OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings

Sunipa Dev
UCLA

Tao Li
University of Utah

Jeff M Phillips
University of Utah

Vivek Srikumar
University of Utah

sunipa@cs.ucla.edu {tli,jeffmp,svivek}@cs.utah.edu

Abstract

Language representations are known to carry certain associations (e.g., gendered connotations) which may lead to invalid and harmful predictions in downstream tasks. While existing methods are effective at mitigating such unwanted associations by linear projection, we argue that they are too aggressive: not only do they remove such associations, they also erase information that should be retained. To address this issue, we propose OSCaR (Orthogonal Subspace Correction and Rectification), a balanced approach of mitigation that focuses on disentangling associations between concepts that are deemed problematic, instead of removing concepts wholesale. We develop new measurements for evaluating information retention relevant to the debiasing goal. Our experiments on gender-occupation associations show that OSCaR is a well-balanced approach that ensures that semantic information is retained in the embeddings and unwanted associations are also effectively mitigated.

1 Introduction

Word embeddings are used extensively across natural language processing (NLP) and succinctly capture not only the syntactic and semantic structure of language, but also word meaning in context. As such, word embeddings are essential building blocks for today's state-of-the-art in NLP. But they are also known to capture a significant amount of stereotypical associations (e.g., Bolukbasi et al., 2016; Zhao et al., 2017; Dev and Phillips, 2019; Sun et al., 2019) related to gender, race, nationality, or religion, which can manifest in unwanted and/or potentially harmful ways in downstream tasks (Webster et al., 2018; Zhao et al., 2019; Dev et al., 2020). Such potentially problematic associations, when embedded in word representations, can lead to incorrect and unfair decisions about large groups of people. While the term "*bias*" has many meanings, in this paper we use it to refer to these *unwanted stereotypical associations*.

Existing methods to mitigate these effects either require expensive retraining of vectors (Lemoine et al., 2018) which can be inefficient, or projecting out information contained along an entire subspace representing a protected concept (such as gender or race) in the embedding space (e.g., Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020). Projective approaches are difficult to control as they are either insufficient: removing a subspace can still leave residual bias (Gonen and Goldberg, 2019; Lauscher et al., 2020), or too aggressive: in the case of gender, also unnecessarily altering the association between the word *pregnant* and words like *female* and *mother*. In tasks such as coreference resolution, removing such associations could hinder reference resolution.

To quantify how much valid information is retained, we look at the output space of models. Following Dev et al. (2020), we use Natural Language Inference (NLI) as an effective quantitative probe. Here, we construct the hypothesis by making minimal edits to a premise, and observe model prediction conditioned on such changes. For example,

Premise: A *matriarch* sold a watch.
Hypothesis: A *woman* sold a watch.

Here, the objective is to determine if the hypothesis is *entailed* by the premise, *contradicted* by it, or neither (*neutral* to it). A GloVe-based NLI (Parikh et al., 2016) model, without any explicit form of bias mitigation, predicts label *entail* with a high probability of 97%; the notion of a *matriarch* being a *woman* is correctly identified by the model. However, after projective debiasing, the model classifies the pair as *neutral* with a probability 62% while the probability of the label *entail* drops to much lower at 16%. That is, *aggressive mitigation of gender representations erases valid gender associations.*

Ideally, we should correct problematic associations without erasing valid ones. To this end, we

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5034–5050
November 7–11, 2021. ©2021 Association for Computational Linguistics

propose OSCAR (Othogonal Subspace Correction and Rectification) which *orthogonalizes* and *rectifies* identified subspaces of concepts that are incorrectly associated in an embedding space. Embeddings outside the subspace are stretched in a graded manner or untouched. Our contributions are:

1. We argue that mitigating unwanted stereotypical associations should go beyond information removal (e.g., projecting out features), and should also preserve pertinent associations.

2. We present OSCAR[1], a completely different method from the existing projective approaches; it uses orthogonalization of subspaces desired not to have interdependence, and so minimal change is made to embeddings to prevent loss of desired associations.

3. We develop a combination of tests based on NLI to evaluate both bias mitigation and information retention.

4. Our experiments show that OSCAR is a well-balanced approach that mitigates biases as good as projective approaches while retaining more valid associations.

Our contributions, focusing specifically on representations rather than classifiers, are important given the preponderance of distributed representations of text across NLP. Predictions from systems that use such representations, if unchecked, could lead to real-world decisions (involving, e.g. hiring) that "*systematically* and *unfairly* discriminate against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996). In an effort to prevent such transition of representational harms into allocational harms (cf. Crawford, 2017; Abbasi et al., 2019; Blodgett et al., 2020), we look at mitigating stereotyping biases at the source, i.e. the embedding space.

2 Bias, Gender and NLP

Bias and Related Work. Social biases in tasks using machine learning have the potential to cause harms (Barocas and Selbst, 2016) and are widely studied. Such biases in language technologies have been detected (Bolukbasi et al., 2016; Gonen and Goldberg, 2019), measured (Caliskan et al., 2017; Webster et al., 2018; Lauscher et al., 2020) and mitigated (Zhao et al., 2018; Ravfogel et al., 2020). Most work focuses on gender bias, and in particular, the stereotypical associations of occupations to males and females (Rudinger et al., 2018; De-Arteaga et al., 2019; Gaut et al., 2019; Dev et al., 2020). In this work too, bias refers to stereotypical associations in language representations.

Treatment of Gender. The concept of gender is a complex one (cf. Larson, 2017; Dev et al., 2021). In our experiments, we identify a subspace in the representation associated with gender, typically derived by terms specifically associated with notions of male and female. This aligns with prominent recent NLP research (Caliskan et al., 2017; Bolukbasi et al., 2016; Dev and Phillips, 2019; Ravfogel et al., 2020) that try to measure and mitigate gender bias (either implicitly or explicitly), directly focusing on male versus female associations.

The underlying goal of this work is to disassociate two concepts. We mainly use concepts male-vs-female and occupations to demonstrate this. There is explicit and measurable bias encoded in word vector representations (Bolukbasi et al., 2016) by the male-vs-female relation to occupations, and standard evaluation metrics (Caliskan et al., 2017; Dev et al., 2020) are designed to measure this. As such many of our experiments are geared towards these male-vs-female issues.

The notion of gender is not restricted to male and female. But because "conceptualizations of gender as binary are reinforced linguistically in English through pronoun conventions" (Blodgett, 2021), gender notions beyond male and female are either missing or poorly represented. As a result, their associations with other concepts are also not robust. Quantifying this effect, and evaluating its attenuation, is beyond the scope of this paper; however the techniques we study for disassociating two concepts are amenable to such tasks in general. We remark that a potential harm towards non-male/female people arises through omission (Cao and Daumé III, 2020), and as a result, we explicitly encourage further research on this topic.

3 Quantifying Bias and Information Retained

Keeping in mind that removing bias and retaining information have to be done in synergy, we present how to obtain aggregated measurements for these two components. We will first describe the de-

[1] https://github.com/sunipa/OSCaR-Orthogonal-Subspace-Correction-and-Rectification/tree/transformer

sign of our evaluation dataset in §3.1, then present evaluation metrics. Then, we will discuss how to measure biases in §3.2 and information retention in §3.3. To be comprehensive, we will use metrics that are both *intrinsic* to embedding geometry and *extrinsic* ones that focus on downstream NLI model outputs.

3.1 Designing Evaluation Data

Inspired by Dev et al. (2020), we use NLI as a probe to assess the impact of embeddings in model outputs. Consider, for instance:

Premise: The *doctor* bought a bagel.
Hypothesis 1: The *woman* bought a bagel.
Hypothesis 2: The *man* bought a bagel.

Both hypotheses are *neutral* with respect to the premise. However, GloVe, using the decomposible attention model (Parikh et al., 2016), deems that the premise entails hypothesis 1 with a probability 84% and contradicts hypothesis 2 with a probability 91%. Models that use contextualized representations (e.g., ELMo, BERT, and RoBERTa) are no better and perpetrate similarly biased associations. It has also been demonstrated that this effect can be mitigated by methods which project word vectors along the relevant subspace (Dev et al., 2020); these incorrect inferences are reduced, implying a reduction in the association encoded.

But what happens if definitionally gendered information is relayed by a sentence pair? The existing tasks and datasets for embedding quality do not directly evaluate that. For instance:

Premise: The *gentleman* drove a car.
Hypothesis 1: The *woman* drove a car.
Hypothesis 2: The *man* drove a car.

The premise should contradict the first hypothesis and entail the second. We thus expand the use of NLI task as a probe not just to measure the amount of the incorrect gender association expressed, but also the amount of *correct* gender information (or other relevant attributes) expressed. This permits a balanced view in evaluating valid correlations retained and information lost in an explicit manner.

3.2 Measuring Stereotypical Associations

For stereotyping biases, we will use the existing WEAT (Caliskan et al., 2017) as the *intrinsic* metric. For *extrinsic* measurement, we will use the existing NLI data in (Dev et al., 2020) which has $\sim$ 1.9 million sentence pairs generated via template instantiation. We will refer to this data as the *neutrality* dataset since examples there are intentionally constructed to have label *neutral*.

3.3 Measures for Information Preservation

To measure information retention in word embeddings, we define **two new metrics** in this work: the *intrinsic* WEAT* measure, a modified version of WEAT, and the *extrinsic* SIRT measure that captures the downstream impact of word embeddings and its bias-mitigated versions.

New intrinsic measure (WEAT*). We modify WEAT (Caliskan et al., 2017) to measure meaningful male-vs-female associations instead of stereotypical ones. Specifically, we use the following two sets of target words (gendered):

X: {man, male, boy, brother, him, his, son}
Y: {woman, female, girl, sister, her, hers, daughter}

and another two sets of definitionally male-vs-female words, A for male (e.g., *gentleman*) and B for female (e.g. *matriarch*). Our WEAT* test has the same formulation as the original WEAT:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

$$s(w, A, B) = \mu_{a \in A} \cos(a, w) - \mu_{b \in B} \cos(b, w)$$

where, μ denotes the average. The score is then normalized by $\text{stddev}_{w \in X \cup Y} s(w, A, B)$.

Different from WEAT, in WEAT*, since the sets A and B are definitionally male-vs-female, the higher the $s(\cdot)$ value, the more meaningful male-vs-female association is captured. We will present more details in §6 and the Supplement D.7.

New extrinsic measure (SIRT). Existing benchmark datasets for NLI do not explicitly evaluate for gendered information retention. As a result, maintaining a high F1 score on these test sets does not necessarily suggest low information loss.

For a more focused evaluation, we extend the use of the NLI-based probe (Dev et al., 2020) to evaluate correctly labeled male-vs-female information to define the Sentence Inference Retention Test (SIRT). Unlike the original neutrality dataset, here sentences are constructed in a way that the ground truth labels are either *entailment* or *contradiction*.

In our *entailment* dataset, a sentence pair would have words conveying the same gender as the subject words in both the premise and the hypothesis:

Premise: The *lady* bought a bagel.

Hypothesis: The *woman* bought a bagel.

We should note here that not all identically gendered words can be used interchangeably in the premise and hypothesis. For instance, the same premise with subject *mother* should entail the hypothesis with *woman*, but not the opposite. Thus such directional subject pairs are excluded in our data. Our data is generated via template instantiation where we use 12 subjects (6 male and 6 female) in the premise, 4 subjects (*man*, *male*, *woman* and *female*) in the hypothesis, 27 verbs, and 184 objects. The verbs are arranged into categories (e.g., commerce verbs like bought or sold; interaction verbs like spoke etc) to appropriately match objects for coherence (i.e. avoiding absurd instantiations like *The man ate a car*).

For our *contradiction* dataset, we use similar example constructions where the premise clearly contradicts the hypothesis, e.g.,

Premise: The *lady* bought a bagel.
Hypothesis: The *man* bought a bagel.

Unlike the case for entailment, here all words of the other out of male/female set can be used interchangebly in the premise and hypothesis sentences, as all combinations should be contradicted. Our template consists of 16 gendered words (8 male and 8 female) in the premise and the hypothesis, and the same verbs and objects as we used in the *entailment* data.

For each dataset, we have $\sim$ 47 thousand NLI examples. We will obtain aggregated statistics over model predictions on these large datasets. Specifically, to measure performances on the *entailment* dataset, we define **Net Entail** as average of predicted probability mass: $(\sum_i P^e(D_i))/|D|$, where $P^e(\cdot)$ denotes the predicted probability on label *entailment* and $|D|$ is the data size. Furthermore, we count the **Fraction Entail** score as the accuracy of model predictions. Similarly, we define **Net Contradict** and **Fraction Contradict** for our *contradiction* dataset. The higher the values these metrics, the more valid male-vs-female information is retained. Finally, for the *neutrality* data (§3.2), we similarly define **Net Neutral** and **Fraction Neutral**, and the higher scores, the more stereotypical association is mitigated in model predictions.

4 Orthogonal Subspace Correction and Rectification (OSCaR)

We describe a new geometric operation, an alternative to the linear projection-based ones in the literature. This operator applies a graded rotation on the embedding space; it rectifies two identified directions (e.g., male-vs-female and occupations) which should ideally be independent of each other, so they become orthogonal, and remaining parts of their span are rotated at a varying rate so the operation is sub-differentiable.

Given d-dimensional embeddings, OSCaR requires us to first identify two subspaces, denoted V_1 and V_2, for which we desire the embedding space to not have interdependence. We will use subspaces representing male-vs-female (V_1) and occupation (V_2) as running examples, following previous work (Bolukbasi et al., 2016; Caliskan et al., 2017; Dev and Phillips, 2019). Specifically we identify 1-dimensional vectors $v_1 \in V_1$ and $v_2 \in V_2$ best capturing these subspaces from word lists; those used for male-vs-female and occupation are listed in the Supplement. We ensure $v_1 \neq v_2$, and normalize so $\|v_1\| = \|v_2\| = 1$.

We restrict the representation adjustment to the subspace $S = (v_1, v_2)$ that is defined by, and spans the male-vs-female and occupation directions. In particular, we can identify an orthogonal basis for S using vectors v_1 and $v_2' = \frac{v_2 - v_1\langle v_1, v_2\rangle}{\|v_2 - v_1\langle v_1, v_2\rangle\|}$ (so the v_1 and v_2' vectors are orthogonal). We then restrict any word vector $x \in \mathbb{R}^d$ (e.g., *job*) to S as two coordinates $\pi_S(x) = (\langle v_1, x\rangle, \langle v_2', x\rangle)$. We will adjust only these coordinates, and leave all d−2 other orthogonal components fixed.

We now restrict our attention to within the subspace S. In detail, we do this by defining a $d \times d$ rotation matrix U. The first two rows are v_1 and v_2'. The next d−2 rows are any set $u_3, \ldots, u_d$ which complete the orthogonal basis. We then rotate all data vectors x by U (as Ux). Then we manipulate the first 2 coordinates (x_1', x_2') to $f(x_1, x_2)$, described below, and then reattach the last d−2 coordinates, and rotate the space back by U^T.

Next we devise the function f which is applied to each word vector $x \in S$ (we can assume now that x is two-dimensional). See illustration in Figure 1. Given male-vs-female v_1 and occupation v_2 subspaces let $\theta' = \arccos(\langle v_1, v_2\rangle)$ be the angle which captures their correlation, and $\theta = \frac{\pi}{2} - \theta'$. Now define a 2 × 2 rotation matrix which would rotate v_2 to v_2' (orthogonal to v_1).

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

However, the function f should be the identity

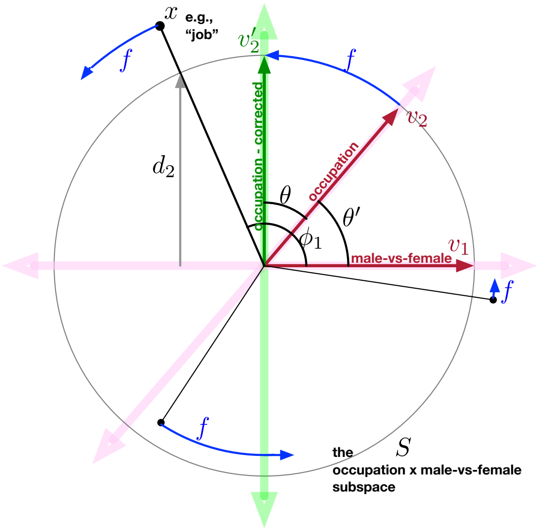

Figure 1: Illustration of OSCaR operation f (blue) in the occupation×male-vs-female subspace S (the span of v_1, v_2).

map for v_1 so $f(v_1) = v_1$, but apply R to v_2 so $f(v_2) = Rv_2$; this maps v_2 to v_2'. And for any other word vector x, it should provide a smooth partial application of this rotation so f is continuous.

In particular, for each data point $x \in S$, we will determine an angle θ_x and apply a rotation matrix $f(x) = R_x x$ defined as

$$R_x = \begin{bmatrix} \cos\theta_x & -\sin\theta_x \\ \sin\theta_x & \cos\theta_x \end{bmatrix}.$$

Towards defining θ_x (how much to rotate a word vector x on this plane), we calculate two other measurements $\phi_1 = \arccos\langle v_1, \frac{x}{\|x\|}\rangle$ and $d_2 = \langle v_2', \frac{x}{\|x\|}\rangle$, which determines which quadrant x is in, in Figure 1. Now we have a case analysis:

$$\theta_x = \begin{cases} \theta\frac{\phi_1}{\theta'} & \text{if } d_2 > 0 \text{ and } \phi_1 < \theta' \\ \theta\frac{\pi-\phi_1}{\pi-\theta'} & \text{if } d_2 > 0 \text{ and } \phi_1 > \theta' \\ \theta\frac{\pi-\phi_1}{\theta'} & \text{if } d_2 < 0 \text{ and } \phi_1 \geq \pi-\theta' \\ \theta\frac{\phi_1}{\pi-\theta'} & \text{if } d_2 < 0 \text{ and } \phi_1 < \pi-\theta'. \end{cases}$$

This effectively orthogonalizes the components of all points exactly along the male-vs-female v_1 and occupation v_2 directions. So, points lying in subspaces especially near v_2 get moved the most, those near v_1 are moved little, and the rest of the embedding space faces a graded rotation. The information contained outside the occupation×male-vs-female subspace S remains the same, thus preserving most of the inherent structure and content of the original embedding representation.

We have an *almost everywhere differentiable* operation applied onto all vectors in the space, enabling us to extend this method to contextualized embeddings. It can now be part of the model specification and integrated with the gradient-based fine-tuning step. Further, it is a post processing step applied onto an embedding space, thus its computational cost is relatively low, and it is easily malleable for a given task for which specific subspaces may be desired as independent.

5 Experimental Methodology

Debiasing methods. Gender association with other concepts and its reduction has been observed on GloVe embeddings (Bolukbasi et al., 2016; Dev and Phillips, 2019; Lauscher et al., 2020; Ravfogel et al., 2020; Rathore et al., 2021, inter alia). However, they are mostly projection-based, and begin by identifying a subspace represented as a vector v. For fair comparison across all these methods, we determine v with embeddings of *he* and *she*. Some methods require an auxillary set of definitionally male/female words G are treated separately (Supplement D.3 and D.4).

Linear Projection (LP): We consider this as the simplest operation. For every word vector w, we project it onto an identified direction v to remove the respective component:

$$w' = w - w\langle w, v\rangle.$$

Afterwards, the w vector (of d-dim) lies in a $(d-1)$-dim subspace, but still retains d coordinates; the subspace v is removed. Lauscher et al. (2020) showed such simple projection reduces the most bias and has the least residual bias.

Hard Debiasing (HD): The original debiasing method, by Bolukbasi et al. (2016), begins with the same projection operation as above, but uses G to only apply it to a subset of all words. First, it trains a linear SVM to separate G from random words. All words labeled as part of G, called $G^\star$, are also assumed definitionally male-vs-female, and not adjusted by projection. The exception is another subset $G' \subset G$ (see Supplement D.3) which come in pairs (e.g., *man-woman*). These words are projected along v, but then are put through another operation called *equalize*. This ensures that after projection, each pair is the same distance apart as they were before projection, but entirely not within the subspace defined by v. As we will observe (similar to Gonen and Goldberg (2019)), equalization and the set $G^\star$ retains certain male-vs-female information in the embedding (compared to projection), but has trouble generalizing when used on words that may carry stereotypical male-vs-female connotations outside of G' (such as proper names).

Iterative Nullspace Projection (INLP): This method (Ravfogel et al., 2020) begins with LP using v on all words except the set G. It then automatically identifies a second set B of most biased words: these are the most extreme words along the direction v (or $-v$). After the operation, it identifies the residual bias by building a linear classifier on B. The normal of this classifier is then chosen as the next direction v_1 on which to apply the next LP operation, removing another subspace. It iterates 35 times, finding v_2 and so on, until no significant residual association can be identified.

OSCAR: Finally, we also apply OSCAR, using *he-she* as v_2, and the subspace defined by an occupation list (see Supplement D.2) as v_1. This subspace is determined by the first principal component of the word vectors in the list. Our code for reproducing experiments will be released upon publication. It is important to note that OSCAR, unlike HD or INLP, does not use any pre-determined lists of words that determine which words to not debias. This is especially advantageous as it is expensive and not possible to demarcate entire vocabularies of languages into two groups—one that has meaningful associations and are not to be debiased, and another list of other words to debias.

Debiasing for contextualized embeddings. The operations above are described for a non-contextualized embedding; we use one of the largest such embeddings GloVe (on 840 B token Common Crawl). They can be applied to contextualized embedding as well; we use RoBERTa (the base version released by Wolf et al. (2019) in (v4.2.1)), the widely adopted state-of-the-art architecture. As advocated by Dev et al. (2020), we only apply the operation (e.g., LP) on the context independent word embeddings that constitute part of the first layer of RoBERTa. Technically, these are subword-embeddings, but all words in G (e.g. *he*, *she*) are full subwords in RoBERTa, and so there is no ambiguity. Ravfogel et al. (2020) extend INLP to contextualized embeddings, but only in a task specific way, whereas we focus on debiasing and evaluating information retention in general. So, for consistency, we only apply INLP on the first (context-independent) layer of RoBERTa, and refer to it as INLP*.

When fine-tuning RoBERTa, we initialize subword embeddings by one of the debiasing methods, and allow gradient to update them. This offers a realistic application scenario where all transformer weights are usually subject to update. It also allows a fair comparison between LP/OSCAR and HD/INLP* since the later are not differentiable.[2]

Extrinsic measurement through NLI. We train our NLI models using the SNLI (Bowman et al., 2015) dataset. While MultiNLI (Williams et al., 2018) contains more complex and diverse sentences than SNLI, making it a good contender for training entailment models, we observed that MultiNLI also carries significantly stronger implicit male-vs-female association. When training on MultiNLI, our model tends to return a significantly higher proportion of stereotyped outcomes (Table S2 in the Supplement). Over 90% of sentence pairs that should have neutral associations are classified incorrectly using both GloVe and RoBERTa. In contrast, using SNLI yields less than 70% unwanted associations and incorrect classifications on the same dataset. Since we focus on the unwanted representational association in word embeddings and not in the datasets used for NLI, using MultiNLI adds a confounding factor. Moreover, MultiNLI has more complex sentences than SNLI, which means that there is a larger probability of having noise via irrelevant information. This would in turn weaken the implication of an incorrect inference of stereotypes expressed.

Keeping train/test separate. Two of the debiasing methods (HD and INLP) use a male-vs-female word list G as an essential step, and the subset $G' \subset G$ of words which are equalized by HD overlap with some words in WEAT (in Supplement D.6), WEAT* (in Supplement D.7), and the extrinsic NLI tests (in Supplement D.5). We allowed this overlap between the training phase for HD and INLP and the testing phase, as the word lists are recommended as ideal for their respective operations by their authors. As a result, HD and INLP may have an advantage in the evaluations.[3]

[2] We also experimented on a more constrained setting where the first layer of embeddings is frozen during fine-tuning. However, we found high variance in the metrics in preliminary experiments.

[3] Importantly, OSCAR does not have any overlap between the words used for training and evaluation.

Since we only use the words *he* and *she* to determine the male-vs-female subspace (which are not used in WEAT or NLI tasks), we can use all other male/female words in the list in the Supplement to generate templates for NLI tasks. We also avoid using *he* and *she* in WEAT*, except for one controlled example (WEAT*(1) in Table S3). For occupations, however, since we use a subset of occupation words to construct the occupation subspace for the OSCAR operation (Supplement D.2), we have disjoint lists of occupations: one set for identifying the subspace and one for testing with WEAT (Supplement D.6) and NLI templates (Supplement D.5).

6 Experimental Results

Intrinsic measurement of bias. Table 1 shows the results on WEAT (Caliskan et al., 2017) between groups of attribute words (i.e. *he* and *she*) and target words (i.e. 3 sets of stereotypes: occupations, work vs. home, and math vs. art). It also shows results for the Embedding Coherence Test (ECT, Dev and Phillips, 2019), showing the association of vectors from X (male/female words) with a list of attribute neutral words Y (occupation) using the Spearman Coefficient. The score ranges in $[-1, 1]$ with 1 being ideal. We see that, across different metrics, OSCAR substantially reduces biases and performs on par with the other best performing methods. Its best performance is on WEAT (1) on occupations, but rectifying occupations still generalizes to other tasks.

	Method	WEAT (1)	WEAT (2)	WEAT (3)	ECT
GloVe	Baseline	1.768	0.535	0.788	0.778
	LP	0.618	0.168	0.282	0.982
	HD	0.241	0.157	0.273	0.942
	INLP	0.495	0.117	0.192	0.844
	OSCAR	0.235	0.170	0.307	0.980

Table 1: **Intrinsic measurement of bias**. Male-vs-female bias contained by embeddings. There are 3 WEAT tests (*a score closer to 0 is better*) results in this table : (1) with stereotypically male/female occupation words, (2) with work versus home related words, and (3) math versus art related words. ECT also measures bias and *a higher score implies less bias*.

Extrinsic measurement of bias. Table 2 shows scores from NLI models with GloVe/RoBERTa before and after they have been debiased by the various approaches. Most debiasing methods, including OSCAR, improve on the neutrality scores (recall sentences should be objectively neutral) without substantially sacrificing the F1 scores on the dev/test sets, except HD on GloVe. INLP appears to be the most successful at debiasing GloVe with multiple steps of projection. OSCAR is always the second best method in each measure. On RoBERTa, OSCAR is significantly more successful at debiasing as compared to the other methods.

	Method	N. Neu	F. Neu	Dev F1	Test F1
GloVe	Baseline	32.1	29.6	87.9	87.3
	LP	38.2	39.7	87.9	87.1
	HD	34.7	32.7	83.4	83.3
	INLP	49.9	53.9	86.4	85.9
	OSCAR	40.0	41.4	87.2	86.9
RoBERTa	Baseline	34.9	32.1	91.2	90.5
	LP	48.9	41.8	91.1	90.8
	HD	45.0	35.6	91.1	90.5
	INLP*	42.8	44.0	91.0	90.8
	OSCAR	56.6	58.8	91.2	90.7

Table 2: **Extrinsic measurement of bias**. Male-vs-female bias expressed downstream by GloVE and RoBERTa embeddings using NLI as a probe. N: Net. F: Fraction. *Higher neutrality scores imply less biases*. OSCAR performs as good as LP on GloVe, and the best on RoBERTa.

Intrinsic metric of male-vs-female information preserved. Table 3 demonstrates how much useful male-vs-female information is retained according to our WEAT* metric after different mitigation methods. To verify our WEAT* evaluation is reliable, we first used random word sets, and after 1000 trials the average WEAT or WEAT* score was 0.001 with standard deviation of 0.33.

The first column uses sets $A = \{he\}$ and $B = \{she\}$ to represent male-vs-female, the same words used to define the mitigation direction. This is the only experiment where the test/train split is not maintained for OSCAR (and other methods). The second column uses two sets of 59 definitionally male and female words as A and B, and the third uses 700 statistically male and female names for each set A and B (both in Supplement D.7). The baseline row (row 1, unmodified GloVe) shows these methods provide similar WEAT* scores.

The LP method retains the least information for tests based on he-she or other male-vs-female words. These scores are followed closely behind by the scores for INLP in these tasks. The single projection of LP retains more info when compared against statistically male and female names, whereas INLP's multiple projections appear to remove more of this information, since only some

	Method	WEAT*(1)	WEAT*(2)	WEAT*(3)
GloVe	Baseline	1.845	1.856	1.874
	LP	0.385	1.207	1.389
	HD	1.667	1.554	0.822
	INLP	0.789	1.368	0.873
	OSCaR	1.361	1.396	1.543

Table 3: **Intrinsic measurement of information retained** using WEAT*. *Larger scores indicate more valid associations expressed.* The test words in columns (1), (2) and (3) respectively, are *he* and *she*, gendered words, and statistically male and female names.

names appear on its pre-specified lists of words to shield from debiasing whereas many male and female words do. HD also performs well for he-she and definitionally male and female words; we suspect the equalize steps allows the information to be retained for these types of words. However, since only some names are pre-specified to be equalized or kept as is, HD loses much more information when compared against statistically male and female names. We use these names in this test as a proxy of many other statistically male and female words which might encode such information but due to the sheer size of the vocabulary, would be impossible to pre-specify exhaustively. Finally, OSCaR performs at or near the best in all evaluations. HD (with its equalize step and pre-specified word lists), performs better for he-she and other male and female words, but this does not generalize to words not specifically set to be adjusted as observed in the names evaluation. In fact, since names do not come in explicit pairs (like uncle-aunt), this equalize step is not even possible.

Extrinsic metric of gendered information preserved. For the new SIRT task we observe the performance using GloVe and RoBERTa in Table 4. GloVe, without any debiasing, which is our baseline, correctly preserves male-vs-female information as seen by the first row. The fraction of correct entails (male premise→male hypothesis, female premise→female hypothesis) and contradicts (male premise→female hypothesis, female premise→male hypothesis) are both high. We see a fall in these scores in all projection based methods (LP, HD and INLP), with a uniform projections step (LP) doing the best among the three. OSCaR does better than all three methods in all four scores measuring valid entailments and contradictions.

RoBERTa, with its different layers learning contextual information, is more equipped at retaining information after being debiased. LP, OSCaR, and HD perform similar to the baseline. Only INLP⋆ with multiple projections registers a noticeable drop in SIRT score.

	Method	N. Ent	F. Ent	N. Con	F. Con
GloVe	Baseline	89.5	96.7	84.0	88.8
	LP	81.0	86.5	71.5	71.3
	HD	66.8	89.1	54.4	75.2
	INLP	74.8	79.3	62.4	63.4
	OSCaR	84.5	91.1	74.7	75.9
RoBERTa	Baseline	94.9	98.4	97.4	97.7
	LP	95.9	99.7	98.9	99.4
	HD	95.1	98.6	98.7	99.3
	INLP⋆	92.8	97.1	95.4	96.4
	OSCaR	95.1	99.0	99.4	99.7

Table 4: **Extrinsic measurement of information retained** using SIRT. N: Net. F: Fraction. *Higher scores indicate more valid information is retained.* Baseline models perform among the best as expected.

7 Discussion & Conclusion

The propagation of undesirable and stereotypical associations learned from data into decisions made by language models maintains a vicious cycle of biases. Combating biases before deploying representations is thus extremely vital. But this poses its own challenges. Word embeddings capture a lot of information implicitly in relatively few dimensions. These implicit associations are what makes them state-of-the-art at tackling different language modeling tasks. Breaking down these associations for bias rectification, thus, has to be done carefully so as to not destroy the structure of the embeddings. OSCaR's rectification of associations helps integrate both these aspects, allowing it to be more suitable at making word embeddings more usable. Moreover, being computationally lightweight and sub-differentiable, it is simple to apply adaptively without extensive retraining.

OSCaR dissociates concepts which may be otherwise lost or overlapped in distributed representations by re-orthogonalizing them. We envision that this method can be used for many types of unwanted associations (age–ability, religion–virtue perception, etc). Since it only decouples specific associations, informative components of these features will remain unchanged. All the experiments in this paper are based on English embeddings, models and tests; extending our analysis to other languages is important future work. Further, we believe OSCAR can extend in a straightforward way beyond NLP to other distributed, vectorized representations (e.g., of images, graphs, or spatial data) which can also exhibit stereotypical associations.

Acknowledgements: We thank the support from NSF grant #2030859 to the Computing Research Association for the CIFellows Project, from NSF #1801446 (SATC), IIS-1927554, NSFCCF-1350888, CNS-1514520, CNS-1564287, IIS-1816149, and Visa Research.

8 Broader Impact

Debiasing word embeddings is a very important and nuanced requirement for making embeddings suitable to be used in different tasks. It is not possible to have exactly one operation that works on all different embeddings and identifies and subsequently reduces all different biases it has significantly. What is more important to note is that this ability is not beneficial in every task uniformly, ie, not all tasks require bias mitigation to the same degree. Further, the debiasing need not apply to all word groups as generally. Disassociating only biased associations in a continuous manner is what needs to be achieved. Hence, having the ability to debias embeddings specifically for a given scenario or task with respect to specific biases is extremely advantageous.

References

Mohsen Abbasi, Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in Representation: Quantifying Stereotyping as a Representational Harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining*.

Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104:671.

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

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

T Bolukbasi, K W Chang, J Zou, V Saligrama, and A Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *ACM Transactions of Information Systems*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

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 Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kate Crawford. 2017. The Trouble with Bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Kimberle Crenshaw. 1991. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

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.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *AAAI*.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies.

Sunipa Dev and Jeff M. Phillips. 2019. Attenuating bias in word vectors. In *AISTATS*, Proceedings of Machine Learning Research, pages 879–887. PMLR.

L Finkelstein, E Gabrilovich, Y Matias, E Rivlin, Z Solan, G Wolfman, and etal. 2002. Placing search in context : The concept revisited. In *ACM Transactions of Information Systems*, volume 20, pages 116–131.

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

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Towards understanding gender bias in relation extraction. *CoRR*, abs/1911.03642.

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 NAACL-HLT 2019*, pages 609–614. 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*.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *AAAI*.

Blake Lemoine, Brian Zhang, and M. Mitchell, editors. 2018. *Mitigating Unwanted Biases with Adversarial Learning*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Technical report, arXiv:1301.3781.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *EMNLP*, pages 2249–2255.

Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, and Bei Wang. 2021. Verb: Visualizing and interpreting bias mitigation techniques for word representations.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL*, pages 8–14.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *CoRR*, abs/1906.08976.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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, Minneapolis, Minnesota. 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, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of EMNLP 2018*.

OSCAR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings
Supplementary Material

A Discussion

Our method of orthogonal correction is easy to adapt to different types of biased associations, such as the good-bad notions attached to different races (Caliskan et al., 2017; Crenshaw, 1991) or religions (Dev et al., 2020) etc. Creating metrics is harder with not as many words to create templates or tests out of, making comprehensive evaluation of bias reduction or information retention harder in these types of biases. We leave that for future exploration.

B Comparison of Debiasing Methods

We have used four distinct debiasing methods in this paper. They differ in some key aspects as listed in Table S1.

While linear projection, hard debiasing and INLP all determine just the gender subspace for debiasing, INLP does so itertively to maximize bias retention. However, the list of words required for the same does not change between iterations. So all of them require a single word list for this segment. OSCAR determines two subspaces - male-vs-female and occupation - in the case of this paper and thus it requires two word lists for this purpose. These word lists are small and contain roughly 10 words.

However, hard debiasing and INLP also require additional word lists for the task of debiasing. These are much longer lists of words. Hard debiaising requires 4 lists: one is a hard coded list of words which helps determine three word lists of which one it debiaises, one it equalizes about the gender axis and the rest it leaves as is. INLP also begins with one hard coded set of words to generate a second set of words it does not debias. This effectively means that both of these methods figure in using a set of hard coded words, a set of words that it protects from debiasing. While effective in many cases, it also is cumbersome in that it is not possible to detect or predict each and every word that might have a meaningful gendered component that should not be debiased. Linear projection and OSCAR do not require any supplementary word lists for the task of debiasing.

Further, since INLP and hard debaisng require these large lists of words for effective debiasing, often in tasks for measuring bias, there is little separation between test and train sets, which could render results for them that are better than in fairer settings.

C Supporting Experiments

C.1 SNLI versus MultiNLI

We compare here the amount of gender bias contained by the templates in SNLI and MultiNLI. While MultiNLI has more complex sentences, it also contains more bias as seen in Table S2 using the standard metrics for neutrality defined by an earlier paper (Dev et al., 2020). Since our work attempts to understand and mitigate the bias in language representation and not the bias in data used in various tasks, we restrict our experiments to SNLI which expresses significantly less bias that MultiNLI.

C.2 Standard Tests for Word Embedding Quality

Non-contextual word embeddings like GloVe or word2vec are characterized by their ability to capture semantic information reflected by valid similarities between word pairs and the ability to complete analogies. These properties reflect the quality of the word embedding obtained and should not be diminished post debiasing. We ascertain that in Table S3. We use a standard word similarity test (Finkelstein et al., 2002) and an analogy test (Mikolov et al., 2013) which measure these properties in word embeddings across our baseline GloVe model and all the debiased GloVe models. All of them perform similarly to the baseline GloVe model, thus indicating that the structure of the embedding has been preserved. While this helps in preliminary

	Linear Projection	Hard Debiasing	INLP	OSCaR
Subspaces Determined	1	1	iterative; hyperparameter	2
Word Lists: Subspaces	1	1	1	2
Word Lists: Debiasing	0	4	3	0

Table S1: Key differences among used debiasing methods.

Embedding	NN	FN	T (0.5)	T(0.7)
GloVe (SNLI)	0.321	0.296	0.186	0.027
GloVe (MNLI)	0.072	0.004	0.0	0.0
RoBERTa (SNLI)	0.338	0.329	0.316	0.139
RoBERTa (MNLI)	0.014	0.002	0.0	0.0

Table S2: Comparison of bias expressed when using two different commonly used datasets - SNLI and MNLI - for natural language inference .

examination of the retention of information in the embeddings, these tests do not contain a large number of sensitive gender comparisons or word pairs. It is thus, not sufficient to claim using this that bias has been removed or that valid gender associations have been retained, leading to the requirement of methods described in the paper and methods in other work (Dev and Phillips, 2019; Dev et al., 2020; Bolukbasi et al., 2016; Caliskan et al., 2017) for the same.

Embedding	GloVe	LP	OSCaR	INLP	HD
WSim	0.697	0.693	0.693	0.686	0.695
GAnalogy	0.674	0.668	0.670	0.663	0.672

Table S3: **Intrinsic Standard Information Tests:** These standard tests evaluate the amount of overall coherent associations in word embeddings. WSim is a word similarity test and Google Analogy is a set of analogy tests.

C.3 Male/Female Names and Debiasing

All these methods primarily use the words ‘he’ and ‘she’ to determine the male-vs-female subspace, though both hard debiasing and INLP use other pre-determined sets of male-vs-female words to guide the process of both debiasing and retention of some male-vs-female information.

Statistically male/female names too have been seen to be good at helping capture the male-vs-female subspace (Dev and Phillips, 2019). We compare in Table S4 the correctly male-vs-female information retention when debiasing is done using projection or correction. We represent simple projection, hard debiasing and INLP by simple projection since it is the core debiasing step used by all three. Both rows have been debiaised using the male-vs-female subspace determined using most common statistically male and female names in Wikipedia (listed in Supplement) and correction uses in addition, the same occupation subspace as used in Table 3. Each value is again a WEAT* calculation where the two sets of words (X and Y) being compared against are kept constant as the same in table 3. The first column of this table, thus represents the association with the subspace determined by ’he - she’ and correction results in a higher association, thus implying that more correctly male-vs-female information retained. We see a similar pattern in columns 2 and 3 which represent other male-vs-female words and male/female names sans the names used to determine the subspace used for debiasing. That correction fares significantly better even among other statistically male and female names reflects the higher degree of precision of information removal and retention.

Embedding	WEAT*(1)	WEAT*(2)	WEAT*(3)
GloVe Proj Names	1.778	1.705	1.701
GloVe OSCaR Names	1.847	1.857	1.875

Table S4: Correctly male-vs-female information contained by embeddings; Larger scores better as they imply more correctly male-vs-female information expressed

C.4 TPR-Gap-RMS

In Table S5, we show the TPR-Gap-RMS metric as used in (Ravfogel et al., 2020) which is an aggregated measurement of male-vs-female bias score over professions. Lower scores imply lesser male-vs-female bias in professions. We refer readers to (Ravfogel et al., 2020) for detailed definition. We follow the same experiment steps, except that we apply different debiasing algorithms to the input word embeddings (instead of the CLS token). This allows us to compare debiasing methods with static use of contextualized embeddings (i.e. without fine-tuning). We see that RoBERTa and RoBERTa HD perform on par while both linear projection and iterative projection methods and our method

OSCaR perform close to each other.

Method	Baselines	LP	HD	INLP	OSCaR
TPR-Gap-RMS	0.19	0.15	0.18	0.13	0.14

Table S5: TPR-Gap-RMS metric of different debiasing method when applied on RoBERTa embeddings.

D Code and Word Lists

D.1 Code

All our code for template generation, debiasing using correction and debiasing contextual embeddings is located at https://github.com/sunipa/OSCaR-Orthogonal-Subspace-Correction-and-Rectification/tree/transformer

D.2 Word Lists for Subspace Determination

Gendered Words for Subspace Determination

he, she

Occupations for Subspace Determination

scientist, doctor, nurse, secretary, maid, dancer, cleaner, advocate, player, banker

D.3 Word Lists for HD

Male-vs-female Words. These are the dictionary-defined male/female words used by the HD operation to determine what words are correctly male-vs-female and which should be neutral in the embedding space. Here is the filtered version of the list G used in our experiments as per our description in §5 about the test/train word list split.

actress, actresses, aunt, aunts, bachelor, ballerina, barbershop, baritone, beard, beards, beau, bloke, blokes, boy, boyfriend, boyfriends, boyhood, boys, brethren, bride, brides, brother, brotherhood, brothers, bull, bulls, businessman, businessmen, businesswoman, chairman, chairwoman, chap, colt, colts, congressman, congresswoman, convent, councilman, councilmen, councilwoman, countryman, countrymen, czar, dad, daddy, dads, daughter, daughters, deer, diva, dowry, dude, dudes, estrogen, fathered, fatherhood, fathers, fella, fellas, females, feminism, fiance, fiancee, fillies, filly, fraternal, fraternities, fraternity, gal, gals, gelding, gentlemen, girlfriend, girlfriends, girls, goddess, godfather, granddaughter, granddaughters, grandma,grandmothers, grandpa, grandson, grandsons, handyman, heiress, hen, hens, her, heroine, hers, herself, him, himself, his, horsemen, hostess, housewife, housewives, hubby, husband, husbands, kings, lad, ladies, lads, lesbian, lesbians, lion, lions, ma, macho, maid, maiden, maids, males, mama, mare, maternal, maternity, men", menopause, mistress, mom, mommy, moms, monastery, monk, monks, motherhood, mothers, nephew, nephews, niece, nieces, nun, nuns, obstetrics, pa, paternity, penis, prince, princes, princess, prostate, queens, salesman, salesmen, schoolboy, schoolgirl, semen, she, sir, sister, sisters, son, sons, sorority, sperm, spokesman, spokesmen, spokeswoman, stallion, statesman, stepdaughter, stepfather, stepmother, stepson, strongman, stud, studs, suitor, suitors, testosterone, uncle, uncles, uterus, vagina, viagra, waitress, widow, widower, widows, wife, witch, witches, wives, womb, women

Equalized Words These male-vs-female words G' are paired (one male, one female) and are equalized by the operation. Here is the filtered version of the list used in our experiments as per our description in §5 about the test/train word list split. Each pair in this list is "equalized".

monastery-convent
spokesman-spokeswoman
priest-nun
Dad-Mom
Men-Women
councilman-councilwoman
grandpa-grandma
grandsons-granddaughters
testosterone-estrogen
uncle-aunt
wives-husbands
Father-Mother
Grandpa-Grandma
He-She
boys-girls
brother-sister
brothers-sisters
businessman-businesswoman
chairman-chairwoman
colt-filly
congressman-congresswoman
dad-mom
dads-moms
dudes-gals
fatherhood-motherhood
fathers-mothers
fella-granny
fraternity-sorority
gelding-mare
gentlemen-ladies
grandson-granddaughter
himself-herself

his-her
king-queen
kings-queens
males-females
men-women
nephew-niece
prince-princess
schoolboy-schoolgirl
son-daughter
sons-daughters

More details about their word lists and code is available at: //github.com/tolga-b/debiaswe

D.4 Word Lists for INLP

Gendered Word List (G) for INLP consists of 1425 words found under https://github.com/Shaul1321/nullspace_projection/blob/master/data/lists/ as the list gender_specific_full.json. This list has been filtered of words used in generating templates (Supplement D.5 and for WEAT (Supplement D.6.

More details about their word lists and code is available at: https://github.com/Shaul1321/nullspace_projection.

D.5 Words Lists for Template Generation

We keep most word lists the same for template generation as used in the paper : For occupations, we change the list to remove those words that we use for subspace determination of occupations for OSCAR. This creates a test/train split for our experiments. We also modify the male-vs-female word lists to create the word lists used in premise/hypothesis for the entail and contradict templates for SIRT.

Male-vs-female words for Neutral Templates

Male: guy, gentleman, man

Female: girl, lady, woman

Male-vs-female words for Entail and Contradict Templates

Male Premise: guy, father, grandfather, patriarch, king, gentleman

Male Hypothesis: man, male

Female Premise: girl, mother, grandmother, matriarch, queen, lady

Female Hypothesis: woman, female

Occupations for Templates accountant, actor, actuary, administrator, advisor, aide, ambassador, architect, artist, astronaut, astronomer, athlete, attendant, attorney, author, babysitter, baker, biologist, broker, builder, butcher, butler, captain, cardiologist, caregiver, carpenter, cashier, caterer, chauffeur, chef, chemist, clerk, coach, contractor, cook, cop, cryptographer, dentist, detective, dictator, director, driver, ecologist, economist, editor, educator, electrician, engineer, entrepreneur, executive, farmer, financier, firefighter, gardener, general, geneticist, geologist, golfer, governor, grocer, guard, hairdresser, housekeeper, hunter, inspector, instructor, intern, interpreter, inventor, investigator, janitor, jester, journalist, judge, laborer, landlord, lawyer, lecturer, librarian, lifeguard, linguist, lobbyist, magician, manager, manufacturer, marine, marketer, mason, mathematician, mayor, mechanic, messenger, miner, model, musician, novelist, official, operator, optician, painter, paralegal, pathologist, pediatrician, pharmacist, philosopher, photographer, physician, physicist, pianist, pilot, plumber, poet, politician, postmaster, president, principal, producer, professor, programmer, psychiatrist, psychologist, publisher, radiologist, receptionist, reporter, representative, researcher, retailer, sailor, salesperson, scholar, senator, sheriff, singer, soldier, spy, statistician, stockbroker, supervisor, surgeon, surveyor, tailor, teacher, technician, trader, translator, tutor, undertaker, valet, veterinarian, violinist, waiter, warden, warrior, watchmaker, writer, zookeeper, zoologist

Objects apple, apron, armchair, auto, bagel, banana, bed, bench, beret, blender, blouse, bookshelf, breakfast, brownie, buffalo, burger, bus, cabinet, cake, calculator, calf, camera, cap, cape, car, cart, cat, chair, chicken, clock, coat, computer, costume, cot, couch, cow, cupboard, dinner, dog, donkey, donut, dress, dresser, duck, goat, headphones, heater, helmet, hen, horse, jacket, jeep, lamb, lamp, lantern, laptop, lunch, mango, meal, muffin, mule, oven, ox, pancake, peach, phone, pig, pizza, potato, printer, pudding, rabbit, radio, recliner, refrigerator, ring, roll, rug, salad, sandwich, shirt, shoe, sofa, soup, stapler, SUV, table, television, toaster, train, tux, TV, van, wagon, watch

Verbs ate, befriended, bought, budgeted for, called, can afford, consumed, cooked, crashed, donated, drove, finished, hated, identified, interrupted, liked, loved, met, owns, paid for, prepared, saved, sold, spoke to, swapped, traded, visited

D.6 Word Lists for WEAT Tests

Male-vs-female Words

Male Gendered: male, man, boy, brother, him, his, son

Female Gendered: female, woman, girl, sister, her, hers, daughter

WEAT 1: Occupations

Stereotypically male: engineer, lawyer, mathematician

Stereotypically female: receptionist, homemaker, nurse

WEAT 2: Work versus Home

Work terms: executive, management, professional, corporation, salary, office, business, career

Home terms: home, parents, children, family, cousins, marriage, wedding, relatives

WEAT 3: Math versus Art

Math terms: math, algebra, geometry, calculus, equations, computation, numbers, addition

Art terms: poetry, art, dance, literature, novel, symphony, drama, sculpture

D.7 Word lists for WEAT* Tests

Definitionally Male-vs-female Words for WEAT*

Male: actor, bachelor, bridegroom, brother, count, czar, dad, daddy, duke, emperor, father, fiance, gentleman, giant, god, governor, grandfather, grandson, headmaster, heir, hero, host, hunter, husband, king, lad, landlord, lord, male, manager, manservant, masseur, master, mayor, milkman, millionaire, monitor, monk, mr, murderer, nephew, papa, poet, policeman, postman, postmaster, priest, prince, shepherd, sir, stepfather, stepson, steward, sultan, uncle, waiter, washerman, widower, wizard,

Female: actress, spinster, bride, sister, countess, czarina, mum, mummy, duchess, empress, mother, fiancee, lady, giantess, goddess, matron, grandmother, granddaughter, headmistress, heiress, heroine, hostess, huntress, wife, queen, lass, landlady, lady, female, manageress, maidservant, masseuse, mistress, mayoress, milkmaid, millionairess, monitress, nun, mrs, murderess, niece, mama, poetess, policewoman, postwoman, postmistress, priestess, princess, shepherdess, madam, stepmother, stepdaughter, stewardess, sultana, aunt, waitress, washerwoman, widow, witch

Statistically male-vs-female Names for WEAT*

There are generated using data curated by social security in USA (https://www.ssa.gov/oact/babynames/). We take the top 1000 names in each male/female gender (as labeled at birth) which are also among the top 100K most frequent words in Wikipedia (to ensure robustly embedded words). Here we present them in lower case.

Male: liam, noah, william, james, logan, benjamin, mason, elijah, oliver, jacob, lucas, michael, alexander, ethan, daniel, matthew, aiden, henry, joseph, jackson, samuel, sebastian, david, carter, wyatt, jayden, john, owen, dylan, luke, gabriel, anthony, isaac, grayson, jack, julian, levi, christopher, joshua, andrew, lincoln, mateo, ryan, nathan, aaron, isaiah, thomas, charles, caleb, josiah, christian, hunter, eli, jonathan, connor, landon, adrian, asher, cameron, leo, theodore, jeremiah, hudson, robert, easton, nolan, nicholas, ezra, colton, angel, jordan, dominic, austin, ian, adam, elias, jose, ezekiel, carson, evan, maverick, bryson, jace, cooper, xavier, parker, roman, jason, santiago, chase, sawyer, gavin, leonardo, jameson, kevin, bentley, zachary, everett, axel, tyler, micah, vincent, weston, miles, wesley, nathaniel, harrison, brandon, cole, declan, luis, braxton, damian, silas, tristan, ryder, bennett, george, emmett, justin, kai, max, diego, luca, carlos, maxwell, kingston, ivan, maddox, juan, ashton, rowan, giovanni, eric, jesus, calvin, abel, king, camden, amir, blake, alex, brody, malachi, emmanuel, jonah, beau, jude, antonio, alan, elliott, elliot, waylon, xander, timothy, victor, bryce, finn, brantley, edward, abraham, patrick, grant, hayden, richard, miguel, joel, gael, tucker, rhett, avery, steven, graham, jasper, jesse, matteo, dean, preston, august, oscar, jeremy, alejandro, marcus, dawson, lorenzo, messiah, zion, maximus, river, zane, mark, brooks, nicolas, paxton, judah, emiliano, bryan, kyle, myles, peter, charlie, kyrie, thiago, brian, kenneth, andres, lukas, aidan, jax, caden, milo, paul, beckett, brady, colin, omar, bradley, javier, knox, jaden, barrett, israel, matias, jorge, zander, derek, holden, griffin, arthur, leon, felix, remington, jake, killian, clayton, sean, riley, archer, legend, erick, enzo, corbin, francisco, dallas, emilio, gunner, simon, andre, walter, damien, chance, phoenix, colt, tanner, stephen, tobias, manuel, amari, emerson, louis, cody, finley, martin, rafael, nash, beckham, cash, reid, theo, ace, eduardo, spencer, raymond, maximiliano, anderson, ronan, lane, cristian, titus, travis, jett, ricardo, bodhi, gideon, fernando, mario, conor, keegan, ali, cesar, ellis, walker, cohen, arlo, hector, dante, garrett, donovan, seth, jeffrey, tyson, jase, desmond, gage, atlas, major, devin, edwin, angelo, orion, conner, julius, marco, jensen, peyton, zayn, collin, dakota, prince, johnny, cruz, hendrix, atticus, troy, kane, edgar, sergio, kash, marshall, johnathan, romeo, shane, warren, joaquin, wade, leonel, trevor, dominick, muhammad, erik, odin, quinn, dalton, nehemiah, frank, grady, gregory, andy, solomon, malik, rory, clark, reed, harvey, jay, jared, noel, shawn, fabian, ibrahim, adonis, ismael, pedro, leland, malcolm, alexis, porter, sullivan, raiden, allen, ari, russell, princeton, winston, kendrick, roberto, lennox, hayes, finnegan, nasir, kade, nico, emanuel, landen, moises, ruben, hugo, abram, adan, khalil, augustus, marcos, philip, phillip, cyrus, esteban, albert, bruce, lawson, jamison, sterling, damon, gunnar, luka, franklin, ezequiel, pablo, derrick, zachariah, cade, jonas, dexter, remy, hank, tate, trenton, kian, drew, mohamed, dax, rocco, bowen, mathias, ronald, francis, matthias, milan, maximilian, royce, skyler, corey, drake, gerardo, jayson, sage, benson, moses, rhys, otto, oakley, armando, jaime, nixon, saul, scott, ariel, enrique, donald, chandler, asa, eden, davis, keith, frederick, lawrence, leonidas, aden, julio, darius, johan, deacon, cason, danny, nikolai, taylor, alec, royal, armani, kieran, luciano, omari, rodrigo, arjun, ahmed, brendan, cullen, raul, raphael, ronin, brock, pierce, alonzo, casey, dillon, uriel, dustin, gianni, roland, kobe, dorian, emmitt, ryland, apollo, roy, duke, quentin, sam, lewis, tony, uriah, dennis, moshe, braden, quinton, cannon, mathew, niko, edison, jerry, gustavo, marvin, mauricio, ahmad, mohammad, justice, trey, mohammed, sincere, yusuf, arturo, callen, keaton, wilder, memphis, conrad, soren, colby, bryant, lucian, alfredo, cassius, marcelo, nikolas, brennan, darren, jimmy, lionel, reece, ty, chris, forrest, tatum, jalen, santino, case, leonard, alvin, issac, bo, quincy, mack, samson, rex, alberto, callum, curtis, hezekiah, briggs, zeke, neil, titan, julien, kellen, devon, roger, axton, carl, douglas, larry, crosby, fletcher, makai, nelson, hamza, lance, alden, gary, wilson, alessandro, ares, bruno, jakob, stetson, zain, cairo, nathanael, byron, harry, harley, mitchell, maurice, orlando, kingsley, trent, ramon, boston, lucca, noe, jagger, randy, thaddeus, lennon, kannon, kohen, valentino, salvador, langston, rohan, kristopher, yosef, lee, callan, tripp, deandre, joe, morgan, reese, ricky, bronson, terry, eddie, jefferson, lachlan, layne, clay, madden, tomas, kareem, stanley, amos, kase, kristian, clyde, ernesto, tommy, ford, crew, hassan, axl, boone, leandro, samir, magnus, abdullah, yousef, branson, layton, franco, ben, grey, kelvin, chaim, demetrius, blaine, ridge, colson, melvin, anakin, aryan, jon, canaan, dash, zechariah, alonso, otis, zaire, marcel, brett, stefan, aldo, jeffery, baylor, talon, dominik, flynn, carmelo, dane, jamal, kole, enoch, kye, vicente, fisher, ray, fox, jamie, rey, zaid, allan, emery, gannon, rodney, sonny, terrance, augustine, cory, felipe, aron, jacoby, harlan

Female: emma, olivia, ava, isabella, sophia, mia, charlotte, amelia, evelyn, abigail, harper, emily, elizabeth, avery, sofia, ella, madison, scarlett, victoria, aria, grace, chloe, camila, penelope, riley, layla, lillian, nora, zoey, mila, aubrey, hannah, lily, addison, eleanor, natalie, luna, savannah, brooklyn, leah, zoe, stella, hazel, ellie, paisley, audrey, skylar, violet, claire, bella, aurora, lucy, anna, samantha, caroline, genesis, aaliyah, kennedy, kinsley, allison, maya, sarah, adeline, alexa, ariana, elena, gabriella, naomi, alice, sadie, hailey, eva, emilia, autumn, quinn, piper, ruby, serenity, willow, everly, cora, lydia, arianna, eliana, peyton, melanie, gianna, isabelle, julia, valentina, nova, clara, vivian, reagan, mackenzie, madeline, delilah, isla, katherine, sophie, josephine, ivy, liliana, jade, maria, taylor, hadley, kylie, emery, natalia, annabelle, faith, alexandra, ximena, ashley, brianna, bailey, mary, athena, andrea, leilani, jasmine, lyla, margaret, alyssa, arya, norah, kayla, eden, eliza, rose, ariel, melody, alexis, isabel, sydney, juliana, lauren, iris, emerson, london, morgan, lilly, charlie, aliyah, valeria, arabella, sara, finley, trinity, jocelyn, kimberly, esther, molly, valerie, cecilia, anastasia, daisy, reese, laila, mya, amy, amaya, elise, harmony, paige, fiona, alaina, nicole, genevieve, lucia, alina, mckenzie, callie, payton, eloise, brooke, mariah, julianna, rachel, daniela, gracie, catherine, angelina, presley, josie, harley, vanessa, parker, juliette, amara, marley, lila, ana, rowan, alana, michelle, malia, rebecca, summer, sloane, leila, sienna, adriana, sawyer, kendall, juliet, destiny, diana, hayden, ayla, dakota, angela, noelle, rosalie, joanna, lola, georgia, selena, june, tessa, maggie, jessica, remi, delaney, camille, vivienne, hope, mckenna, gemma, olive, alexandria, blakely, catalina, gabrielle, lucille, ruth, evangeline, blake, thea, amina, giselle, melissa, river, kate, adelaide, vera, leia, gabriela, zara, jane, journey, miriam, stephanie, cali, ember, logan, annie, mariana, kali, haven, elsie, paris, lena, freya, lyric, camilla, sage, jennifer, talia, alessandra, juniper, fatima, amira, arielle, phoebe, ada, nina, samara, cassidy, aspen, allie, keira, kaia, amanda, heaven, joy, lia, laura, lexi, haley, miranda, kaitlyn, daniella, felicity, jacqueline, evie, angel, danielle, ainsley, dylan, kiara, millie, jordan, maddison, alicia, maeve, margot, phoenix, heidi, alondra, lana, madeleine, kenzie, miracle, shelby, elle, adrianna, bianca, kira, veronica, gwendolyn, esmeralda, chelsea, alison, skyler, magnolia, daphne, jenna, kyla, harlow, annalise, dahlia, scarlet, luciana, kelsey, nadia, amber, gia, carmen, jimena, erin, christina, katie, ryan, viviana, alexia, anaya, serena, ophelia, regina, helen, remington, cadence, royalty, amari, kathryn, skye, jada, saylor, kendra, cheyenne, fernanda, sabrina, francesca, eve, mckinley, frances, sarai, carolina, tatum, lennon, raven, leslie, winter, abby, mabel, sierra, april, willa, carly, jolene, rosemary, selah, renata, lorelei, briana, celeste, wren, leighton, annabella, mira, oakley, malaysia, edith, maryam, hattie, bristol, demi, maia, sylvia, allyson, lilith, holly, meredith, nia, liana, megan, justice, bethany, alejandra, janelle, elisa, adelina, myra, blair, charley, virginia, kara, helena, sasha, julie, michaela, carter, matilda, henley, maisie, hallie, priscilla, marilyn, cecelia, danna, colette, elliott, cameron, celine, hanna, imani, angelica, kalani, alanna, lorelai, macy, karina, aisha, johanna, mallory, leona, mariam, karen, karla, beatrice, gloria, milani, savanna, rory, giuliana, lauryn, liberty, charli, jillian, anne, dallas, azalea, tiffany, shiloh, jazmine, esme, elaine, lilian, kyra, kora, octavia, irene, kelly, lacey, laurel, anika, dorothy, sutton, julieta, kimber, remy, cassandra, rebekah, collins, elliot, emmy, sloan, hayley, amalia, jemma, jamie, melina, leyla, wynter, alessia, monica, anya, antonella, ivory, greta, maren, alena, emory, cynthia, alia, angie, alma, crystal, aileen, siena, zelda, marie, pearl, reyna, mae, zahra, jessie, tiana, armani, lennox, lillie, jolie, laney, mara, joelle, rosa, bridget, liv, aurelia, clarissa, elyse, marissa, monroe, kori, elsa, rosie, amelie, eileen, poppy, royal, chaya, frida, bonnie, amora, stevie, tatiana, malaya, mina, reign, annika, linda, kenna, faye, reina, brittany, marina, astrid, briar, teresa, hadassah, guadalupe, rayna, chanel, lyra, noa, laylah, livia, ellen, meadow, ellis, milan, hunter, princess, nathalie, clementine, nola, simone, lina, marianna, martha, louisa, emmeline, kenley, belen, erika, lara, amani, ansley, salma, dulce, nala, natasha, mercy, penny, ariadne, deborah, elisabeth, zaria, hana, raina, lexie, thalia, annabel, christine, estella, adele, aya, estelle, landry, tori, perla, miah, angelique, romina, ari, jaycee, kai, louise, mavis, belle, lea, rivka, calliope, sky, jewel, paola, giovanna, isabela, azariah, dream, claudia, corinne, erica, milena, alyson, joyce, tinsley, whitney, carolyn, frankie, andi, judith, paula, amia, hadlee, rayne, cara, celia, opal, clare, gwen, veda, alisha, davina, rhea, noor, danica, kathleen, lindsey, maxine, paulina, nancy, raquel, zainab, chana, lisa, heavenly, patricia, india, paloma, ramona, sandra, abril, vienna, rosalyn, hadleigh, barbara, jana, brenda, casey, selene, adrienne, aliya, miley, bexley, joslyn, zion, breanna, melania, estrella, ingrid, jayden, kaya, dana, legacy, marjorie, courtney, holland