From Inclusive Language to Inclusive AI: A Proof-of-Concept Study into Pre-Trained Models

Marion Bartl[1,2,*], Susan Leavy[1,2]

[1]Insight SFI Research Centre for Data Analytics
[2]School of Information and Communication Studies, University College Dublin, Belfield, Dublin 4, Ireland

Abstract

Pre-trained language models are central to today's AI landscape. However, harmful and outdated gender stereotypes can be learned from training data and ingrained into these models. Since pre-trained models are used in many everyday language-based technologies, the deployment of unchecked systems risks the perpetuation of stereotypical and heteronormative conceptualizations of gender in society and could result in biased AI-driven decisions. In this work, we present a study into the effects of data curation to mitigate such gender bias. We use language that counteracts male-centric expressions and structures in favor of inclusivity across all gender identities. This line of interdisciplinary research has received little attention in NLP in the past, despite the fact that gender-inclusive language has been a central tenet within feminist linguistics over five decades. For this study we rewrite gender-specific pronouns using the gender-neutral *they* pronoun and replace gendered role nouns for gender-inclusive variants. Our findings show a reduction in gender stereotyping for English word embedding models and a disruption of latent gender associations of gender-neutral words. This work demonstrates how incorporating principles of gender inclusive language can mitigate risks of bias in AI.

Keywords

gender-inclusive language, feminist AI, gender bias, pre-trained models

1. Introduction

Language models significantly impact society. They are ubiquitous in applications ranging from search engines to hiring systems. State-of-the-art models like GPT-4 [1] and LLama2 [2] dominate current research due to their high performance. However, earlier models, such as classic pre-trained embeddings (Word2Vec [3], GloVe [4]) and smaller-scale language models (BERT [5]), remain in industrial use for their cost-effectiveness due to fast computation and memory efficiency [6].

All these pre-trained representations present a significant issue: they encode concepts of gender derived from training data that mirror existing patterns of inequality and discrimination [7]. These biased representations can reinforce discriminatory patterns through generated language or influence hiring decisions that perpetuate gender imbalances based on stereotypes [8]. To build trustworthy and fair language technologies, we must therefore ensure that the training language is fair from the outset.

One approach to achieving this is through training with gender-inclusive language, which has three main aims: (1) Avoiding the use of masculine terms generically to refer to people of unknown gender or groups of mixed gender (e.g. *mankind→humankind*, *to man→to staff*), (2) eliminating irrelevant gender distinctions such as in *headmistress/headmaster→headteacher* [9], and (3) establishing a trans-inclusive model of gender that includes references beyond binary categories, such as the use of singular *they* or neopronouns [10]. Gender-inclusive language has a long research tradition in feminist linguistics [11, 12, 13] and has recently become a focus in research on gender bias in NLP. Examples include gender-neutral rewriting models [14, 15] and gender-inclusive language as a means of combating misgendering in translation [16] or as a fine-tuning strategy for reducing gender stereotyping in Large Language Models (LLMs) [17, 18].

Most of these works, however, make an assumption that equality-promoting effects of gender-inclusive language, such as reduction in gender stereotyping and discrimination, can be directly picked

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain
marion.bartl@insight-centre.org (M. Bartl); susan.leavy@ucd.ie (S. Leavy)
0000-0002-8893-4961 (M. Bartl); 0000-0002-3679-2279 (S. Leavy)

original text	As a **fireman**, Zachary is always ready to help people, but since **his** parents' relationship was marked by conflict, **he** is opposed to commitments.
after rewriting	As a **firefighter**, Zachary is always ready to help people, but since **their** parents' relationship was marked by conflict, **they** are opposed to commitments.

Table 1
Example sentence from Wikipedia subcorpus before and after gender-neutral rewriting

up by LLMs through fine-tuning [17]. However, fine-tuning an LLM necessarily invites interference from the pre-trained model, which might obscure conclusions on how gender-inclusive language is incorporated into model representations of gender. This work therefore presents a foundation-level proof-of-concept study with classic pre-trained embeddings. These allow us to train a model with gender-neutral text from scratch. By contrast, training an LLM from scratch goes beyond our and many other institution's resources [8]. Further, word embedding models might still be used in small-scale industry settings due to their low computational costs, which makes them relevant [6].

We train two Word2Vec embedding models [3] on unchanged vs. gender-neutral English text, additionally comparing against a common post-hoc debiasing technique [19]. The code for our experiments is openly accessible[1]. In the experiments we find that the use of gender-neutral terminology reduces gender stereotyping as measured by the Word Embedding Association Test [20] and the Embedding Coherence Test [21] as well as reducing latent gender information in the embeddings of gender-neutral words. These results demonstrate how incorporating principles of gender-inclusive language, which were designed to help people avoid bias or discrimination in how they speak or write, can have the same effect on how gender is represented in word embedding models.

2. Methodology

2.1. Data Collection

Our experiments were conducted on a corpus introduced by [22], the *Small Heap*. The corpus is made up of random subsections of three popular LLM training corpora: OpenWebText2 [23] (50%), CC-News [24] (30%) and English Wikipedia (20%). The final sub-corpus contains ~250 million tokens, or 1.5 GB of text.

2.2. Gender-neutral Rewriting

The corpus was edited using the *NeuTralRewriter* [14]. This involved replacing gender-specific pronouns (*he*, *she*, *him* etc.) with the corresponding variant of the gender-neutral pronoun *they*. Additionally, 91 gender-specific nouns (*headmaster*, *mankind*, etc.) including plural and spelling variants were replaced by neutral versions (*principal*, *humankind*, etc.; for full set see 14). Table 1 shows an example of rewritten text.

There are two implementations of the *NeuTral Rewriter*, a rule-based version and a neural, machine translation-based model. While the neural model performed better in the original experiments [14], it proved to be very susceptible to noise in our data (email addresses, digits, etc.) as well as low-frequency words, often translating them into unintelligible text. We therefore used the rule-based implementation, which uses a combination of word, part-of-speech and dependency information to derive the correct replacement of pronouns.

2.3. Embedding Models

In order to evaluate the effects of gender-neutral language on representations of gender within the corpus, we built three different Word2Vec models [3]. The first was trained on the original and the second on the rewritten corpus. Each Word2Vec model was trained using the Continuous Bag of Words (CBOW) algorithm with the default hyperparameters of the gensim library's *Word2Vec* class [25]. The

[1] https://github.com/marionbartl/ILIA

third model was created by performing *hard debiasing* [19] on our original model in order to compare our method to an existing, model-based debiasing method. Hard debiasing modifies embeddings in such a way that gender-neutral words (e.g. *babysit*) are equidistant to gendered word pairs (e.g. *grandfather* – *grandmother*). Additionally, the gendered component of embeddings of gender-neutral words, as defined by what is termed the 'gender subspace', is set to zero.

2.4. Bias Evaluation

The three trained embedding models were analyzed for underlying gender bias using three methods. Previous research found that bias measures are not always consistent [26, 27]. Using a composition of metrics therefore allows for a more comprehensive evaluation.

The **Word Embedding Association Test (WEAT)** is one of the most commonly applied bias measures for word embeddings [20]. The test is modelled after a psychological assessment, the Implicit Associations Test [28], and measures bias by computing the mean association between two sets of target and attribute words. We used a WEAT implementation by Lauscher et al. [29]. Each WEAT test (i.e. the specific combination of target and attribute terms) is identified as WN with N corresponding to its position in the original WEAT paper. W9 and W10 were added by Lauscher et al. [29]. We additionally added two tests using attribute words related to male- and female-stereotypical professions (WA) as well as words related to computer science and childcare (WB, cf. Table 4).

Clustering and Classification into two groups was used by Gonen and Goldberg [26] to show that embedding spaces retain gender information despite application of debiasing. We measured cluster integrity after K-Means clustering (averaged over 50 runs) as well as classification accuracy with an SVM (trained for 20 epochs) in order to find out how well gender information can be recovered from the embedding space. We use the original model's 500 most male-/female-biased words according to their similarity to the element-wise mean of the male/female attribute embeddings A_1 and A_2 of W8 (cf. Table 5).

The **Embedding Coherence Test (ECT)** calculates distances between two sets of gendered target words T_1,T_2 and a set of attribute words A that relate to a societal gender imbalance (e.g. *captain*, *football*) [21]. Instead of relying on the absolute distances, the ECT calculates the Spearman coefficient between the ranked distances for T_1 and A vs. T_2 and A. A high coefficient indicates similar ranks between the two gendered sets, signifying reduced bias. We used the ECT implementation by [29].

Semantic Quality is evaluated following Lauscher et al. [29], by using the similarity benchmarks SimLex-999 [30] and WordSim-353 [31] and computing the Pearson and Spearman correlation coefficients between the benchmark term-pair similarities and cosine similarities of the corresponding embedding pairs of the respective models.

3. Findings and Discussion

We will discuss the results for our chosen gender bias metrics: WEAT [20], ECT [21] and Clustering and Classification [26]. Finally, we contextualize these findings with the performance of our models on two semantic quality benchmarks.

WEAT: Results indicated a reduction in gender bias related to stereotypical associations of women with arts, domestic work, and childcare, and men with (computer) science, maths, and careers, respectively. All five tests measuring gender bias, W6, W7, W8, WA, and WB, showed a reduction in the statistic after rewriting (Table 2). WA additionally shows that there is a reduction in the association of feminine/masculine words with traditionally gendered professions. Comparing the WEAT scores after rewriting to the hard debiased embeddings, one can see that the scores for the hard debiased embeddings are approaching zero, which indicates equal association of male/female attributes with the respective targets. Thus, on one hand, training with gender-neutral language generally leads to a reduction in WEAT bias scores, indicating that this change in the language can lead to a reduction in associations based on stereotypes. On the other hand, stereotyped associations can be specifically targeted and mostly removed post-hoc.

			Cohen's *d*			effect size		
#	**targets**	**attributes**	**pre**	**post**	**h.d.**	**pre**	**post**	**h.d.**
W6	M vs F names	career vs family	1.28*	1.06*	0*	1.95	1.92	1.69
W7	math vs arts	M vs F	0.26*	0.23*	0*	1.37	1.51	1.42
W8	science vs arts	M vs F	0.2*	0.19*	0*	0.86	1.04	1.18
W*A*	M professions vs F professions	M vs F	0.97*	0.76*	0.19*	1.49	1.48	0.42
W*B*	CS vs childcare	M vs F	0.81*	0.69*	0*	1.87	1.18	0.97
W1	flowers vs insects	pleasant vs unpleasant	1.95*	1.82*	2.03*	1.51	1.44	1.52
W2	instruments vs weapons	pleasant vs unpleasant	2.97*	2.81*	2.98*	1.62	1.62	1.62
W9	disease vs health	uncontrollable vs controllable	0.53*	0.58*	0.53*	1.47	1.51	1.48
W10	young vs old names	pleasant vs unpleasant	0.12	0.11	0.13	0.39	0.34	0.46

Table 2
Results for WEAT before and after rewriting. Results marked * significant with $p < 0.05$. Zero values indicate $d \approx 0$. h.d. = hard debiased; CS = computer science.

ECT			
attributes	**pre**	**post**	**h.d.**
arts vs science	0.51*	0.50*	0.6*
M vs F professions	0.47	0.56*	0.8*
CS vs childcare	0.17	0.09	0.15
Clustering & Classification			
K-Means	0.66	0.64	0.93
SVM	0.98	0.96	1.0

(a) ECT results marked * significant with $p < 0.05$. CS = computer science.

		pre	**post**	**h.d.**
Pearson	**SimLex 999**	0.38	0.37	0.38
	WordSim 353	0.72	0.71	0.72
Spearman	**SimLex 999**	0.36	0.37	0.36
	WordSim 353	0.71	0.71	0.71

(b) Semantic quality of W2V embeddings before and after re-writing. All results significant with $p < 0.01$.

Table 3
Results for ECT (Spearman's rank correlation *r*), Clustering & Classification accuracy, and semantic quality. h.d. = hard debiased.

ECT: A reduction in bias was demonstrated in relation to gendered associations with professions. Table 3a shows that for arts vs science and profession attributes, the ECT scores increase, both after rewriting and hard debiasing. However, the ECT scores show a higher increase for the hard debiased model, suggesting an advantage of this method over neutral rewriting. For the computer science vs. childcare attributes however, both methods show a reduction in ECT scores, which could indicate that neither neutral rewriting nor hard debiasing are sufficiently affecting words in these semantic fields.

Clustering and Classification: Our results demonstrated that while the embeddings clearly encode gender information that is very salient to a binary classifier, rewriting with gender-neutral terminology has a more comprehensive effect than focusing on removing a limited 'gender subspace', as done by hard debiasing. In fact, hard debiasing improves the clustering accuracy, indicating that gender in word embeddings is more intricately encoded than can be captured by a gender subspace. These results mirror the findings of Gonen and Goldberg [26]. Both clustering and classification accuracy marginally decline in the model trained on gender-neutral text, as shown in Table 3a. The SVM shows a very high accuracy of 98% in separating words with male and female direct bias in the unchanged and hard-debiased model, which is reduced to 96% in the gender-neutral model.

Semantic Quality: The semantic quality of the word embeddings as measured by the SimLex 999 [30] and WordSim 353 [31] benchmarks dropped only minimally by at most 0.01 points after rewriting (Table 3b). Overall, these results fall only slightly behind larger embedding models. According to the SimLex 999 leaderboard, a Word2Vec model trained on one billion words of Wikipedia text reached a

Spearman correlation of 0.37[2], which is similar to our model.

4. Limitations and Future Work

The first limitation pertains to the **size of the data** used. Our corpus contains 250 million tokens, which is less than 25% of the training data size for a common embedding model [3]. We will explore in future work whether our findings hold for larger datasets and whether the measured reduction in gender stereotyping in the embedding model can translate to LLMs if fine-tuned on gender-inclusive text. Secondly, our research is focused on gender-inclusive language in **English** and not directly applicable to other languages. The *NeuTralRewriter* [14] was specifically developed for English, and since the specific characteristics of gender-neutral terminology are language-dependent, applying the method to other languages would require the development of a language-specific version of the *Rewriter*. We leave this to future research. A third limitation of our research lies in the **erasure of word embeddings for *he*/*she* pronouns** due to the replacement with *they*. However, since we are presenting a proof-of-concept study and the measurement of gender stereotyping is not dependent on these pronouns, we accepted this. Future work could rewrite pronouns only in a percentage of cases or only in cases where masculine/feminine pronouns are used generically. Lastly, our research is limited by a narrow focus on **binary male and female genders** when assessing model bias. There is a significant gap in NLP research regarding the incorporation of non-binary gender identities in both measuring and mitigating bias [32]. Due to the nature of this proof-of-concept study, we adhered to commonly employed binary metrics. Future work will need to examine progress made regarding the integration of non-binary gender identities in embedding models through inclusive terminology.

5. Conclusion

This research explored the effects of gender-neutral language on gender stereotyping and latent gender information in classic embedding models. We found that training on text with gender-neutral singular pronouns and role nouns effected a reduction in stereotyping as measured by WEAT [20] and ECT [21]. These reductions do not surpass those that can be achieved by targeted, post-hoc debiasing [19]. However, gender-neutral training data showed an advantage when measuring latent gender information in embeddings through classification and clustering. This demonstrates a more comprehensive effect of gender-neutral language in the removal of unnecessarily gendered associations, which is in line with the aims of gender-inclusive language.

While future work will need to investigate whether our results hold at scale and can be transferred to LLMs, our exploratory findings suggest that adjusting training data to be more gender-inclusive can improve gender representations in pre-trained models toward a more equitable conceptualization of gender. This research presents a promising approach to the incorporation of principles of gender-inclusive language to ensure fairness and inclusivity in AI systems.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 12/RC/2289_P2. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

[2] https://fh295.github.io/simlex.html

References

[1] OpenAI, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774, arXiv:2303.08774 [cs].

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. URL: http://arxiv.org/abs/2302.13971. doi:10.48550/arXiv.2302.13971, arXiv:2302.13971 [cs].

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, 2013. URL: http://arxiv.org/abs/1301.3781, arXiv:1301.3781 [cs].

[4] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: http://aclweb.org/anthology/D14-1162. doi:10.3115/v1/D14-1162.

[5] J. Devlin, M.-W. Chang, L. Kenton, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[6] S. Arora, A. May, J. Zhang, C. Ré, Contextual Embeddings: When Are They Worth It?, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2650–2663. URL: https://aclanthology.org/2020.acl-main.236. doi:10.18653/v1/2020.acl-main.236.

[7] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and Fairness in Large Language Models: A Survey, Computational Linguistics (2024) 1–79. URL: https://doi.org/10.1162/coli_a_00524. doi:10.1162/coli_a_00524.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623. doi:10.1145/3442188.3445922, conference Proceedings.

[9] E. Kramer, Feminist Linguistics and Linguistic Feminisms, in: Ellen Lewin, Leni M. Silverstein (Eds.), Mapping Feminist Anthropology in the Twenty-First Century, Rutgers University Press, 2016, p. 65. URL: https://go.exlibris.link/J2p0HbgK.

[10] A. C. Saguy, J. A. Williams, A Little Word That Means A Lot: A Reassessment of Singular *They* in a New Era of Gender Politics, Gender & Society 36 (2022) 5–31. URL: http://journals.sagepub.com/doi/10.1177/08912432211057921. doi:10.1177/08912432211057921.

[11] R. Lakoff, Language and Woman's Place, Language in Society 2 (1973) 45–80. URL: http://www.jstor.org/stable/4166707, publisher: Cambridge University Press.

[12] A. Pauwels, Linguistic Sexism and Feminist Linguistic Activism, in: J. Holmes, M. Meyerhoff (Eds.), The Handbook of Language and Gender, Blackwell Publishing Ltd, Oxford, UK, 2003, pp. 550–570. doi:10.1002/9780470756942.ch24.

[13] S. F. Kiesling, Language, gender, and sexuality: an introduction, Book, Whole, 1;1st; ed., Routledge, London;New York;, 2019. doi:10.4324/9781351042420.

[14] E. Vanmassenhove, C. Emmery, D. Shterionov, NeuTral Rewriter: A Rule-Based and Neural Approach to Automatic Rewriting into Gender Neutral Alternatives, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 8940–8948. URL: https://aclanthology.org/2021.emnlp-main.704.

[15] C. Amrhein, F. Schottmann, R. Sennrich, S. Läubli, Exploiting Biased Models to De-bias Text: A Gender-Fair Rewriting Model, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4486–4506. URL: https://aclanthology.org/2023.acl-long.246. doi:10.18653/v1/2023.acl-long.246.

[16] A. Piergentili, D. Fucci, B. Savoldi, L. Bentivogli, M. Negri, Gender Neutralization for an Inclusive

Machine Translation: from Theoretical Foundations to Open Challenges, 2023. URL: http://arxiv.org/abs/2301.10075. doi:10.48550/arXiv.2301.10075, arXiv:2301.10075 [cs].

[17] H. Thakur, A. Jain, P. Vaddamanu, P. P. Liang, L.-P. Morency, Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 340–351. URL: https://aclanthology.org/2023.acl-short.30.

[18] Z. Fatemi, C. Xing, W. Liu, C. Xiong, Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1249–1262. URL: https://aclanthology.org/2023.acl-short.108.

[19] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in: Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

[20] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186. Publisher: American Association for the Advancement of Science.

[21] S. Dev, J. Phillips, Attenuating Bias in Word vectors, in: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 879–887. URL: https://proceedings.mlr.press/v89/dev19a.html, iSSN: 2640-3498.

[22] M. Bartl, S. Leavy, From 'Showgirls' to 'Performers': Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs, in: A. Faleńska, C. Basta, M. Costa-jussà, S. Goldfarb-Tarrant, D. Nozza (Eds.), Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 280–294. URL: https://aclanthology.org/2024.gebnlp-1.18.

[23] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, The Pile: An 800GB Dataset of Diverse Text for Language Modeling, 2020. URL: http://arxiv.org/abs/2101.00027. doi:10.48550/arXiv.2101.00027, arXiv:2101.00027 [cs].

[24] J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, A. Moffat, CC-News-En: A Large English News Corpus, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, Virtual Event Ireland, 2020, pp. 3077–3084. URL: https://dl.acm.org/doi/10.1145/3340531.3412762. doi:10.1145/3340531.3412762.

[25] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50.

[26] H. Gonen, Y. Goldberg, Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 609–614. URL: https://aclanthology.org/N19-1061. doi:10.18653/v1/N19-1061.

[27] S. Goldfarb-Tarrant, R. Marchant, R. Muñoz Sánchez, M. Pandya, A. Lopez, Intrinsic Bias Metrics Do Not Correlate with Application Bias, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 1926–1940. URL: https://aclanthology.org/2021.acl-long.150. doi:10.18653/v1/2021.acl-long.150.

[28] A. G. Greenwald, D. E. McGhee, J. L. Schwartz, Measuring individual differences in implicit cognition: the implicit association test, Journal of personality and social psychology 74 (1998) 1464. Publisher: American Psychological Association.

[29] A. Lauscher, G. Glavaš, S. P. Ponzetto, I. Vulić, A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 8131–8138. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6325. doi:10.1609/aaai.v34i05.6325, number: 05.
[30] F. Hill, R. Reichart, A. Korhonen, SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation, 2014. URL: http://arxiv.org/abs/1408.3456. doi:10.48550/arXiv.1408.3456, arXiv:1408.3456 [cs] version: 1.
[31] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: the concept revisited, ACM Transactions on Information Systems 20 (2002) 116–131. URL: https://doi.org/10.1145/503104.503110. doi:10.1145/503104.503110.
[32] H. Devinney, J. Björklund, H. Björklund, Theories of "Gender" in NLP Bias Research, in: ACM FAccT Conference 2022, Conference on Fairness, Accountability, and Transparency, Hybrid via Seoul, Soth Korea, June 21-14, 2022, 2022.
[33] T. Manzini, L. Yao Chong, A. W. Black, Y. Tsvetkov, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 615–621. URL: https://aclanthology.org/N19-1062. doi:10.18653/v1/N19-1062.

A. WEAT Target and Attribute Terms

	category	words
$\mathbf{W}A$	**male-dominated professions**	manager, executive, doctor, lawyer, programmer, scientist, soldier, supervisor, rancher, janitor, firefighter, officer
	female-dominated professions	secretary, nurse, clerk, artist, homemaker, dancer, singer, librarian, maid, hairdresser, stylist, receptionist, counselor
$\mathbf{W}B$	**computer science**	firmware, gui, programmer, hardware, notebook, database, router, pc
	childcare	children, babysitter, daycare, homemaker, newborn, baby, toddler, parenting

Table 4
Target words for additional WEAT-inspired testing. Professions were taken from Manzini et al. [33]

W7	**A1**	male, man, boy, brother, son, he, him, his
	A2	female, woman, girl, sister, daughter, she, her, hers
W8	**A1**	brother, father, uncle, grandfather, son, he, his, him
	A2	sister, mother, aunt, grandmother, daughter, she, her, hers
present study	**A1**	brother, father, uncle, grandfather, son, boy, man, male
	A2	sister, mother, aunt, grandmother, daughter, girl, woman, female

Table 5
Attribute words used in WEAT 7, WEAT 8 and in the present study. We replaced the original lists due to the fact that gender-specific pronouns were removed by the rewriting process and we decided to keep the original length of seven attribute words.