Bias Identification in Language Models is Biased

Fanny Ducel[1], Aurélie Névéol[2], Karën Fort[1]

[1]) Sorbonne Université/LORIA, France
[2]) Université Paris-Saclay, CNRS, LISN, France

The issue of biases in language models (LMs) has recently gained significant attention within the Natural Language Processing (NLP) community. Addressing this problem is crucial due to its potential sociological, ethical, and political implications, especially considering the widespread use of language models by the general public. There exists now a significant amount of literature aimed at evaluating and mitigating these biases. In this work, we perform a non-exhaustive survey of this literature and find that these efforts are not immune to biases themselves. We then outline a meta-data based analysis of these works to partly explain this phenomenon.

To conduct this study, we manually annotated 66 NLP research articles written in English between 2018 and 2023 that focus on bias evaluation or bias mitigation in language models. We will present in further details our methodology, and the overlap with the papers studied in [Blodgett et al., 2020]. Amongst them, 53 articles propose debiasing technique or metrics, while 13 articles are surveys and position papers.

As a starting point, we note some common limitations mentioned in multiple approaches and surveys we studied, namely: English is the target language, the perspective is US-centric, and the only type of studied bias is gender, and more precisely, binary gender.

There are 23 different languages targeted amongst the 53 experimental articles we studied. However, we find that 96% (51/53) of articles focus on English and that 83% (44/53), exclusively so (see Figure 1). We also note that 33% (17/51) of papers do not explicitly state that they work on English. As pointed out in [Ducel et al., 2022], it is important to do so, as English is not a "default language" and approaches proposed for it cannot be trivially adapted for other languages. However, we want to acknowledge the recent efforts that are being made to work on more diverse languages. Some of the articles (9/53) even propose multilingual solutions [Lauscher et al., 2021, Nozza et al., 2021, Arora et al., 2023].

Further, we claim that the perspective of a vast majority of these articles is US-centric. Our corpus contain 237 different authors that are based in 21 different countries. Nonetheless, similarly to the distribution of studied languages, we can see on Figure 2 that 56% of papers (37/66) contain at least one author affiliated to the United States. This rises to 72% (37+11 out of 66) when extrapolating country of residence from affiliated companies, in cases where countries are not specified. This can be problematic in that biases are cultural, so the biases that are taken into account by US authors are specific to their country.

Therefore, it is likely that a language model that has been supposedly "debiased" using an American interpretation of bias would contain lots of other biases, which we could neither detect, nor mitigate [Malik et al., 2022].

We also study the proportion of industry affiliations present in the papers. We find that 42% (28/66) of them have at least one author affiliated to a company (see Figure 3). In total, 13 companies are represented, and we find the most famous BigTech amongst them : Microsoft, Google, Facebook and Amazon. We will explain in more details the potential consequences of this, relying in particular on [Abdalla and Abdalla, 2020].

Finally, we turn our attention to the type of bias that is being studied. For this part, we exclude again the 13 survey and position papers. We find that 88% (47/53) of the articles focus on gender biases (see Figure 4), and 95% of them (45/47) on binary gender more specifically. Nonetheless, 58% (31/53) of the total of papers deal with several biases in parallel, and 11% (6/53) are intersectional, studying different types of biases simultaneously. Efforts towards intersectionality are very valuable as biases emerge from different sources, take different forms and individuals can suffer from different kinds of biases at once [Cao et al., 2022, Crenshaw, 1989].

To conclude, we would like to emphasize that our goal is not to belittle the efforts that have been produced, but to shed light on the biases inherent to the research. We would like to make simple recommendations, especially to encourage researchers to write a short paragraph stating where they write from, as it is usually done in social sciences.

We also want to highlight that the current research on biases in language models does not mirror the reality of biases. Biases are not universal, they depend heavily on language and culture. Moreover, gender is not the only source of bias, and it is well known that gender is not only binary, assuming that can even be harmful [Larson, 2017, Dev et al., 2021]. We advocate for more diverse resources to better profile these other types of biases in language models, so that we can try and tackle the notion of bias in all its complexity.

Figure 1: Distribution of studied languages among papers

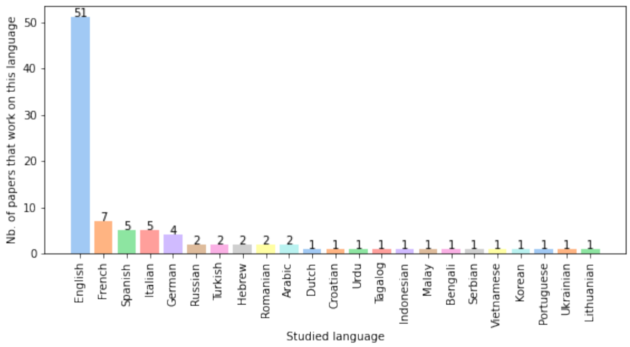


Figure 2: Distribution of authors' countries affiliations among papers

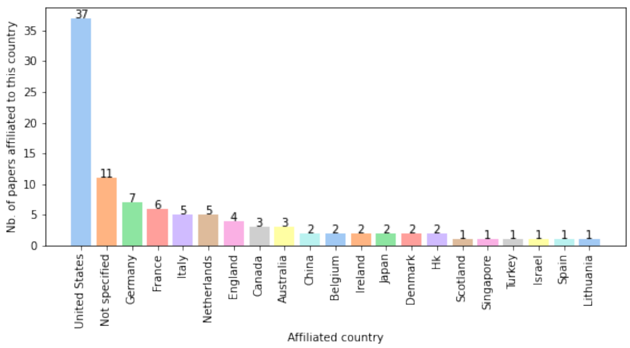


Figure 3: Distribution of authors' companies affiliations among papers

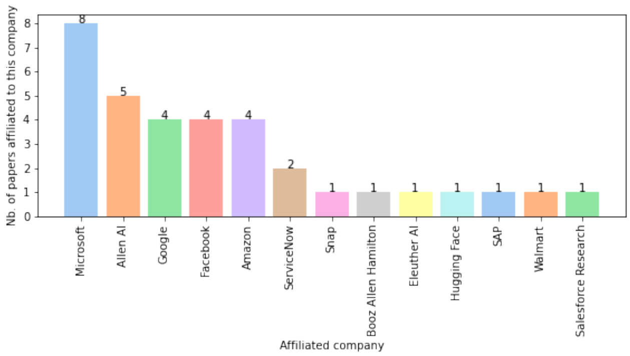
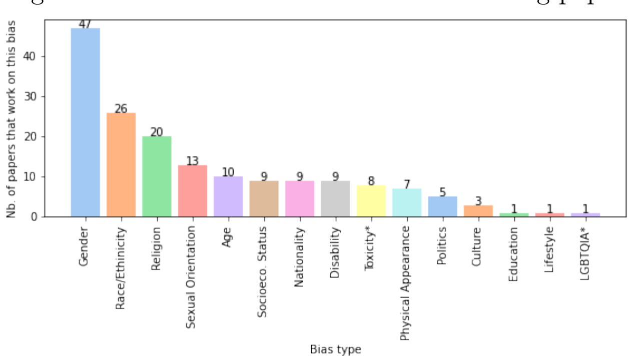


Figure 4: Distribution of studied biases among papers



References

[Abdalla and Abdalla, 2020] Abdalla, M. and Abdalla, M. (2020). The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. *CoRR*, abs/2009.13676.

[Akyürek et al., 2022] Akyürek, A. F., Paik, S., Kocyigit, M., Akbiyik, S., Runyun, S. L., and Wijaya, D. (2022). On Measuring Social Biases in Prompt-Based Multi-Task Learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 551–564, Seattle, United States. Association for Computational Linguistics.

[An et al., 2023] An, H., Li, Z., Zhao, J., and Rudinger, R. (2023). SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models. arXiv:2210.07269 [cs].

[Arora et al., 2023] Arora, A., Kaffee, L.-A., and Augenstein, I. (2023). Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. arXiv:2203.13722 [cs].

[Blodgett et al., 2020] Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

[Blodgett et al., 2021] Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

[Borchers et al., 2022] Borchers, C., Gala, D., Gilburt, B., Oravkin, E., Bounsi, W., Asano, Y. M., and Kirk, H. (2022). Looking for a Handsome Carpenter! Debiasing GPT-3 Job Advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. Association for Computational Linguistics.

[Bordia and Bowman, 2019] Bordia, S. and Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

[Cao et al., 2022] Cao, Y., Sotnikova, A., Daumé III, H., Rudinger, R., and Zou, L. (2022). Theory-Grounded Measurement of U.S. Social Stereotypes in English Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

[Cao and Daumé III, 2020] Cao, Y. T. and Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics, pages 4568–4595, Online. Association for Computational Linguistics.

[Choenni et al., 2021] Choenni, R., Shutova, E., and van Rooij, R. (2021). Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Chuang et al., 2023] Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., and Jegelka, S. (2023). Debiasing Vision-Language Models via Biased Prompts. arXiv:2302.00070 [cs].

[Crenshaw, 1989] Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167.

[de Vassimon Manela et al., 2021] de Vassimon Manela, D., Errington, D., Fisher, T., van Breugel, B., and Minervini, P. (2021). Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

[Delobelle and Berendt, 2022] Delobelle, P. and Berendt, B. (2022). FairDistillation: Mitigating Stereotyping in Language Models. arXiv:2207.04546 [cs].

[Delobelle et al., 2022] Delobelle, P., Tokpo, E., Calders, T., and Berendt, B. (2022). Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

[Dev et al., 2019] Dev, S., Li, T., Phillips, J., and Srikumar, V. (2019). On Measuring and Mitigating Biased Inferences of Word Embeddings. arXiv:1908.09369 [cs].

[Dev et al., 2021] Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J., and Chang, K.-W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Dev et al., 2022] Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., and Chang, K.-W. (2022). On Measures of Biases and Harms in NLP. arXiv:2108.03362 [cs].

[Dhamala et al., 2021] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). BOLD: Dataset and

Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872. arXiv:2101.11718 [cs].

[Dinan et al., 2020] Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., and Weston, J. (2020). Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

[Ducel et al., 2022] Ducel, F., Fort, K., Lejeune, G., and Lepage, Y. (2022). Do we name the languages we study? the #BenderRule in LREC and ACL articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, Marseille, France. European Language Resources Association.

[Gaci et al., 2022a] Gaci, Y., Benatallah, B., Casati, F., and Benabdeslem, K. (2022a). Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Gaci et al., 2022b] Gaci, Y., Benatallah, B., Casati, F., and Benabdeslem, K. (2022b). Masked Language Models as Stereotype Detectors? In *EDBT 2022*, Edinburgh, United Kingdom.

[Gehman et al., 2020] Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462 [cs].

[Hutchinson et al., 2020] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

[Jia et al., 2020] Jia, S., Meng, T., Zhao, J., and Chang, K.-W. (2020). Mitigating Gender Bias Amplification in Distribution by Posterior Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.

[Jin et al., 2021] Jin, X., Barbieri, F., Kennedy, B., Mostafazadeh Davani, A., Neves, L., and Ren, X. (2021). On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

[Johnson et al., 2022] Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., and Bertulfo, D. J. (2022). The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv:2203.07785 [cs].

[Kaneko and Bollegala, 2021] Kaneko, M. and Bollegala, D. (2021). Unmasking the Mask – Evaluating Social Biases in Masked Language Models. arXiv:2104.07496 [cs].

[Kaneko et al., 2022] Kaneko, M., Bollegala, D., and Okazaki, N. (2022). Debiasing isn't enough! – On the Effectiveness of Debiasing MLMs and their Social Biases in Downstream Tasks. arXiv:2210.02938 [cs].

[Kirk et al., 2021] Kirk, H., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., and Asano, Y. M. (2021). Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. arXiv:2102.04130 [cs].

[Kurita et al., 2019] Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

[Lalor et al., 2022] Lalor, J., Yang, Y., Smith, K., Forsgren, N., and Abbasi, A. (2022). Benchmarking Intersectional Biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, Seattle, United States. Association for Computational Linguistics.

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

[Lauscher et al., 2021] Lauscher, A., Lueken, T., and Glavaš, G. (2021). Sustainable Modular Debiasing of Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Li et al., 2020] Li, T., Khot, T., Khashabi, D., Sabharwal, A., and Srikumar, V. (2020). UnQovering Stereotyping Biases via Underspecified Questions. arXiv:2010.02428 [cs].

[Liang et al., 2021] Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards Understanding and Mitigating Social Biases in Language Models. arXiv:2106.13219 [cs].

[Liu et al., 2021] Liu, R., Jia, C., Wei, J., Xu, G., Wang, L., and Vosoughi, S. (2021). Mitigating Political Bias in Language Models Through Reinforced Calibration. arXiv:2104.14795 [cs].

[Lucy and Bamman, 2021] Lucy, L. and Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

[Ma et al., 2020] Ma, X., Sap, M., Rashkin, H., and Choi, Y. (2020). PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

[Malik et al., 2022] Malik, V., Dev, S., Nishi, A., Peng, N., and Chang, K.-W. (2022). Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

[Meade et al., 2022] Meade, N., Poole-Dayan, E., and Reddy, S. (2022). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

[Miceli et al., 2021] Miceli, M., Posada, J., and Yang, T. (2021). Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? arXiv:2109.08131 [cs].

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

[Nangia et al., 2020] Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

[Nozza et al., 2021] Nozza, D., Bianchi, F., and Hovy, D. (2021). HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

[Nozza et al., 2022a] Nozza, D., Bianchi, F., and Hovy, D. (2022a). Pipelines for Social Bias Testing of Large Language Models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

[Nozza et al., 2022b] Nozza, D., Bianchi, F., Lauscher, A., and Hovy, D. (2022b). Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

[Névéol et al., 2022] Névéol, A., Dupont, Y., Bezançon, J., and Fort, K. (2022). French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

[Ousidhoum et al., 2021] Ousidhoum, N., Zhao, X., Fang, T., Song, Y., and Yeung, D.-Y. (2021). Probing Toxic Content in Large Pre-Trained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

[Parrish et al., 2022] Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. (2022). BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

[Sap et al., 2020] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

[Schick et al., 2021] Schick, T., Udupa, S., and Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. arXiv:2103.00453 [cs].

[Schramowski et al., 2022] Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., and Kersting, K. (2022). Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do. arXiv:2103.11790 [cs].

[Selvam et al., 2022] Selvam, N. R., Dev, S., Khashabi, D., Khot, T., and Chang, K.-W. (2022). The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks. arXiv:2210.10040 [cs].

[Sheng et al., 2019] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3405–3410, Hong Kong, China. Association for Computational Linguistics.

[Sheng et al., 2020] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2020). Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

[Sheng et al., 2021] Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2021). Societal Biases in Language Generation: Progress and Challenges. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4275–4293, Online. Association for Computational Linguistics.

[Si et al., 2023] Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., and Wang, L. (2023). Prompting GPT-3 To Be Reliable. arXiv:2210.09150 [cs].

[Silva et al., 2021] Silva, A., Tambwekar, P., and Gombolay, M. (2021). Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

[Smith and Williams, 2021] Smith, E. M. and Williams, A. (2021). Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models. arXiv:2109.03300 [cs].

[Stanczak and Augenstein, 2021] Stanczak, K. and Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing. arXiv:2112.14168 [cs].

[Stanovsky et al., 2019] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

[Sun et al., 2022] Sun, T., He, J., Qiu, X., and Huang, X. (2022). BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Talat et al., 2022] Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., and Van Der Wal, O. (2022). You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

[Tan and Celis, 2019] Tan, Y. C. and Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. arXiv:1911.01485 [cs, stat].

[van der Wal et al., 2022a] van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., and Schulz, K. (2022a). Undesirable biases in NLP: Averting a crisis of measurement. arXiv:2211.13709 [cs].

[van der Wal et al., 2022b] van der Wal, O., Jumelet, J., Schulz, K., and Zuidema, W. (2022b). The Birth of Bias: A case study on the evolution of gender bias in an English language model. arXiv:2207.10245 [cs].

[Vig et al., 2020] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. M. (2020). Investigating gender bias in language models using causal mediation analysis. In *Neural Information Processing Systems*.

[Webster et al., 2018] Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

[Zmigrod et al., 2019] Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2019). Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.