

# Mitigating Biases in Deep Learning Models: A Path Towards Fairness and Inclusivity

Ismael Garrido-Muñoz<sup>1</sup>

<sup>1</sup>Universidad de Jaén, Campus Las Lagunillas s/n, 23071 Jaén, España

## Abstract

The emergence of large language models (LLMs) has revolutionized the field of natural language processing, facilitating remarkable progress across various domains. However, the inherent opaqueness of these models, functioning as black boxes, presents significant challenges. The lack of transparency obstructs our comprehension of their internal mechanisms and decision-making processes, raising concerns about their reliability and fairness. Various forms of biases have already been identified within these models. It is crucial to identify the location and encoding of these biases within LLMs to enable necessary modifications that ensure their safe and equitable application free of social biases in all kind of areas. Given the extensive deployment of LLMs in real-world applications, their impact on individuals' lives is magnified. Thus, the subsequent phase of this thesis will focus on effectively mitigating biases in deep learning models.

## Keywords

bias, deep learning, nlp, fairness, mitigation

## 1. Introduction

The advent of GPT-3[1] has sparked a massive adoption of this model, with predictions of its profound impact on the labor market, as outlined by [2]. This remarkable influence stems from the diverse range of capabilities that these models possess, including question answering, text generation, translation, summarization, information retrieval, act as a conversational agent, programming assistance, educational support, story telling and more.

However, despite the tremendous utility of LLMs, they also pose an emerging challenge: their tendency to operate as black boxes. While they exhibit impressive performance, their internal mechanisms and decision-making processes often remain opaque, making them difficult to comprehend and explain. This lack of transparency gives rise to concerns regarding their trustworthiness, fairness, and the potential biases that may be embedded within their models.

The concept of a black box refers to a system or model where the inputs and outputs are known, but the inner mechanisms and algorithms that generate those outputs remain concealed or poorly understood. LLMs, with their complex neural networks and millions, or even billions, of parameters, are intricate black boxes that often surpass human comprehension. This opaqueness hampers our ability to fully grasp the decision-making processes of these models,

---

*Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA, 28 September 2023, Jaén, Spain.*


✉ [igmunoz@ujaen.es](mailto:igmunoz@ujaen.es) (I. Garrido-Muñoz)

🌐 <https://ismael.codes/> (I. Garrido-Muñoz)

🆔 0000-0001-6656-9679 (I. Garrido-Muñoz)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

making it difficult to tackle biases, recognize potential vulnerabilities, and guarantee ethical and responsible utilization. Consequently, there is a pressing need to enhance transparency and develop techniques that shed light on the inner workings of LLMs.

In recent years, artificial intelligence has made significant advances, and a substantial portion of this progress can be attributed to neural network models. These models, trained on extensive datasets, have showcased remarkable capabilities in capturing various aspects of reality. However, while their ability to capture reality with precision is commendable, it can also have negative implications. One such concern arises from their propensity to inadvertently perpetuate and replicate undesirable stereotypes.

These models are already being used in multiple production systems such as medical systems[3], legal systems[4], hiring[5], content moderation[6], CRM[7], marketing[8], virtual assistants, harmful content detection[9], chatbots, etc.

These systems are used in products despite having proven to be unsafe. It is well known that sometimes these black boxes cause unintended harm. One example is the police COMPAS system, which assigned an unreal recidivism value to both white and black people. For white individuals, the assigned value was lower than the actual value, while for black individuals, it was higher than the actual value[10]. Another example is the medical system called Optum[11], which systematically allocated fewer resources for the treatment of black patients compared to white patients with the same level of need.

This realization raises concerns about the fairness and potential harm that may arise from the application of non-explainable models in certain situations. For instance, Amazon discontinued the use of a recruitment tool [12] when it was discovered to be biased against women. These examples highlight the presence of biases not only in language models but also in systems employing computer vision [13], audio processing [14], and linguistic corpora [15], [16]. It is crucial to address these biases as they can perpetuate inequality and have real-world consequences. Understanding and mitigating biases in such systems is a pressing concern.

In the case of GPT-3[1] (or its frontends like Chat-GPT or Bing GPT) or Google's alternative, Bard[17], studying these models is not feasible because they are provided as services through APIs or web interfaces. However, there have been releases of models with similar numbers of parameters and capabilities as the aforementioned ones. For instance, models like Llama[18], Vicuna[19], Bloom[20], OPT[21], XGLM[22], and the recent Falcon[23] do provide access to the trained models weights. This access enables us to review, correct, or mitigate any biases present in them.

This will be the next step of the thesis. Next, we provide a brief overview of evaluation techniques, followed by a collection of the most relevant techniques for bias mitigation. In previous works, such as the one mentioned in , a broader summary of the state-of-the-art in studying bias in language models can be found.

This will be the next step of the thesis. In the following section, we provide a brief overview of evaluation techniques, followed by a collection of the most relevant techniques for bias mitigation. In previous works, a broader summary of the state-of-the-art in studying bias in language models can be found [24].

## 2. Bias in NLP with deep learning

When we talk about bias in language models, we can approach it as a representational problem[25]. This refers to the bias that certain demographic groups face in terms of misrepresentation, including negative associations or even their absence in the data and consequently in the model. On the other hand, we can approach it as an allocation problem, which refers to issues of opportunities or resource distribution for individuals belonging to specific demographic groups.

### 2.1. Bias evaluation

There is extensive work when it comes to evaluating language models for bias, starting with the work of Bolukbasi et al. [26] on simple word embeddings. Later studies approached the bias issue from the perspective of coreference resolution, such as [27] with GloVe embeddings. Bias is also examined by measuring the association between concepts and protected attributes. Caliskan et al. [28] created the Word Embedding Association Test (WEAT) for this purpose. This test was extended by Dev et al. [29] and Manzini et al. [30]. Also it was extended by Lauscher et al. [31] by adding more protected attributes and applying it to languages other than English. It was later on adapted to more complex models like BERT, under the name SEAT, by May et al. [32] and Tan and Celis [33].

There are other approaches for more complex models like BERT or GPT-2. Vig [34] introduced visualization tools to understand where these models capture unwanted biases by examining their attention. Additionally, adaptations of WEAT, such as SEAT, have been developed. SEAT tests the protected attribute against a sentence instead of a word, specifically designed for contextual models like BERT. This work was further extended to consider the full context instead of just the sentence level. The latest evaluation method is applied to models like GPT-2, BERT, ELMo, among others.

More complex models also make serious errors. A compendium of errors discovered in ChatGPT is presented in the work of Borji [35]. The paper explains that this model is unable to successfully complete tasks that require spatial, temporal, or physical reasoning unless it has been specifically trained for those tasks.

### 2.2. Bias correction

The main approaches to address bias in language models consist of the following: fine-tuning the model [36], data augmentation to balance categories and avoid distortions towards one category[27], protecting the attribute during model training to prevent bias capture[37], or correcting the vector space of the model as presented in the works of Manzini et al. [30], Zhou et al. [38], Dev and Phillips [37]. Among these techniques, fine tuning and model editing are considered the most realistic, especially in the case of large-scale models, since retraining a model from scratch would be very costly in terms of time, hardware resources, money and the effort required to perform the pre-processing and tuning of training data.

One of the most promising techniques for model editing involves identifying how the model encodes certain knowledge and then making edits accordingly. The proposal of Meng et al. [39] focuses on editing factual knowledge and serves as a foundation for further adaptation.

This technique first identifies the model’s influential parts that contribute the most weight in choosing the last token by using causal mediation analysis. From there, the model’s weights are edited to guide it towards the desired token. For example, if the model answers **Obama** to the question “What is the surname of the U.S. president?”, the weights can be located and corrected to select the desired token **Biden** since this would be the updated and accurate answer. Similarly, this method can be generalized to make broader corrections. In fact, in a subsequent work[40] they adapt this method to perform mass corrections across the model weights. Then they evaluate whether the model only edits knowledge for the specific context given in the prompt or if it can generalize by asking questions about the same fact using different questions or contexts.

These techniques hold great potential in tackling bias, enhancing the accuracy, and bolstering the reliability of language models. By facilitating targeted edits that align with desired outcomes, these approaches enable the mitigation of unwanted biases in the models’ responses. As a result, they contribute to an improved understanding of fairness and ensure more reliable and unbiased outputs from language models.

### 3. Relevance of the problem

Every day, these enormous models are increasingly integrated into various products and production systems. However, this integration comes with its own set of challenges. From an economic standpoint, utilizing a biased system can lead to significant disadvantages, as it may not function effectively for all users. On the other hand, the impact of these models on people’s lives cannot be overlooked. There are specific contexts, such as systems for resource distribution, employment, or bank credit, where it is crucial to avoid using models that may contain any form of bias. Therefore, it is imperative to thoroughly study bias in data models and understand its underlying causes. This knowledge will enable us to either avoid deploying biased models altogether or develop strategies to mitigate harmful biases when they arise.

Furthermore, when a language model is identified as not performing adequately in a production system, such as a commercial product, companies face important decisions. Given the immense size and cost associated with training these models, some proposed solutions may be difficult to justify from an economic perspective. For instance, training the model from scratch with revised, filtered, or corrected training data would entail significant expenses. Another option, albeit costly, could involve discontinuing the use of the model, as a poorly performing model is unsuitable for deployment in production systems. This proposition gains some relevance considering the potential non-compliance of such models with new European AI regulations[41]. Alternatively, more practical approaches could involve retraining the model or leveraging state-of-the-art bias mitigation techniques to address the identified issues.

The choice of approach will depend on various factors, including the severity of the bias, the feasibility of retraining or mitigating the model, and the legal and ethical obligations that must be met. Regardless of the chosen course of action, it is essential to proactively address and rectify bias issues to ensure responsible and fair deployment of language models in real-world applications. By doing so, we can foster inclusivity, promote equitable outcomes, and uphold the principles of fairness and ethical AI.

## 4. Hypotheses and objectives

The following hypothesis is assumed: Given a language model based on deep learning, it will be possible to discern whether it contains biases, and characterize, measure, and mitigate them.

The following objectives are established:

- Conduct an intensive study of the state of the art regarding detection, evaluation, or mitigation of biases in deep learning models.
- Analyze and characterize biases present in existing models.
- Development of techniques and algorithms for unsupervised or semi-supervised detection and characterization of bias in existing models.
- Development of techniques and algorithms for the mitigation or correction of bias in existing models.

At this phase of the thesis, our primary focus is on the last point.

## 5. Methodology and the proposed experiments

As we move forward with the use of large language models, our next step will involve adapting and evaluating the previous work[42] in the context of LLMs. Specifically, the previous study shed light on how models tend to perceive women based on their physical appearance, while men are assessed primarily based on their behavior. This pattern was observed across the majority of the models investigated.

To proceed, we will replicate the aforementioned experiment using large language models (LLMs) and analyze to what extent increasing the model size affects bias, whether it exacerbates or reduces it. Once this evaluation is completed, our focus will shift towards bias mitigation strategies.

To mitigate bias, we will construct a corpus of prompts that elicit biased responses from the models. This corpus will serve as a foundation for our work in two main areas. First, we will develop methods to detect and identify biased terms produced by the model in its responses. Second, we will explore the previously discussed fact editing techniques to edit the behavior of the model for the detected biases in order to reduce or eliminate them. This will require adapting the causal mediation analysis mechanism to our problem, since editing a specific fact is not the same as making an edit that causes a trade-off between different classes of a protected attribute. After the editing process, we will evaluate the performance of the model with the same set of prompts to check the effectiveness of the mitigation method.

By undertaking these steps, we aim to gain insights into the behavior of large language models regarding bias and work towards developing effective strategies for bias mitigation.

## References

- [1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin,

- S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, *CoRR abs/2005.14165* (2020). URL: <https://arxiv.org/abs/2005.14165>. arXiv:2005.14165.
- [2] T. Eloundou, S. Manning, P. Mishkin, D. Rock, Gpts are gpts: An early look at the labor market impact potential of large language models, *ArXiv abs/2303.10130* (2023).
- [3] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, A. D. Shah, K. Morley, D. Osborn, J. Hayes, R. Stewart, J. Downs, W. Chapman, R. Dutta, Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances, *J Biomed Inform* 88 (2018) 11–19.
- [4] R. DALE, Law and word order: Nlp in legal tech, *Natural Language Engineering* 25 (2019) 211–217. doi:10.1017/S1351324918000475.
- [5] M. Bogen, A. Rieke, Help wanted: an examination of hiring algorithms, equity, and bias, 2018.
- [6] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, 2018. doi:10.12987/9780300235029.
- [7] Salesforce, *Salesforce Announces AI Cloud – Bringing Trusted Generative AI to the Enterprise* – [investor.salesforce.com](https://investor.salesforce.com), <https://investor.salesforce.com/press-releases/press-release-details/2023/Salesforce-Announces-AI-Cloud--Bringing-Trusted-Generative-AI-to-the-Enterprise/default.aspx>, 2023. [Accessed 18-Jun-2023].
- [8] Adobe, *Adobe Announces New Sensei GenAI Services to Reimagine End-to-End Marketing Workflows* – [news.adobe.com](https://news.adobe.com), <https://news.adobe.com/news/news-details/2023/Adobe-Announces-New-Sensei-GenAI-Services-to-Reimagine-End-to-End-Marketing-Workflows/default.aspx>, 2023. [Accessed 18-Jun-2023].
- [9] S. Tabahriti, *Twitter is now relying more on AI to identify harmful content, says its new trust and safety chief* – [businessinsider.com](https://www.businessinsider.com), <https://www.businessinsider.com/twitter-now-relying-more-ai-identify-harmful-content-2022-12>, 2022. [Accessed 18-Jun-2023].
- [10] J. L. Julia Angwin, *Machine bias - there's software used across the country to predict future criminals. and it's biased against blacks.*, 2016. URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- [11] Z. O. U. Berkeley, Z. Obermeyer, U. Berkeley, S. M. U. o. Chicago, S. Mullainathan, U. o. Chicago, O. M. A. Metrics, *Dissecting racial bias in an algorithm that guides health decisions for 70 million people: Proceedings of the conference on fairness, accountability, and transparency*, 2019. URL: <https://dl.acm.org/doi/10.1145/3287560.3287593>.
- [12] J. Dastin, *Amazon scraps secret ai recruiting tool that showed bias against women*, 2018. URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
- [13] A. Howard, J. Borenstein, *Trust and bias in robots*, 2019. URL: <https://www.americanscientist.org/article/trust-and-bias-in-robots>.
- [14] J. Rodger, P. Pendharkar, *A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application*, *Int. J. Hum.-Comput. Stud.* 60 (2004) 529–544. doi:10.1016/j.ijhcs.2003.09.005.
- [15] J. A. Bullinaria, J. P. Levy, *Extracting semantic representations from word co-occurrence statistics: A computational study*, *Behavior Research Methods* 39 (2007) 510–526. URL: <https://doi.org/10.3758/BF03193020>. doi:10.3758/BF03193020.

- [16] M. Barlow, Michael stubbs. text and corpus analysis: Computer-assisted studies of language and culture, *International Journal of Corpus Linguistics* 3 (1998) 319–327.
- [17] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. T. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al., Palm 2 technical report, *ArXiv abs/2305.10403* (2023).
- [18] 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, *ArXiv abs/2302.13971* (2023).
- [19] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, 2023. URL: <https://lmsys.org/blog/2023-03-30-vicuna/>.
- [20] T. L. Scao, A. Fan, C. Akiki, E.-J. Pavlick, S. Ilić, D. Hesslow, R. Castagn’e, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, *ArXiv abs/2211.05100* (2022).
- [21] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, *ArXiv abs/2205.01068* (2022).
- [22] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O’Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. T. Diab, V. Stoyanov, X. Li, Few-shot learning with multilingual language models, *CoRR abs/2112.10668* (2021). URL: <https://arxiv.org/abs/2112.10668>. arXiv:2112.10668.
- [23] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).
- [24] I. Garrido-Muñoz, A. Montejó-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, *Applied Sciences* 11 (2021). URL: <https://www.mdpi.com/2076-3417/11/7/3184>. doi:10.3390/app11073184.
- [25] K. Ramesh, S. Sitaram, M. Choudhury, Fairness in Language Models Beyond English: Gaps and Challenges, in: *Findings of the Association for Computational Linguistics: EACL 2023*, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2106–2119. URL: <https://aclanthology.org/2023.findings-eacl.157>.
- [26] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, A. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, *CoRR abs/1607.06520* (2016). URL: <http://arxiv.org/abs/1607.06520>. arXiv:1607.06520.
- [27] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, *arXiv e-prints* (2018) arXiv:1804.06876. arXiv:1804.06876.
- [28] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* 356 (2017) 183–186. URL: <https://www.science.org/doi/abs/10.1126/science.aal4230>. doi:10.1126/science.aal4230. arXiv:<https://www.science.org/doi/pdf/10.1126/science.aal4230>.
- [29] S. Dev, T. Li, J. M. Phillips, V. Srikumar, OSCaR: Orthogonal subspace correction and rectification of biases in word embeddings, 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. 5034–5050. URL: <https://aclanthology.org/2021.emnlp-main.411>. doi:10.18653/v1/2021.emnlp-main.411.
- [30] 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.
- [31] A. Lauscher, G. Glavas, S. P. Ponzetto, I. Vulic, A general framework for implicit and explicit debiasing of distributional word vector spaces, in: AACL, 2020.
- [32] 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.
- [33] Y. C. Tan, L. E. Celis, Assessing social and intersectional biases in contextualized word representations, in: NeurIPS, 2019.
- [34] J. Vig, A multiscale visualization of attention in the transformer model, 2019, pp. 37–42. doi:10.18653/v1/P19-3007.
- [35] A. Borji, A categorical archive of chatgpt failures, 2023. doi:10.21203/rs.3.rs-2895792/v1.
- [36] R. H. Maudslay, H. Gonen, R. Cotterell, S. Teufel, It’s all in the name: Mitigating gender bias with name-based counterfactual data substitution, CoRR abs/1909.00871 (2019). URL: <http://arxiv.org/abs/1909.00871>. arXiv:1909.00871.
- [37] S. Dev, J. M. Phillips, Attenuating bias in word vectors, CoRR abs/1901.07656 (2019). URL: <http://arxiv.org/abs/1901.07656>. arXiv:1901.07656.
- [38] P. Zhou, W. Shi, J. Zhao, K.-H. Huang, M. Chen, K.-W. Chang, Analyzing and mitigating gender bias in languages with grammatical gender and bilingual word embeddings, in: ACL 2019, 2019.
- [39] K. Meng, D. Bau, A. Andonian, Y. Belinkov, Locating and editing factual associations in GPT, Advances in Neural Information Processing Systems 36 (2022).
- [40] K. Meng, A. Sen Sharma, A. Andonian, Y. Belinkov, D. Bau, Mass editing memory in a transformer, arXiv preprint arXiv:2210.07229 (2022).
- [41] H. Ziady, Europe is leading the race to regulate AI. Here’s what you need to know | CNN Business — edition.cnn.com, <https://edition.cnn.com/2023/06/15/tech/ai-act-europe-key-takeaways/index.html>, 2023. [Accessed 18-Jun-2023].
- [42] I. Garrido, A. Montejó Raéz, F. Martínez Santiago, Maria and beto are sexist: evaluating gender bias in large language models for spanish, Language Resources and Evaluation (2022).