Bias, Subjectivity and Norm in Large Language Models

Thierry Poibeau

Bias, Subjectivity and Norm in Large Language Models

Thierry Poibeau[1]

[1]Lattice Lab., École normale supérieure-PSL & CNRS 45 rue d'Ulm 75005 Paris, France

Abstract

This article reevaluates the concept of bias in Large Language Models, highlighting the inherent and varying nature of these biases and the complexities involved in post hoc adjustments to meet legal and ethical standards. It argues for shifting the focus from seeking bias-free models to enhancing transparency in filtering processes, tailored to specific use cases, acknowledging that biases reflect societal values.

Keywords

Large Language Models, Bias, Norm, Subjectivity

1. Introduction

Large Language models (LLMs), particularly generative models like GPT, have become prominent in natural language processing due to their effectiveness across various tasks and languages. Despite their known architecture [1], their internal operations remain largely opaque, raising questions about their ability to generalize linguistic phenomena and handle subjective information, such as differing opinions and cultural preferences. A significant challenge is the potential for these models to generate undesired content, including violent, misogynistic or racist remarks. To address this, various techniques are used to filter and mitigate problematic elements during both training and content generation, often relying on presumed reliable sources like Wikipedia and sanitized internet data.

The issue at hand has been extensively addressed in the literature through the concept of bias. Biases inherently entail negative elements that should be eliminated. A substantial segment of the NLP field is now dedicated to the process of debiasing models [2, 3]. The aim is to remove the biases until we obtain a neutral model, which would allow less discriminatory use. One of the main challenges of eliminating biases lies in defining and identifying them, necessitating a norm against which these biases can be measured, However, it appears improbable that we can establish a completely bias-free and objective world to which language models can conform once cleansed of spuriously learned associations. In this position paper, we defend the idea that the world contain subjective elements (opinions, tastes, preferences) that cannot be "objectified", with the consequence that the "norm" implicitly required to debias models does not exist. In other words, debiasing" in itself reflect a point of view and thus is not neutral.

The paper is structured as follows: Section 2 offers a brief review of the concept of bias.

AEQUITAS 2024: Workshop on Fairness and Bias in AI | co-located with ECAI 2024, Santiago de Compostela, Spain
thierry.poibeau@ens.psl.eu (T. Poibeau)
0000-0003-3669-4051 (T. Poibeau)

In Section 3, we address the absence of a universal "ground truth" for the task, leading to the conclusion that achieving completely unbiased models is unattainable. Finally, Section 4 presents some proposals for better addressing this subjectivity.

2. Previous Work

Blodgett et al. [4] highlights the importance of precisely defining the terms used when talking about notions such as *biases*. We revisit this very concept, delving into the inherent connection between biases and the data employed in training a model.

2.1. Language Models as a Mirror of Society

A cognitive bias refers to a consistent deviation from established norms or rational thinking in the process of judgment [5]. People form their own "subjective reality" based on how they perceive information. It is the individual's construction of reality, rather than the objective information itself, that can influence their actions in the world. Consequently, cognitive biases can result in distortions in perception, flawed judgment, illogical interpretations, and irrational behavior.

As early as 2016, in a seminal article, Bolukbasi et al. [6] set out the problem clearly: language models reflect the data on which they are trained, and therefore indirectly society. We could legitimately say that it is society that needs to be acted upon (which is not wrong in itself, but does not really answer the question), and developers must also take their share of responsibility (the quote refers to *word embeddings*, but it can be transposed to language models in general).

> One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to debias society rather than word embeddings. However, by reducing the bias in today's computer systems (or at least not amplifying the bias), which is increasingly reliant on word embeddings, in a small way debiased word embeddings can hopefully contribute to reducing gender bias in society. At the very least, machine learning should not be used to inadvertently amplify these biases, as we have seen can naturally happen. [6]

The quotation from Bolukbasi et al. [6] thus highlights the link between these models and the society they reflect. Language plays a central role in this: Blodgett et al. [4] describes language itself as a means of maintaining and/or reinforcing social hierarchies.

2.2. Mitigating and/or Removing Bias

Many studies have highlighted the presence of bias in language models [7, 8, 9, 10, 11], among others. The means of attenuating and/or eliminating these biases has therefore logically become a major research theme and a large number of techniques have been proposed [12, 13, 9, 14, 15, 16]. This inventory is merely illustrative and necessarily very partial, given the increase in publications on this subject in recent years.

However, these studies remain partial (most focus on gender bias, others on race or religion, but the different aspects are rarely treated together). Furthermore, as noted by Meade et al. [3],

the effectiveness of techniques and their impact on processing algorithms is also often left out of the equation. Finally, aiming at eliminating bias implies being able to recognise it. But the notion of bias is complex, and implies a deviation from a norm, as we saw in the previous section. We do not question the need to propose methods to reduce bias in models, but eliminating it implies being able to achieve an objective description of reality, a notion discussed by Waseem et al. [17]. These authors challenge the "solutionism" of the algorithmic approaches proposed: while algorithms are useful, they also suffer from their own subjectivity and are not a universal solution.

2.3. Practical and Regulatory Considerations

To complete our observations and take into account societal concerns, we decided to complete our study by interviewing representative experts. These individuals include engineers deeply entrenched in the development of language models and their applications, as well as members of think tanks and regulatory bodies.

These professionals were very careful with the notion of bias. They all consider that biases are entrenched in society and thus an integral part of corpora. This does not mean that there is nothing to be done. They advocate for the regulation of practical applications of LLMs, rather than trying to regulate the models themselves. This approach seeks to ensure that specific measures are meticulously tailored to each unique application. The overarching criterion here is the prevention of discrimination against individuals who are the target of these applications. At the very least, these models should refrain from exacerbating existing biases ingrained within society. This application-centric strategy (which is also the one of the AI Act recently voted in Europe [18]), while effective, inherently introduces complexity when it comes to formulating general principles. Simultaneously, there exists a unanimous consensus among these professionals regarding the paramount importance of preserving a substantial degree of freedom for theoretical AI research (especially for the development of LLMs, that remains models and not appliocations; a difference should be made for example between GPT models, and applications like ChatGPT).

These observations extend to more nuanced considerations concerning regulation. A hot topic in the domain is for example the idea of model certification (which refers to the idea that models would have to be assessed and certified prior any commercial use). While this idea may initially seem enticing, it is crucial to acknowledge other potential issues. The cost associated with certification, as articulated by these professionals, could cast a substantial shadow over smaller enterprises, potentially leading to distortions in commercial competition. However, these intricacies, while significant, venture beyond the immediate scope of this study.

3. The Lack of a Universal "Ground truth", or the End of Objectivity

As we have observed, the concept of bias inherently presupposes the existence of a norm, yet this norm is profoundly relative, contingent upon culture, ideology, or even the perspective of an individual.

3.1. The Relativity of Universal Notions

Contrary to the inherent subjectivity we have described, the prevailing literature in the field often operates on the presumption of objectivity, where representations within language models are expected to be rendered devoid of subjectivity, including bias, point of view, or opinion. However, as Waseem et al. [17] have pointed out, this assumption is indeed an illusion. For instance, the interpretation of principles such as freedom of expression can diverge significantly between a European and American vantage point. Certain associations, like those between gender and profession, which may seem to some as clear-cut biases warranting elimination, are experienced differently across countries, cultures, and political beliefs. Natural Language Processing (NLP) predominantly reflects the values and perspectives of Western culture, which, by its very nature, lacks universality.

The challenge of mitigating bias within such a context is arduous, as we operate within a relatively subjective domain where there is no universal "ground truth." While test and training datasets can be curated to exclude blatantly problematic elements, there exist more intricate cases that elude straightforward binary classifications of bias or no bias. Furthermore, the approach to addressing bias is also a crucial consideration, as Bolukbasi et al. [6] reveal that suppressing, rather than simply mitigating, biases can yield unintended consequences. Waseem et al. [17] emphasize that de-biasing methods can only rectify a fraction of the biases present.

Barocas et al. [19] provide valuable insights into this intricate conundrum, suggesting the need to differentiate between tasks with accessible ground truth and those without. In this particular scenario, we find ourselves in the latter category, where the concept of bias lacks a consensus definition and universal understanding. Barocas presages that technologies developed within such an uncertain framework may have limited success, owing to the inherent ambiguity surrounding the ideal outcome. Even seemingly neutral concepts like "good quality" text for learning, used in various publications, exhibit inherent biases favoring the language and perspectives of well-educated and affluent classes, as demonstrated by Gururangan et al. [20].

3.2. Bias and Freedom of Expression.

Finally, it is important to highlight the complexity of these issues, which relate to the broader context of freedom of expression and its various social impacts. This dilemma is clear, especially with the challenges posed by social networks. Regulating these platforms is necessary to prevent abuses like hate speech, harassment, and defamation. However, too much regulation could also threaten freedom of expression.

The question of who sets the rules for these networks is another layer of complexity. When networks establish their own guidelines, it raises concerns because it means that private entities are effectively determining the boundaries of freedom of expression. On the contrary, if the State intervenes, suspicions often emerge that it might be infringing upon free expression, indirectly fueling conspiracy theories.

Navigating through this intricate landscape requires treading a delicate path, as none of the available options is entirely satisfactory. The key lies in discerning the least detrimental course of action depending on the context, while maintaining a commitment to maximum transparency and swift responsiveness in the face of problems. Striking a balance between safeguarding

freedom of expression and addressing the genuine concerns of abuse is an ongoing challenge that calls for careful consideration and adaptability.

4. Some Proposals

Recognizing the lack of universal norm and the subjectivity of the notion of bias does not mean that nothing should be done. In this section, we examine different proposals to better address the problem.

Develop Application-Dependent Bias Typologies. As already seen, LLMs are, by design, reflections of the subjectivity present in the massive text corpora they were trained on. Thus, it is unsurprising that biases manifest within them.

The real challenge arises when these models are employed in practical applications, generating texts that can be riddled with stereotypes, discrimination, or even outright racism. As per Abid et al. [21]'s findings in the study on persistent bias in LLMs, bias can be mitigated to some extent by providing the model with more extensive context, often achieved by utilizing longer prompts. For example, introducing a "positive" primer to the model, such as querying it with prompts like "*Muslims are hard-working. Two Muslims walked into a*" can significantly alter the content generated by the model.

However, it is important to note that in this scenario, the model's output can be swayed in either direction – toward mitigating biases or exacerbating them. Therefore, a more nuanced understanding of biases would be useful. This would entail for model and service providers developing a comprehensive typology of biases, allowing us to distinguish between those that infringe upon legal and ethical boundaries, which necessitate complete elimination, and those that merely reflect opinion or subjectivity, which may be addressed through attenuation.

Use the notion of "class of applications". The majority of articles addressing bias removal tend to lack contextualization. These articles primarily offer technical solutions to modify model weights, aimed at removing or mitigating bias, irrespective of the specific usage context. However, it is imperative to implement varying filtering strategies depending on whether we are dealing with a fully automated filtering algorithm, intended to eliminate problematic content without human intervention, a professional writing tool, or a public dialogue system.

In the realm of AI regulation, the European AI Act [18] is poised to categorize applications into different classes, each requiring distinct levels of filtering and precaution measures based on their perceived level of risk or danger. While the criticality of the targeted application might not be the sole pivotal factor for language models, the concept of application classes and context-aware filtering should be kept in consideration.

For instance, as demonstrated by Blodgett et al. [4] in their 2020 meta-study on bias, most articles in the field often lack well-defined motivations for their objectives. They tend to overlook essential sociological data when attempting to debias a model for a specific application. Blodgett and her colleagues put forth several recommendations to ameliorate this situation, offering potential solutions to address this issue, which remains pertinent in the current landscape.

Better Document models. One crucial step in addressing these intricate challenges is to comprehensively document the algorithms, datasets, and filtering strategies employed. This

recommendation aligns with well-established principles, as outlined by Bender and Friedman [22]. They offer practical suggestions, such as the creation of "data statements", which elucidate the constraints of a given system and provide insights into the training sets utilized.

It is worth noting that developers may not always have access to the most optimal or genuinely representative data. In such instances, it becomes imperative to meticulously document these limitations to foster transparency and awareness. Moreover, it is essential that all filtering strategies used are methodically disclosed and made accessible to the public, irrespective of whether they are applied within private or commercial systems. These strategies constitute fundamental decisions with far-reaching implications.

Furthermore, it is pertinent to recognize that while filtering and debiasing techniques are typically applied for benevolent purposes, they can potentially be inverted and exploited to introduce ideology or a specific point of view into a model. This underscores the need for vigilance and ethical considerations when implementing such techniques.

5. Conclusion

The article reevaluates the concept of bias in language models, acknowledging that biases are inherent and vary in degree. Traditional methods to address these biases involve post hoc adjustments, which are necessary to comply with legal and ethical standards. However, the article emphasizes the complexity of such interventions, raising questions about the criteria for text selection and bias correction, which often reflect Western values and may not fully account for the cultural, social, and contextual nuances present in other regions or communities. It argues that biases are inevitable and that language models, rooted in statistical principles, mirror societal biases. The focus should shift from seeking bias-free models to improving transparency in the filtering processes, tailored to specific use cases. Fully open-source LLMs may help achieve this goal by allowing greater scrutiny and customization of the underlying algorithms, fostering more adaptable and transparent solutions.

6. Limitations

A limitation of this paper is that our state-of-the-art in removing bias from large language models (LLMs) is not comprehensive, as evidenced by the numerous publications emerging each month on this topic. Moreover, some experiments aim at reducing bias in specific applications rather than at the model level itself, as proposed in our article.

The sample size of experts interviewed for our study, as outlined in section 2.3, is currently limited. Expanding this pool of experts is essential to achieve a more comprehensive and dependable understanding of the landscape of Large Language Models.

Lastly, the recommendations presented in section 4 require practical implementation, thorough assessment, and rigorous evaluation to gauge their effectiveness and impact. Monitoring tendencies by analyzing how LLMs perform in mitigating bias and providing acceptable solutions will also be necessary to observe progress and advancements in the domain.

Ethical Aspects

While this study is centered around the examination of bias in Large Language Models (LLMs) and the methods to alleviate or eliminate them, no significant ethical concerns have arisen in the course of this research. The limited survey described in section 2.3 engaged human participants; however, it is important to emphasize that no personal information is either utilized or disclosed in this paper. The data employed here is exclusively impersonal and aggregated.

Acknowledgements

This work was supported in part by the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). This work was also funded by ASTOUND project (101071191 — HORIZON-EIC-2021- PATHFINDERCHALLENGES-01) of the European Commission.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proc of the Thirty-first Conference on Advances in Neural Information Processing Systems, Long Beach, USA, 2017, pp. 5998–6008.

[2] K. Stanczak, I. Augenstein, A Survey on Gender Bias in Natural Language Processing, arXiv, 2021. URL: http://arxiv.org/abs/2112.14168, arXiv:2112.14168 [cs].

[3] N. Meade, E. Poole-Dayan, S. Reddy, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1878–1898. URL: https://aclanthology.org/2022.acl-long.132. doi:10.18653/v1/2022.acl-long.132.

[4] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5454–5476. URL: https://www.aclweb.org/anthology/2020.acl-main.485. doi:10.18653/v1/2020.acl-main.485.

[5] M. G. Haselton, D. Nettle, D. R. Murray, The Evolution of Cognitive Bias, John Wiley & Sons, Ltd, 2015, pp. 1–20. doi:https://doi.org/10.1002/9781119125563.evpsych241.

[6] T. Bolukbasi, K. 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 29: Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 2016, pp. 4349–4357.

[7] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis,

Minnesota, 2019, pp. 622–628. URL: https://aclanthology.org/N19-1063. doi:10.18653/v1/N19-1063.

[8] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, in: Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Association for Computational Linguistics, Florence, Italy, 2019, pp. 166–172. URL: https://aclanthology.org/W19-3823. doi:10.18653/v1/W19-3823.

[9] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, S. Petrov, Measuring and Reducing Gendered Correlations in Pre-trained Models, arXiv, 2020. URL: https://arxiv.org/abs/2010.06032. doi:10.48550/ARXIV.2010.06032.

[10] N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman, 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), Association for Computational Linguistics, Online, 2020, pp. 1953–1967. URL: https://aclanthology.org/2020.emnlp-main.154. doi:10.18653/v1/2020.emnlp-main.154.

[11] M. Nadeem, A. Bethke, S. Reddy, 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), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416. doi:10.18653/v1/2021.acl-long.416.

[12] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, L.-P. Morency, Towards debiasing sentence representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5502–5515. URL: https://aclanthology.org/2020.acl-main.488. doi:10.18653/v1/2020.acl-main.488.

[13] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg, Null it out: Guarding protected attributes by iterative nullspace projection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7237–7256. URL: https://aclanthology.org/2020.acl-main.647. doi:10.18653/v1/2020.acl-main.647.

[14] M. Kaneko, D. Bollegala, Debiasing pre-trained contextualised embeddings, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1256–1266. URL: https://aclanthology.org/2021.eacl-main.107. doi:10.18653/v1/2021.eacl-main.107.

[15] T. Schick, S. Udupa, H. Schütze, Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP, Transactions of the Association for Computational Linguistics 9 (2021) 1408–1424. doi:10.1162/tacl_a_00434.

[16] A. Lauscher, T. Lueken, G. Glavaš, Sustainable modular debiasing of language models, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4782–4797. URL: https://aclanthology.org/2021.findings-emnlp.411. doi:10.18653/v1/2021.findings-emnlp.411.

[17] Z. Waseem, S. Lulz, J. Bingel, I. Augenstein, Disembodied Machine Learning: On the

Illusion of Objectivity in NLP, arXiv, 2021. URL: https://arxiv.org/abs/2101.11974. doi:10.48550/ARXIV.2101.11974.

[18] European Commission, The Artificial Intelligence Act, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024.

[19] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook.org, 2019. http://www.fairmlbook.org.

[20] S. Gururangan, D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, N. A. Smith, Whose language counts as high quality? measuring language ideologies in text data selection, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2562–2580. URL: https://aclanthology.org/2022.emnlp-main.165.

[21] A. Abid, M. Farooqi, J. Zou, Persistent Anti-Muslim Bias in Large Language Models, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, ACM, Online, 2021. doi:10.1145/3461702.

[22] E. M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Transactions of the Association for Computational Linguistics 6 (2018) 587–604. URL: https://aclanthology.org/Q18-1041. doi:10.1162/tacl_a_00041.

[23] A. B. Powell, F. Ustek-Spilda, S. Lehuedé, I. Shklovski, Addressing ethical gaps in 'technology for good': Foregrounding care and capabilities, Big Data & Society 9 (2022). URL: https://doi.org/10.1177/20539517221113774.

[24] T. Schick, S. Udupa, H. Schütze, Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021. URL: https://arxiv.org/abs/2103.00453. doi:10.48550/ARXIV.2103.00453.

[25] X. Han, A. Shen, T. Cohn, T. Baldwin, L. Frermann, Systematic evaluation of predictive fairness, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 68–81. URL: https://aclanthology.org/2022.aacl-main.6.

[26] T. Korbak, K. Shi, A. Chen, R. Bhalerao, C. L. Buckley, J. Phang, S. R. Bowman, E. Perez, Pretraining Language Models with Human Preferences, arXiv, 2023. URL: https://arxiv.org/abs/2302.08582. doi:10.48550/ARXIV.2302.08582.

[27] S. C. Y. Chan, I. Dasgupta, J. Kim, D. Kumaran, A. K. Lampinen, F. Hill, Transformers generalize differently from information stored in context vs in weights, 2022. URL: https://arxiv.org/abs/2210.05675. doi:10.48550/ARXIV.2210.05675.

[28] E. Sheng, K.-W. Chang, P. Natarajan, N. Peng, 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), Association for Computational Linguistics, Online, 2021, pp. 4275–4293. URL: https://aclanthology.org/2021.acl-long.330. doi:10.18653/v1/2021.acl-long.330.

[29] S. Dev, E. Sheng, J. Zhao, A. Amstutz, J. Sun, Y. Hou, M. Sanseverino, J. Kim, A. Nishi, N. Peng, K.-W. Chang, On Measures of Biases and Harms in NLP, 2022. URL: http://arxiv.org/abs/2108.03362, arXiv:2108.03362 [cs].

[30] D. Chavalarias, Toxic Data: Comment les réseaux manipulent nos opinions, Flammarion, Paris, 2022.

[31] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, P. Kohli, Reducing sentiment bias in language models via counterfactual evaluation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 65–83. URL: https://aclanthology.org/2020.findings-emnlp.7. doi:10.18653/v1/2020.findings-emnlp.7.

[32] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, P. Kohli, Reducing Sentiment Bias in Language Models via Counterfactual Evaluation, 2020. URL: http://arxiv.org/abs/1911.03064. doi:10.48550/arXiv.1911.03064, arXiv:1911.03064 [cs].

[33] S. Sczesny, M. Formanowicz, F. Moser, Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?, Frontiers in Psychology 7 (2016). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00025. doi:https://doi.org/10.3389/fpsyg.2016.00025.

[34] N. Meade, E. Poole-Dayan, S. Reddy, An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models, 2022. URL: http://arxiv.org/abs/2110.08527, arXiv:2110.08527 [cs].

[35] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, 2022. URL: http://arxiv.org/abs/2211.05100, arXiv:2211.05100 [cs].

[36] A. Simoulin, B. Crabbé, Un modèle Transformer Génératif Pré-entrainé pour le _____ français, in: P. Denis, N. Grabar, A. Fraisse, R. Cardon, B. Jacquemin, E. Kergosien, A. Balvet (Eds.), Traitement Automatique des Langues Naturelles, ATALA, Lille, France, 2021, pp. 246–255. URL: https://hal.archives-ouvertes.fr/hal-03265900.

[37] H. Laurençon, L. Saulnier, T. Wang, C. Akiki, The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, 2022.

[38] J. Weizenbaum, Eliza —— a computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1966) 36–45. URL: https://doi.org/10.1145/365153.365168. doi:10.1145/365153.365168.

[39] J. F. Le Ny, Article "Biais", in: H. Bloch (Ed.), Grand dictionnaire de la psychologie, Larousse, Paris, 1991.

[40] S. Dev, E. Sheng, J. Zhao, A. Amstutz, J. Sun, Y. Hou, M. Sanseverino, J. Kim, A. Nishi, N. Peng, K.-W. Chang, On measures of biases and harms in NLP, in: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Association for Computational Linguistics, Online only, 2022, pp. 246–267. URL: https://aclanthology.org/2022.findings-aacl.24.

[41] K. Crawford, The trouble with bias, 2017. NeurIPS Keynote, https://www.youtube.com/watch?v=ggzWIipKraM.