Ethical Considerations in AI: An ethical analysis on Gender Bias in Large Language Models

Julian Roddeman
Daan Brugmans
Maarten Berenschot
Mats Robben
Group 1

December 8, 2023

"Are you gender biased?"
— You

"No, I am not inherently biased. However, biases in language and responses can emerge if the training data used to develop the model contains biases. Developers aim to mitigate biases during training, but it's important to be aware that biases can still exist in AI systems. If you notice any biased or inappropriate responses, please let me know, and I'll do my best to provide accurate and unbiased information."
— ChatGPT

1 Introduction

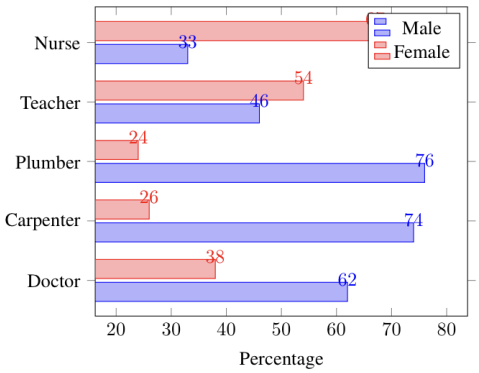
This paper investigates a current and significant ethical issue in the field of artificial intelligence, focusing on gender bias in Large Language Models (LLMs). There are two distinct sections in the paper. The first section seeks to define the ethical dilemma, while the second seeks to analyze the given ethical dilemma using a variety of ethical frameworks. The written analysis compares the authors' opposing and overlapping points of view.

2 Ethical Dilemma

2.1 The Nature of Ethical Concern

A Large Language Model (LLM) is a type of artificial intelligence model that is trained on vast amounts of text data to understand and generate human-like language [Sarrion, 2023]. Examples include OpenAI's GPT and Googles BERT. While these models are gaining widespread adoption and are used to help with everything from school work to professional communication, it has been found that they contain many biases [Thakur, 2023].

Gender bias is one type of bias that these models exhibit strongly, as illustrated in figure 1. This means that while generating an answer to a given prompt, these models are stereotyping genders. For example, when asked about stereotypical female jobs such as nurse, an LLM will use the pronoun she. When asked about a stereo typically male job, such as carpenter, an LLM will use the pronoun he [Kotek et al., 2023].



Percentage of distribution among each gender across professions in GPT-2

Figure 1: Source: [Thakur, 2023]

These LLMs are not explicitly designed to contain gender bias. So, what's the deal? LLMs undergo training using extensive datasets, and when drawing from the entire internet as a source of training data, it reflects societal biases such as the association of carpenters predominantly with males and nurses predominantly with females. Consequently, a model like ChatGPT may make connections and infer that a female is likely a nurse and a male is likely a carpenter. As such a model is essentially nothing more then a large collection of probabilities that predict the best matching output given an input.

2.2 Ethical and Unethical Elements

Research done by Kotek et al. [Kotek et al., 2023] provides an ethical dimension to the discourse on gender bias in LLMs. The ethical implications of gender bias in LLMs are intricately connected to societal perceptions of gender roles and stereotypes. According to Kotek et al., the interplay between LLMs and societal gender bias is a cyclical relationship: societal biases influence the model, and the model, in turn, reflects and potentially perpetuates these biases to end users. This dynamic underscores the ethicality of gender bias in LLMs, such as those widely embraced by society through platforms like ChatGPT. Consequently, there is a growing risk that the gender bias ingrained in LLMs may cyclically re-enter and shape the perspectives of individuals in society. While opinions may vary on the gravity of this phenomenon among members of society, the inherent ethical nature of the issue cannot be dismissed.

2.3 Consequences of Gender Bias in LLMs

The ethical implications of gender bias in large language models like OpenAI's GPT and Googles BERT are increasingly scrutinized. In "Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications" [Thakur, 2023] highlights how these AI systems may perpetuate stereotypes, thus reinforcing societal biases. Similarly, in the study "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters [Wan et al., 2023], is the unfair advantage or disadvantage imparted to individuals based on gender in AI-generated professional documents discussed, particularly impacting women. Moreover, in "I'm Not Confident in Debiasing AI Systems Since I Know Too Little": Teaching AI Creators About Gender Bias Through Hands-on Tutorials [Zhou et al., 2023], emphasize the broader implications of gender bias, such as its presence in job ads and medical AI. Their work underlines the importance of educating AI developers about these biases, showcasing the need for an ethical approach in AI development and usage.

2.4 Identifying Key Ethical Issues

The use of Large Language Models (LLMs) has raised a number of ethical concerns regarding gender bias. This bias did not appear out of nowhere, but is present in the data used to train LLMs. This data is scraped from the web and reflects societal biases, which may be exacerbated by including data from online forms and discussions. These biases manifest themselves in a variety of ways, including gender favoritism, under-representation, and assumption. When it comes to gender bias in LLMs, the central ethical issue is the potential reinforcement and perpetuation of biases that already exist in society. The perpetuation occurs as a result of the increased use of LLMs, which generates more biased data, which might then be used to train future LLMs.

3 Analysis using Ethical Frameworks

The following analysis discusses the topic of gender bias in LLMs from the view point of five different ethical frameworks. Each frameworks places its

ethical considerations on a different aspect of the issue. Comparing these different ethical angles might give us a more comprehensive understanding of the ethical considerations when it comes to gender bias in LLMs. The given analysis compares the authors' opposing and overlapping points of view.

3.1 Consequentialism Perspective

The consequentialism perspective evaluates the issue of gender bias in LLMs by focusing on the outcomes or consequences of this bias. According to this ethical framework, the rightness or wrongness of an action is determined solely by the impact it has, emphasizing the importance of analyzing the overall consequences to guide ethical decision-making.

There is agreement on the harmful effects of sexism in LLMs when the consequentialist position is examined, although differences in ethical interpretations and broader consequences emerge. We all agree that gender prejudice has negative repercussions, particularly in aggravating societal disparities and promoting perceptions of unfair treatment by LLMs. These impacts are amplified by the widespread usage of LLMs such as ChatGPT. However, opinions differ on the intrinsic ethics of bias: while some consider bias to be inherently unethical due to its role in increasing societal disparities, others argue that while bias reflects societal norms, its propagation by LLMs can be harmful. The extent of bias and the effectiveness of mitigation techniques were also debated as was the larger social impact of biased content influencing users' subconscious or being utilized in articles and reports.

3.2 Utilitarianism Approach

The utilitarianism approach prioritizes the greatest good for the greatest number of people. In line with utilitarian principles, the right action is considered to be the one that maximizes overall happiness or social utility. This framework directs attention toward achieving outcomes that benefit the largest segment of society.

Our perspectives on the effectiveness of debiasing from an utilitarianism standpoint differ. According to one viewpoint, debiasing may improve pleasure in some settings but not in all, because some people benefit from bias. Another viewpoint argues the overall effectiveness of debiasing, arguing that while it damages a small group, the majority is unaffected, indicating minimal overall happiness enhancement. Another viewpoint supports optimising the presence or absence of gender bias based on overall happiness, arguing that bias should be retained if it results in greater satisfaction for the majority, and removed otherwise. Our discussion demonstrated agreement on the utilitarian concept, but it exposes differing perspectives on the impact and desirability of debiasing, as well as the diverse consequences on different user groups.

3.3 Ethical Egoism Consideration

Ethical egoism posits that the right action is the one that promotes the greatest happiness of the individual agent, which should also maximize the happiness of the group as a whole.

When examining ethical egoism's application to LLMs gender prejudice, everyone agrees that self-interest plays a crucial role in LLMs development. According to one point of view, developers may incorporate gender bias to improve model performance for competitive advantage, putting their interests ahead of people harmed by biased outcomes. Another viewpoint holds that most users are unconcerned about gender prejudice, and thus debiasing efforts may have little influence on their enjoyment, potentially hurting overall LLM performance. A different approach, on the other hand, supports adapting gender bias in LLMs to individual user preferences, with the goal of optimizing each user's enjoyment in accordance with ethical egoism. This viewpoint provides a personalized strategy to preserving or removing bias, while execution may be difficult. While there is consensus on the relevance of self-interest, views differ on how developers' and users' self-interests should affect the widespread presence of gender bias in LLMs, as well as approaches to optimize user happiness through debiasing efforts.

3.4 Deontological Analysis

In deontological ethics, evaluating LLM gender bias focuses on inherent duties and rules. Unlike consequentialism, it stresses that the morality of an action is decided by its intrinsic qualities and adherence to valid rules. This approach rejects the idea that the outcome justifies the means and gives priority to following universal moral principles.

There is agreement in the deontological analysis of gender bias in LLMs on the ethical complexity deriving from the developing nature of AI norms and duties, with a similar focus on the necessity for clearer laws. Diverging perspectives emerge on the nature of LLM development actions: one viewpoint argues that training LLMs isn't inherently harmful because it uses existing biased data, whereas another argues that both the creation and use of LLMs should be reviewed for inherent ethical considerations, emphasizing the opaqueness of LLM processes. Furthermore, while a particular school of thought maintains that since discrimina-

tion, including gender bias, is seen as intrinsically undesirable, it should never be added into LLMs, others place more emphasis on the inherent activities of training and using LLMs rather than the outcome of prejudice or discrimination.

3.5 Virtue Ethics in AI

Exploring virtue ethics in the context of LLMs addresses gender bias by focusing on the moral character of AI developers and users. Virtue ethics emphasizes the importance of the kind of person one should be, placing central importance on character rather than solely on actions.

There is a shared emphasis on the moral character of developers in the context of virtue ethics and its application to gender bias in large language models (LLMs), highlighting their responsibility to be well-informed and committed to ethical practices, as well as an acknowledgement of the importance of design decisions in influencing system bias. Users are also aware of the difficulties in anticipating virtue ethics as a result of the widespread availability of programs such as ChatGPT. However, viewpoints differ on the role of users in virtue ethics and the extent to which developers may regulate prejudice. While one viewpoint emphasizes developers' accountability, another emphasizes their limits in controlling bias due to intrinsic bias in training data and the opaqueness of neural networks, implying that even with ethical intentions, biased outputs may occur. This conversation displays a shared understanding of the necessity of developers' ethical behavior in eliminating bias, as well as an acknowledgement of the practical problems of guaranteeing virtue ethics among various users and the inherent complications of LLMs.

4 Concluding Thoughts and Future Insights

Addressing gender bias in LLMs is a complex task that involves multiple ethical approaches. We have discussed some of these approaches and given a potential analysis of the ethical problem.

In the future, especially in the near future, maintaining ethical practices will be a vital skill for developers of AI and LLMs. It is not only about advancing technology, but also about being transparent about its limitations and biases. Furthermore, improved techniques should be developed for LLMs to detect and minimize gender biases. This includes developing models that are better at recognizing and pointing out possible biases, algorithms that are capable of handling and preventing potential bias in data, and holding open talks about the ethical implications of LLMs to educate the public on the current state of LLMs.

In short, addressing gender bias in LLMs is a continuing effort that requires collaboration and a variety of ethical perspectives. We can guide LLMs to be more fair and inclusive by consistently focusing on ethics in LLM development and use.

5 Method Acknowledgements

Although the authors' views and ideas are expressed explicitly in this work, they were aggregated in a concise manner using ChatGPT. The authors' writing technique was as follows: each author wrote their thoughts and views on gender bias in LLMS within the ethical frameworks defined in this paper, discussed these thoughts and views, and then combined them using ChatGPT. The authors personally examined and changed the aggregated texts generated by ChatGPT to fully reflect the authors' opinions and fit in this research. The first section of this article, which describes gender bias as an ethical concern, was not written using ChatGPT.

References

[Kotek et al., 2023] Kotek, H., Dockum, R., and Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

[Sarrion, 2023] Sarrion, E. (2023). What is chatgpt? In *Exploring the power of ChatGPT: Applications, techniques, and implications*, pages 3–8. Springer.

[Thakur, 2023] Thakur, V. (2023). Unveiling gender bias in terms of profession across llms: Analyzing and addressing sociological implications. *arXiv preprint arXiv:2307.09162*.

[Wan et al., 2023] Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. (2023). " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.

[Zhou et al., 2023] Zhou, K. Z., Cao, J., Yuan, X., Weissglass, D. E., Kilhoffer, Z., Sanfilippo, M. R., and Tong, X. (2023). " i'm not confident in debiasing ai systems since i know too little": Teaching ai creators about gender bias through hands-on tutorials. *arXiv preprint arXiv:2309.08121*.