w, v_B \rangle v_B$ is the projection of w onto v_B.

Partial Projection (PP) To improve on LP, Partial Projection (PP) was developed to allow the extent of projection to vary based on the component of the given word vector which is orthogonal to the bias subspace. Intuitively, only words with unintended bias (e.g., "nurse" or "doctor"), and not words which are definitional to the bias concept (e.g., "man" or "woman") will have a large orthogonal component to the bias subspace v_B. For all words $w \in W$,

$$w' = \mu + r(w) + \beta \cdot f(\|r(w)\|) \cdot v_B$$
$$\beta = \langle w, v_B \rangle - \langle \mu, v_B \rangle$$

where μ is the mean embedding of words used to define v_B, $r(w) = w - \langle w, v_B \rangle v_B$ is the bias-orthogonal component, and $f(.)$ is a smoothing function which helps to remove unintended bias and keep definitional bias. We use $f(\eta) = \frac{1}{(\eta+1)^2}$ (see Dev and Phillips, 2019).

3.3 Static Word Embedding Benchmarks

Embedding Coherence Test Given a set of tuples $A = \{(a_1^1, a_1^2, ...), (a_2^1, a_1^2, ...)\}$ where a_j^i denotes the j^{th} word for i^{th} subgroup of an attribute (e.g., {("father", "mother"), ...} for binary gender), and a set of professions $P = \{p_1, ..., p_m\}$, the Embedding Coherence Test (ECT; Dev and Phillips, 2019) is the Spearman rank correlation between the rank order of cosine similarities of professions with each subgroup's average embedding. Bias is completely removed when subgroups have identical ordering of associations with professions (ECT $= 1$).

Embedding Quality Test Word analogies are one of the main methods for evaluating word embeddings (Mikolov et al., 2013). The EQT (Dev and Phillips, 2019) quantifies the improvement in unbiased analogy generation after debiasing. Similar to ECT, EQT requires a set of word pairs A and a set of professions P. For each word pair (a_i^+, a_i^-) the analogy $a_i^+ : a_i^- :: p_j$ is completed, if the answer is p_j or plurals or synonyms of p_j (via NLTK; Bird et al. 2009), it is counted as unbiased. EQT is the ratio of unbiased analogies to all analogies. An ideal unbiased model would achieve EQT$= 1$ while lower values indicate a more biased model.

3.4 Proposed Method

To identify a *group-agnostic* bias subspace, we use the warmth and competence dictionaries from (Nicolas et al., 2021). To construct the poles of the dimensions, "high" and "low" word pairs (e.g., "able"–"unable" for competence and "sociable"–"unsociable" for warmth) were selected by downsampling to 15 word pairs, per dimension. We use word pairs for each SCM dimension to identify an SCM subspace (see Section 3.1), and subsequently apply the methods from Sec. 3.2.

We test whether SCM-based debiasing can substitute group-specific debiasing simultaneously for gender, race, and age. This is broken down into two related research questions. First, does SCM-based debiasing remove a comparable amount of bias relative to group-specific debiasing? And second, does SCM-based debiasing have more or less of a negative effect on embedding utility (Bolukbasi et al., 2016)? We compare SCM-based debiasing to group-specific debiasing using previous debiasing methods, specifically HD, Sub, LP, and PP (Section 3.2), and evaluate bias as measured by ECT and EQT following Dev and Phillips (2019). In addition, we evaluate the performance of each set of debiased embeddings on established word embedding benchmarks (Jastrzebski et al., 2017).

3.5 Results - Bias Reduction

We investigate whether SCM-based debiasing can simultaneously debias word embeddings with respect to gender, race, and age. For a given bias dimension, we established baselines by applying HD, Sub, LP, and PP using the respective word pair list (e.g., for gender bias we used gender word pairs), denoted with the subscript "same." To place an upper bound on removed bias, we perform PP using gender, race, and age word lists (PP$_{\text{G+R+A}}$).

	Vanilla	HD$_{\text{same}}$	Sub$_{\text{same}}$	Sub$_{\text{SCM}}$	LP$_{\text{same}}$	LP$_{\text{SCM}}$	PP$_{\text{same}}$	PP$_{\text{SCM}}$	PP$_{\text{G+R+A}}$
ECT$_{\text{gender}}$	0.83	0.92	**0.83**	**0.83**	0.82	**0.83**	**0.99**	0.97	0.99
ECT$_{\text{race}}$	0.69	-	**0.51**	**0.52**	0.70	**0.74**	**0.99**	0.96	0.99
ECT$_{\text{age}}$	0.30	-	0.23	**0.34**	**0.60**	0.34	**0.96**	0.95	0.99
EQT$_{\text{gender}}$	0.075	0.056	**0.071**	**0.072**	**0.081**	0.073	**0.063**	0.049	0.059
EQT$_{\text{race}}$	0.042	-	0.032	**0.036**	**0.051**	0.044	**0.061**	0.056	0.073
EQT$_{\text{age}}$	0.052	-	**0.043**	0.041	**0.062**	0.051	**0.063**	0.047	0.057

Table 1: ECT and EQT for gender, race, and age. Subscript "same" denotes the debiasing was performed with respect to the corresponding dimension (e.g. PP$_{\text{same}}$ denotes PP was applied to gender for ECT$_{\text{gender}}$.) and subscript "SCM" refers to debiasing with respect to the SCM subspace. Debiasing was repeated 30 times for each method, and bold values indicate higher scores (per method) with non-overlapping 95% confidence intervals. HD was limited to gender because of other dimensions' lack of equality sets. For experiment with other social groups see §A.6.

	Analogy ↑		Similarity ↑	
	Google	MSR	WS353	RG-65
Vanilla	0.39	0.45	0.50	0.50
PP$_{\text{Gender}}$	0.31	0.36	0.49	0.37
PP$_{\text{Gender+Race}}$	0.27	0.33	0.43	0.30
PP$_{\text{G+R+A}}$	0.25	0.30	0.40	0.27
PP$_{\text{SCM}}$	0.29	0.34	0.42	0.33

Table 2: Embedding utility for debiased models.

For race and age we used the lists from Caliskan et al. (2017), while gender lists were taken from Bolukbasi et al. (2016). All methods were repeatedly applied using 30 different word pair samples, and we report each measure's average and compare values using 95% confidence intervals. Implementation details are provided in the Appendix.

Table 1 shows the results of our experiments. Overall, SCM-based debiasing performs comparably to social-group-specific debiasing across methods. Specifically for ECT, SCM-based debiasing was either better than, or not statistically different from, LP$_{\text{same}}$ and Sub$_{\text{same}}$, while SCM-based debiasing was only slightly out-performed by PP$_{\text{same}}$ (0.01–0.03). In other words, these results demonstrate that warmth and competence dimensions can simultaneously capture gender, race, and age bias in word embeddings. For the EQT, results are somewhat similar to those of ECT; however, we caution against interpreting small differences in EQT due to its definition of biased analogies relying on NLTK to compile comprehensive sets of synonyms and plural forms of words (Dev and Phillips, 2019).

3.6 Results - Word Embedding Utility

Table 1 shows that PP$_{\text{G+R+A}}$ outperformed all other methods on bias evaluations. However, one trade-off is the reduction in word embedding utility. Table 2 shows that PP$_{\text{SCM}}$ preserves more embedding utility than PP$_{\text{G+R+A}}$, using established benchmarks for analogy and similarity (Jastrzebski et al., 2017). Due to the information removed in the debiasing process, as the number of social attributes increases, the quality of embeddings for group-specific debiasing deteriorates; however, this is not the case for PP$_{\text{SCM}}$ showing that PP$_{\text{SCM}}$ preserves some of the definitional biases (e.g. gender bias of actor vs. actress). These results indicate that our proposed approach for SCM-based bias mitigation is a better solution especially when our goal is to remove social biases for as many groups as possible.

4 SCM-Based Bias Mitigation for Contextualize Language Models

Similar to the previous section, before discussing our proposed method, we briefly review bias mitigation efforts in language models.

4.1 Methods of post-hoc bias mitigation

Similar to bias mitigation in static word embeddings, mitigating bias in language models requires a definition of the bias subspace. The bias subspace is defined via a set of *attribute* word tuples $A = \{(a_1^1, a_1^2, ...), (a_2^1, a_1^2, ...)\}$ where a_j^i denotes the j^{th} word for i^{th} subgroup of an attribute (e.g., for religion { ("Muslim"-"Jewish"-"Chrisitan"), ...}). In addition, bias mitigation for language models requires a set of neutral *target* words T (e.g., occupations such as "doctor", "nurse", etc.) Given a language model M_θ the goal is to find $M_{\theta'}$ such that the difference in association of each neutral target word with all subgroups is minimized.

Our goal is to provide a theory-driven framework for bias mitigation that generalizes to many social groups. Therefore, we chose to focus on DPCE and ADEPT, two top-performing post-hoc methods that do not require significant hand-designed resource development, as our baselines.

Problem Definition: Both DPCE and ADEPT view bias mitigation as a downstream task and use a loss $L = L_{bias} + L_{representation}$ that balances bias mitigation – via L_{bias} – and preserves a model's representational power via $L_{representation}$. The formal definitions of L_{bias} and $L_{represenation}$ are provided when we discuss ADEPT and DPCE. In addition, both algorithms collect a set of sentences S^w for each word w to capture the contextual representation of w. We use $E_i(w, s; \theta)$ to show the embedding of a word w in sentence s in the i^{th} layer of the model parameterized by θ. Then the layer-prototype of a word w in i^{th} layer of a model, $e_i(w)$, is defined as the average of $E_i(w, s; \theta)$ for all $s \in S^w$. The prototype of a word is then defined as the average of all layer-prototypes.

DPCE: proposed by Kaneko and Bollegala (2021) is a fine-tuning approach for mitigating biases in a language model. L_{bias} for DPCE is designed to minimize the inner product of attribute word layer-prototypes with embeddings of target words across all model layers.

$$L_{bias} = \sum_{t \in T} \sum_{s \in S^t} \sum_{a \in A} (e_i(a)^T E_i(t; s; \theta_e))^2$$

$L_{representation}$ is defined to minimize the L_2 norm of embeddings before and after debiasing across all layers.

$$L_{representation} =$$
$$\sum_{s \in S^a} \sum_{w \in s} \sum_{i=1}^{N} ||(E_i(w, s; \theta) - E_i(w, s; \theta')||^2$$

ADEPT: Yang et al. (2023) proposed "A Debiasing PrompT" framework (ADEPT). Let $p_{t_j|a^i}$ quantify how much of attribute a^i's (e.g., male gender's) information can be recovered from neutral target word t_j. Also let $P^{a^i} = [p_{t_1|a^i}, p_{t_2|a^i}, ...]$. ADEPT's L_{bias} is designed to reduce the difference, measured by Jensen-Shannon divergence, between relative distances of different attributes to neutral target words. Formally,

$$L_{bias} = \sum_{i,j \in \{1,...,d\}, i<j} \{JS(P^{a(i)} || P^{a(j)})\}$$

$L_{representation}$ for ADEPT is defined to minimize the KL divergence of embeddings before and after debiasing across all layers.

$$L_{representation} = KL(M_\theta(S) || M_{\theta'}(S))$$

4.2 Bias Benchmarks

4.2.1 SEAT

The Sentence Encoder Association Test (SEAT; May et al., 2019) is the extended version of the Word Embedding Association Test (WEAT; Caliskan et al., 2017), which places the WEAT words to the pre-determined sentences and computes the effect size and p-value. Effect sizes closer to zero indicate lower magnitude of bias.

4.2.2 CrowS-Pairs

Crowdsourced Stereotype Pairs (CrowS-Pairs; Nangia et al., 2020) evaluates whether a model gives a higher probability to the stereotypical sentences over the anti-stereotypical sentences. CrowS-Pairs test set consists of pairs of sentences that target explicit expressions of stereotypes by changing the stereotyped word with an anti-stereotype word. The ideal model should achieve a score of 50.

4.2.3 StereoSet

StereoSet (Nadeem et al., 2021) provides three scores, Stereotype Score (SS), Language Modeling Score (LMS), and Idealized CAT (ICAT) Score. For each item in Stereoset, the model should choose between stereotypical, anti-stereotypical or unrelated variations. The SS is the percentage of sentences in which the model prefers stereotypical ones. The LMS is the percentage of sentences the model prefers, stereotypical or anti-stereotypical sentences, over unrelated. The ideal model should get a SS of 50% and LMS of 100%. The ICAT combines the SS and LMS, and the ideal model should get 100% on ICAT.

4.3 Proposed Method

Similar to the SCM-based bias mitigation for static word embeddings, we propose that SCM's warmth and competence can be used to define a social-group-agnostic bias subspace in language models. We operationalize the warmth and competence dimensions by 16 pairs of "high" and "low" words for each dimension from Nicolas et al. (2021). We hypothesize that similar to static word embeddings, SCM-based debiasing in contextualized language models will: 1. reduce bias for multiple social groups comparable to group-specific bias mitigation and generalizes to understudied social groups or attributes (§4.4), and 2. maintain the expressiveness of language models (§4.5). To test our hypotheses, we mitigate bias with DPCE and ADEPT relying on SCM and compare it with social-group-specific debiasing on gender and religion on multiple bias benchmarks (§4.2). In addition to bias benchmarks, we compare the models on the GLUE benchmark (Wang et al., 2018) to evaluate whether the debiased models have the same expressiveness as the original model.

		Race			Gender			Religion	
		S3	S4	S5	S6	S7	S8	SR1	SR2
$\text{BERT}_{\text{LARGE}}$		0.42*	0.41*	0.89*	0.37	0.42*	-0.26	0.01	-0.16
DPCE	Gender	0.39*	0.61*	0.74*	0.72*	-0.20	-0.26	-0.18	-0.20
	Religion	0.30*	0.76*	0.79*	-0.43	0.69*	0.67*	0.01	-0.12
	SCM	0.25*	0.41*	0.32*	0.72*	-0.42	-0.08	-0.24	-0.15
ADEPT	Gender	0.35*	0.25	0.82*	0.70*	-0.34	0.15	0.26	-0.21
	Religion	0.64*	0.54*	0.88*	0.66*	0.44*	0.51*	0.61*	0.17
	SCM	0.37*	0.21	0.63*	0.81*	0.82*	0.50	0.50*	-0.10

Table 3: SEAT effect sizes for selected tests on Race (S3, S4, S5), Gender (S6, S7, S8), and Religion (SR1, SR2). *: $p< 0.01$

Datasets and Experiment Details: For all experiments, we use the same neutral target words as previous debiasing methods (Kaneko and Bollegala, 2021). To mitigate gender bias, we use binary gender words from Zhao et al. (2018), and to mitigate bias on religion, we use ternary religion words from Liang et al. (2020). Words for SCM-based debiasing are included in Appendix 7. We use two corpora to collect sentences for each word: 1. News-Commentary v15 (NC-v15)[2] and 2. BookCorpus (BC)[3]. We collect the sentences for gender words from NC-v15, and for religion and SCM words we use NC-v15 combined with BC. We extract 58,252 neutral sentences and 14,688 attribute sentences for each gender subgroup, 4,949,126 neutral sentences and 6,485 attribute sentences for each religion subgroup, and 4,650,778 neutral sentences and 35,064 attribute sentences for each SCM subgroup. All models are repeatedly debiased using a random sample of sentences.

We debias $\text{BERT}_{\text{LARGE}}$ (Devlin et al., 2019) with DPCE and ADEPT algorithms. All DPCE experiments were conducted on a GeForce GTX 3090 Ti GPU with the same hyperparameters reported in Kaneko and Bollegala (2021). Our experiments show that the DPCE algorithm is sensitive to the number of sentences collected for each attribute and neutral word sentences. For each setting (gender, religion, and SCM), we run DPCE with $|S| = \{100, 500, 1000, 5000\}$ sentences and chose the model with the highest overall ICAT score ($|S| = 100$ for religion and $|S| = 500$ for SCM and gender, see §B.3 for results). All ADEPT experiments use the same hyperparameters proposed by Yang et al. (2023) and were run on a 16GB Tesla V100. The original ADEPT algorithm was proposed for debiasing in a single bias dimension at a time, which we follow for each of the gender- and religion-debiased models. For the SCM-based model, we adopt a coordinate descent-based iteration approach to debias the warmth and competence dimensions together: at each epoch, we first debias the warmth dimension with respect to the neutral words and then debias the competence dimension with respect to the neutral words (see §B.4).

4.4 Results - Bias Reduction

We compare variations of each algorithm (DPCE or ADEPT) separately to disentangle the effect of SCM-based debiasing from the debiasing algorithm (i.e., fine-tuning or prompt-tuning). We evaluate biases of our models using three benchmarks (§4.2) spanning four social groups/attributes of race, gender, religion, and profession.

DPCE + SCM: On Stereoset, our results show that SCM-based bias mitigation achieves a higher ICAT score compared to group-specific debiasing for all categories of gender, profession, race, religion, and overall (Table 4). Table 3 shows our result on SEAT also reflects the same pattern. With the exception of S6, in all cases, SCM-based DPCE results in smaller effect sizes (or is insignificant). Finally, as shown in Table 4, for CrowS-pairs, SCM-based bias mitigation with DPCE achieves a better score on two of the three categories. Altogether, these results demonstrate that for DPCE, SCM-based debiasing can mitigate social biases on multiple social groups/attributes on par with or better than group-specific debiasing even when evaluated on the explicitly targeted group.

ADEPT + SCM: SCM-based bias mitigation with ADEPT performs better than mitigating biases on religion in almost all cases for Stereoset, except for SS on gender. When compared to mitigating gender bias using ADEPT, SCM achieves comparable LMS, SS, and ICAT scores overall. Surprisingly, we observe that SCM-based deibiasing, compared to other models, achieves better scores on race on

[2] https://www.statmt.org/wmt20/translation-task.html.

[3] https://huggingface.co/datasets/bookcorpus

Benchmark	Task	Metric	BERT$_{\text{LARGE}}$	DPCE			ADEPT		
				Gender	Religion	SCM	Gender	Religion	SCM
StereoSet	Gender	LMS ↑	86.5	84.7	75.5	83.3	86.0	85.2	85.4
		SS → 50	63.2	59.9	57.9	58.7	58.3	59.7	59.5
		ICAT ↑	63.6	68	63.5	68.8	71.7	68.7	69.2
	Profession	LMS ↑	84.8	82.2	76.7	82.6	85.1	85.1	85.3
		SS → 50	59.4	57.3	55.7	55.5	56.7	56.3	56.2
		ICAT ↑	68.8	70.3	67.9	73.5	73.6	74.4	74.7
	Race	LMS ↑	83.6	82.9	81	83.5	84.2	83.6	83.6
		SS → 50	57.1	56.3	55.8	55.9	53.0	55.0	54.1
		ICAT ↑	71.7	72.6	71.7	73.7	79.3	75.3	76.6
	Religion	LMS ↑	84.3	80.1	79.9	83.5	82.7	81.9	81.9
		SS → 50	59.9	57.8	58.8	57.9	55.9	58.1	56.3
		ICAT ↑	67.5	67.7	65.8	70.3	72.8	68.6	71.5
	Overall	LMS ↑	84.4	82.8	78.7	83.1	84.7	84.3	84.4
		SS → 50	58.8	57.1	56.1	56.1	55.1	56.2	55.6
		ICAT ↑	69.5	71	69	72.9	76.0	73.9	74.8
CrowS-Pairs	Gender	Score → 50	55.7	53.4	55.7	52.7	50.4	55.73	54.2
	Race	Score → 50	60.4	59	58.4	68.6	61.6	61.94	61.2
	Religion	Score → 50	67.6	69.5	65.7	60.2	65.7	63.81	64.8

Table 4: StereoSet and CrowS-pairs benchmarks on BERT$_{LARGE}$, and debiased models. We debiased the original model by DPCE and ADEPT with three different sets of attributes, Gender, Religion, and SCM. The best score on each algorithm based on the significant level on ten runs is in bold. → 50 means the ideal model score is 50.

SEAT. However, it is the gender-debiased model that achieves the best results on race for stereoset. This could be in part caused by the fact that race items in stereoset contain multiple instances of gender related stereotypes (e.g. "woman often notice how <blank> the egyptian men was" with "sexiest", "generous", and "lobster"). On CrowS-Pairs, SCM achieves better or comparable scores on race and religion to the best models, but mitigating biases on gender seems to be more effective for gender. As for SEAT, on average ADEPT with SCM gets lower effect sizes for race and comparable effect sizes for religion, but ADEPT with gender results in lower effect sizes on gender.

Across all benchmarks, our SCM-based approach consistently achieves a better score compared to at least one of the group-specific models. This suggests that our SCM-base approach is a better solution for minimizing the potential harms of language models for all social groups/attributes, and their potential intersections.

4.5 Results - Natural Language Understanding

Following Kaneko and Bollegala (2021), we used five tasks from GLUE benchmark, Stanford Sentiment Treebank (SST-2; Socher et al., 2013), Microsoft Research Paraphrase Corpus (MRPC; Dolan and Brockett, 2005), Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017), Recognising Textual Entailment (RTE; Dagan et al., 2006; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and Winograd Schema Challenge (WNLI; Levesque et al., 2012). Table 5 shows that debiased models preserve their expressiveness. All models achieve comparable or better scores than the original model on all five tasks. We speculate that this is due to the $L_{representation}$ loss component in both algorithms. This analysis suggest that for LMs SCM-based debiasing not only doesn't deteriorate the models but also improves their performance on some benchmarks.

		SST-2	MRPC	STS-B	RTE	WNLI
BERT$_{\text{LARGE}}$		91.2	90.7/86.8	90.2/90.0	73.3	56.3
DPCE	Gender	92.4	91.3/87.7	90.3/89.8	61.4	47.9
	Religion	93.6	90.2/86.0	90.4/90.0	68.2	50.7
	SCM	93.2	89.4/84.8	90.8/90.5	72.6	47.9
ADEPT	Gender	93	90.5/86.5	89.9/89.6	72.6	56.3
	Religion	93.2	91.0/87.5	89.9/89.6	73.3	56.3
	SCM	93.7	89.6/85.5	89.8/89,7	72.9	56.3

Table 5: GLUE benchmark for Language Models.

5 Conclusion

In this work, we demonstrated the viability of a theory-driven approach to debiasing pre-trained word embeddings and language models. By replacing the ad hoc, social-group-specific component of existing debiasing techniques with a general, theory-driven, social-group-agnostic counterpart, we have solved two concrete problems with prior debiasing work and opened the door for more research into theory-driven approaches. First, SCM-based debiasing was shown to sufficiently reduce bias across social attributes, without relying on any manually constructed group-specific resources. Second, it is evident from our results that SCM-based debiasing is scalable with respect to generic social attributes – and embeddings or LMs debiased with respect to SCM can be thought of as *generally* debiased. Importantly, SCM-based debiasing results in improvements on the quality of the respective word embeddings or language models.

6 Acknowledgement

We would like to thank our anonymous reviewers for their feedback. This research was supported by NSF CAREER BCS-1846531 and DARPA INCAS HR001121C0165. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

7 Limitations

The word embeddings and language models used in this work are trained on contemporary English language, and our social contexts overly contain explicit stereotypes encoded in English. Stereotypes for a specific group can be quite different depending on the language and culture. Although out of the scope of the present work, cross-societal differences in human stereotyping have been shown to be explainable using the SCM framework (Cuddy et al., 2009). Thus, it is fair to posit that our SCM-based framework generalizes to social group biases beyond those in English. Future research is encouraged to replicate our study in non-English languages.

Furthermore, we would like to point out that there exists a catalogue of bias measurements for word embeddings and language models in the field. However, the current catalogue is far from comprehensive in covering social groups even in the contemporary English/American context, with few resources for the intersectionality of groups and attributes (Subramanian et al., 2021; Dhamala et al., 2021). Additionally, some of these measures have been shown to fail robustness checks. Although our current work uses some of the most recently developed ECT and EQT, we believe that few, if any, of these measurements are completely sound nor complete. In our experiments for language models, we tried to measure bias for the same social group or attribute using multiple benchmarks but still found some substantial differences in results across benchmarks. Therefore, we caution against interpreting low bias measurements as evidence of complete bias removal. While developing a new bias measurement scale is not within the scope of this work, we are optimistic that the social psychological theory in which our approach is grounded provides the bedrock for the current evidence of SCM efficacy to hold on future benchmarks.

Unlike bias mitigation methods for static word embeddings, such as partial projection, the post hoc methods of debiasing for large language models can't be trivially applied to mitigate biases for multiple social attributes simultaneously. For DPCE, the formulation allows for mitigating biases on multiple social attributes, but collecting enough sentences from each attribute that do not include any words from other attributes or neutral words (i.e. mutually exclusive sentences) was not possible with the corpora we experimented with. This problem is exacerbated as the number of social attributes grow due to the mutual exclusivity condition. For ADEPT on the other hand, the formulation did not trivially handle multiple dimensions. Hence, we employed a coordinate-descent modification in our experiments to apply ADEPT to SCM (more info in §B.4). We encourage future work to devise data-efficient methods that can mitigate biases on multiple dimensions at the same time.

References

Andrea E Abele and Bogdan Wojciszke. 2007. Agency and communion from the perspective of self versus others. *Journal of personality and social psychology*, 93(5):751.

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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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. *Advances in neural information processing systems*, 29:4349–4357.

Hege H Bye and Henrik Herrebrøden. 2018. Emotions as mediators of the stereotype–discrimination relationship: a bias map replication. *Group Processes & Intergroup Relations*, 21(7):1078–1091.

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

Aylin Caliskan and Molly Lewis. 2022. Social biases in word embeddings and their relation to human cognition. In Morteza Dehghani and Ryan L. Boyd, editors, *Handbook of Language Analysis in Psychology*, pages 478–493. Guilford.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tessa ES Charlesworth, Aylin Caliskan, and Mahzarin R Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

Amy JC Cuddy, Susan T Fiske, Virginia SY Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, et al. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Naomi Ellemers, Susan T Fiske, Andrea E Abele, Alex Koch, and Vincent Yzerbyt. 2020. Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *Proceedings of the National Academy of Sciences*, 117(14):7561–7567.

Naomi Ellemers, Stefano Pagliaro, and Manuela Barreto. 2013. Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, 24(1):160–193.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.

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.

Dmitry Grigoryev, Susan T Fiske, and Anastasia Batkhina. 2019. Mapping ethnic stereotypes and their antecedents in russia: The stereotype content model. *Frontiers in psychology*, 10:1643.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: 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.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Stanisław Jastrzebski, Damian Leśniak, and Wojciech Marian Czarnecki. 2017. How to evaluate word embeddings? on importance of data efficiency and simple supervised tasks. *arXiv preprint arXiv:1702.02170*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The abc of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, 110(5):675.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Lemeng Liang, Yongai Jin, Jie Zhou, and Yu Xie. 2022. Stereotype contents, emotions, and public attitudes: How do chinese people stereotype nations and national groups? *Chinese Journal of Sociology*, 8(1):52–78.

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.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Merrick Osborne, Ali Omrani, and Morteza Dehghani. 2022. The sins of the parents are to be laid upon the children: biased humans, biased data, biased models. *PsyArXiv*.

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.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Mina Schütz, Christoph Demus, Jonas Pitz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2021. DeTox at GermEval 2021: Toxic comment classification. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 54–61, Duesseldorf, Germany. Association for Computational Linguistics.

MohammadHasan Sharifian, Javad Hatami, Seyed Amir Hossein Batouli, and Mohammad Mahdi Fathian Boroujeni. 2022. Citizens of the world: National stereotypes do not affect empathic response in the presence of individuating information. *International Journal of Psychology*, 57(2):251–260.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

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.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ke Yang, Charles Yu, Yi Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. *Proc. Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI2023)*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Proc. The 61st Annual Meeting of the Association for Computational Linguistics (ACL2023) Findings*.

Vincent Yzerbyt. 2018. The dimensional compensation model: Reality and strategic constraints on warmth and competence in intergroup perceptions. In *Agency and communion in social psychology*, pages 126–141. Routledge.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. 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.

A Implementation Details for Static Word Embeddings.

A.1 Training Word Embeddings

We used the Gensim (Rehurek and Sojka, 2011) implementation of Skip Gram with Negative Sampling variant of Word2Vec (Mikolov et al., 2013) to train a 300 dimensional word embedding model on the WikiText-103 (Merity et al., 2017) with a 5-word window. Words with fewer than 5 occurances in the corpus were dropped. Training was done for 5 iterations with 48 threads on a single AMD Ryzen Threadripper 2990WX CPU.

A.2 Debiasing Algorithms

We used Bolukbasi et al. (2016)'s gender word sets and implementation of for HD. For Sub, LP and PP we follow Dev and Phillips (2019). The implementation can be found in the project repository.

A.3 Bias Subspace

The bias subspace used in HD is identical to that of Bolukbasi et al. (2016). Each dimensions' bias subspace for Sub, LP, and PP was the first principal component of C constructed using 8 randomly sampled word pairs from the corresponding dimensions' word pair list (Section A.5).

A.4 Experiments

Each debiasing algorithm for each dimension was conducted 30 times using a random sample of 8 pairs from the corresponding word list.

A.5 Word Pairs

A.5.1 Social Group Word Pairs

Gender nephews, nieces - nephew, niece - males, females - boys, girls - man, woman - sons, daughters - brother, sister - boy, girl - father, mother - guy, gal - male, female - uncle, aunt - himself, herself - uncles, aunts - fathers, mothers - his, her - son, daughter - him, her - men, women - his, hers - he, she - brothers, sisters - from Bolukbasi et al. (2016).

Race Brad, Darnell - Brendan, Hakim - Geoffrey, Jermaine - Greg, Kareem - Brett, Jamal - Neil, Rasheed - Neil, Rasheed - Todd, Tyrone - Allison, Aisha - Anne, Ebony - Carrie, Keisha - Emily, Kenya - Laurie, Latoya - Meredith, Tamika - from Caliskan et al. (2017).

Age Tiffany, Ethel - Michelle, Bernice - Cindy, Gertrude - Kristy, Agnes - Brad, Cecil - Eric, Wilbert - Joey, Mortimer - Billy, Edgar - from Caliskan et al. (2017).

A.5.2 SCM Word Pairs for Static Word Embeddings

Warmth pleasant, unpleasant - liked, disliked - outgoing, shy - sensitive, insensitive - friendliness, unfriendliness - sociable, unsociable - warm, cold - warmth, coldness - honest, dishonest - fair, unfair - loyal, disloyal - right, wrong - criminal, innocent - genuine, fake - reliable, unreliable - from Nicolas et al. (2021).

Competence smart, stupid - competent, incompetent - intelligent, dumb - able, unable - rational, irrational - capable, incapable - aggressive, docile - resilient, nonresilient - motivated, unmotivated - ambitious, unambitious - independent, dependent - determined, inactive - secure, insecure - clever, foolish - dominant, submissive - from Nicolas et al. (2021).

A.6 SCM-based Debiasing for More Social Groups

We replicate our results with pre-trained embeddings, and (2) include additional social groups, we present our results for Word2Vec trained on Google News including additional groups of Asians, Hispanics, and fat vs. thin (Table 6). The results show that our proposed framework succeeds in handling these additional dimensions and generalizes to other embeddings.

	Gender	Black	Asian	Hispanic	Age	Fat-Thin
PP_{same}	0.86	0.56	0.95	0.91	0.51	0.91
PP_{SCM}	0.85	0.78	0.95	0.91	0.78	0.91

Table 6: ECT for each group when the model is debiased along the same dimension (PP_{same}, i.e. debiasing on gender for gender) and using SCM (PP_{SCM})

B Implementation Details for Debiasing Language Models

B.1 SCM Words for Language Models

We used the following words as SCM attributes in DPCE and ADEPT algorithm to mitigate the bias in the models:

Warm: social, warm, popular, good, right, kind, loyal, pleasant, friendly, funny, moral, fair, sympathetic, sensitive, cooperative, innocent, liked, responsible, genuine, polite, trustworthy, reliable, caring, helpful, thoughtful.

Cold: antisocial, cold, unpopular, bad, wrong, mean, treacherous, unpleasant, unfriendly, boring, immoral, unfair, unsympathetic, insensitive, selfish, criminal, disliked, irresponsible, fake, rude, untrustworthy, unreliable, uncaring, unhelpful, inconsiderate.

Competent: able, bright, brilliant, competent, capable, wise, rational, practical, dominant, dependent, confident, active, efficient, ambitious, determined, critical, secure, daring, educated, aggressive, motivated, intelligent, graceful, creative, energetic.

Incompetent: unable, stupid, dumb, incompetent, incapable, unwise, irrational, impractical, submissive, independent, insecure, inactive, inefficient, lazy, doubtful, naive, vulnerable, cautious, uneducated, docile, unmotivated, unintelligent, clumsy, unimaginative, lethargic.

B.2 Gender and Religion words

We used the same word lists for the gender and religion dimensions as Kaneko and Bollegala (2021); Yang et al. (2023) did for our experiments.

B.3 DPCE Sample Size

After collecting sentences from corpora for the three different settings (gender, religion, and SCM - we separated 1000 samples for the evaluation and trained the model on the $|S| = \{100, 500, 1000, 5000\}$. We ran the StereoSet benchmark for different $|S|$ and chose the best model based on the overall ICAT score (Table 7).

B.4 ADEPT + SCM

To debias with ADEPT using SCM, we adopted a coordinate descent-based algorithm. At each epoch, we first neutralized warmth words with respect to the neutral words, then neutralized competence words with respect to those neutral words. These results are reported in the main paper (section 4).

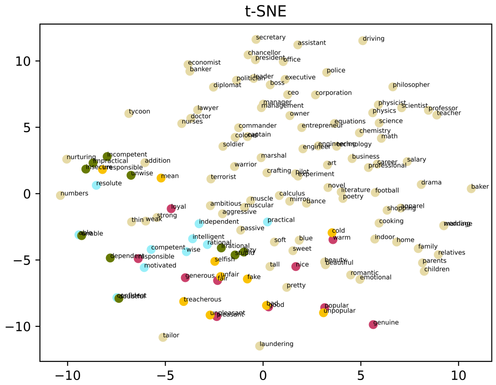

Figure 2: t-SNE plot after running ADEPT based on SCM. Some pairwise words, mostly in the warmth dimension (red/yellow) cluster together, but others, mostly in the competence dimension (green/cyan) do not.

We also experimented with this coordinate descent-based algorithm without using any explicit neutral words. Instead, at each epoch, we neutralized warmth words with respect to competence words, then neutralized competence words with respect to warmth words. In the perfect case of orthogonal warmth and competence axes, this debiasing procedure would hopefully retain all other attributes of words. In our experiments, we found no discernible difference when comparing this to the version with a separate set of neutral words, so we don't report these results.

Figure 2 shows that many of the warmth and competence words indeed do not appear as close as they could be with respect to all the other words used. We speculate that with more effective adaptations of the ADEPT algorithm, SCM-based debiasing with this algorithm might be able to achieve better results on various bias benchmarks.

		$\|S\|$=100			$\|S\|$=500			$\|S\|$=1000			$\|S\|$=5000		
		Gender	Religion	SCM	Gender	Religion	SCM	Gender	Religion	SCM	Gender	religion	SCM
Gender	LMS ↑	47	75.5	84.4	84.7	54.7	83.3	81.7	42.7	81.1	77	58.7	35.6
	SS → 50	52.5	57.9	62.3	59.9	48.4	58.7	57.6	49.9	59.9	57.3	49	50.9
	ICAT ↑	44.7	63.5	63.6	68	53	68.8	69.4	42.7	65.1	65.7	57.5	35
Profession	LMS ↑	49.6	76.7	80.8	82.2	57.7	82.6	72.7	47.3	79.6	56.4	57.8	36.9
	SS → 50	52.5	55.7	58.1	57.3	50.4	55.5	54.8	50.6	53.9	50.7	45.9	50.4
	ICAT ↑	47.2	67.9	67.7	70.3	57.2	73.5	65.8	46.7	73.3	55.5	53.1	36.6
Race	LMS ↑	47.1	81	80.9	82.9	74.8	83.5	72.3	46.2	82.8	57.4	57.2	34.8
	SS → 50	50.2	55.8	59.1	56.3	52.1	55.9	52	46.6	56.6	59.8	54.2	43
	ICAT ↑	47	71.7	66.1	72.6	71.7	73.7	69.4	43.1	71.8	46.1	52.4	29.9
Religion	LMS ↑	41.6	79.9	79.7	80.1	80.6	83.5	72.2	73.7	81.6	54.1	73.7	27.8
	SS → 50	47.7	58.8	60.5	57.8	57.6	57.9	50.1	55.5	55.4	60	58.4	46.8
	ICAT ↑	39.7	65.8	63	67.7	68.3	70.3	72	65.5	72.7	43.3	61.4	26.1
Overall	LMS ↑	47.8	78.7	81.3	82.8	66.2	83.1	73.6	47.2	81.3	59.3	58.2	35.4
	SS → 50	51.2	56.1	59.2	57.1	51.2	56.1	53.6	48.8	56	56.2	50.6	46.9
	ICAT ↑	46.7	69	66.3	71	64.5	72.9	68.3	46.1	71.6	52	57.5	33.2

Table 7: StereoSet benchmark for DPCE with different sample sizes..

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☒ A2. Did you discuss any potential risks of your work?
Our work is on a theory driven, generalizable bias mitigation for language models and in fact addresses some of the potential risks of previous methods of debiasing such as ignoring understudied social groups.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4, appendix

☑ B1. Did you cite the creators of artifacts you used?
3,4,appendix

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We used open-sourced or publicly available corpora, pre-trained word embeddings, and language models

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We only tested pre-trained word embeddings and language models on common NLP benchmarks

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3, 4, appendix

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

3,4,appendix

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
3,4,appendix

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
3,4,appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3,4,appendix

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3,4,appendix

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.