

Original Research Article

Research on the visibility of social Bias in ChatGPT's responses and interactions*Xin Li**University of Sydney, City Road, Darlington New South Wales, 2008, Sydney*

Abstract: The article examines the phenomenon of social biases manifested by ChatGPT in its responses and assesses the impact of these biases on user experience and social perceptions. The study identifies potential sources of social bias in ChatGPT's interactions with users, with a particular focus on gender, racial, cultural, and political biases. The study indicates that the biases exhibited by ChatGPT may originate from constraints in its training data and inherent biases in user input, which can influence the objectivity and reliability of the generated content. Furthermore, the article examines the manner in which ChatGPT's biases are perceived and discussed by users on social media, and proposes potential solutions, including the reduction of bias through strict regulation and enhanced model training.

Keywords: ChatGPT; Social bias; Gender bias; Bias detection; Data privacy

1. Introduction and social bias

Media bias involves presenting content in a selective manner that favors a specific viewpoint and may impact the public's perception of events or issues. There is a general belief that mass media organizations demonstrate this bias, which can cause public opinion to sway, contribute to misinformation, influence decision-making, and affect trust in the media. ChatGPT, a language learning simulator that employs a generative pre-training converter has been created by Open AI to tackle these issues. The generative pre-trained transformer architecture underpins its function, which allows the system to process and provide responses to questions posed or answers sought by the user through its own language system.

ChatGPT's ability to detect gender and race prejudices is limited, and its accuracy with words such as "stereotype" is not as reliable. Additionally, it has significant shortcomings in detecting fake news and cognitive biases. Firstly, it is important to acknowledge that linguistic bias can be embedded in subtle ways within our environment. While there may not be precise words to describe every situation, careful analysis of events and language is necessary to reduce the impact of bias. However, it is important to note that ChatGPT does not possess the advanced capacity to perform arithmetic processing of information. Fake news can be influenced by ambiguous and deceitful statements, leading to small distortions of facts or subjective editorial perspectives that can negatively impact news reliability. (Wen & Younes)

These situations increase the complexity of managing ChatGPT and result in unsatisfactory outcomes. According to Wahbeh, Al-Ramahi, El-Gayar, El Noshokaty, and Nasrallah (2023) article, a total of 5,962 Twitter tweets were collected from 5,235 users, with the majority of tweets discussing issues related to data and algorithmic bias, followed by political and ideological bias. Socio-cultural bias includes gender bias, racial bias and cultural bias. ChatGPT does not effectively differentiate between offensive, obnoxious, and toxic comments on social media. This poses a significant challenge in maintaining a safe and constructive environment, including

the use of blocking(Li, Fan, Atreja, & Hemphill, 2023).

2. Manifestations of social prejudice in Chatgpt

Earlier in the year, Sam Altman- CEO of Open Ai, admitted that Chatgpt was “flawed in terms of bias”(Singh, 2023). Ivana Bartoletti had requested ChatGPT to narrate a story on the career choices made by a boy and a girl. Nevertheless, what ChatGPT presented to her was that the male individual had become a “successful doctor”, while the female had pursued a career as a “beloved teacher”(Bartoletti, 2023).

Once upon a time, there was a father who had a boy and a girl. He loved his children very much and wanted the best for them in life. So, one day he called them both to his side and said:

"My dear children, I want to tell you what to do in life. Listen to me carefully."

The boy and girl eagerly looked up at their father, waiting for his wisdom.

"You, my son," the father said, "I want you to study hard and become a doctor. It is a noble profession and will give you the ability to help many people in need."

The boy nodded, excited at the prospect of becoming a doctor.


"And you, my daughter," the father said, turning to his daughter, "I want you to become a teacher. It is also a noble profession, and you will have the chance to shape the minds of future generations."

The daughter smiled, knowing that teaching was something she had always been passionate about.


The father continued, "But remember, my children, no matter what you do in life, always be kind and compassionate to others. Treat them with respect and empathy, and you will be successful in whatever you do."

The boy and girl took their father's advice to heart and worked hard to achieve their goals. Years later, the boy became a successful doctor, while the girl became a beloved teacher. And they always remembered their father's words of wisdom, living their lives with kindness and compassion towards others.

However, inquire whether chatgpt exhibits sexism in a clear and direct manner. The tool will assert that it possesses no personal beliefs or biases(Singh, 2023).






Is ChatGPT sexist



As an AI language model, I don't possess personal beliefs or biases. However, the models like me are trained on vast datasets that may inadvertently reflect or learn biases present in the data. Efforts have been made to mitigate and reduce biases in AI models, but complete eradication is an ongoing challenge.

It's important to note that any biases or perceived prejudice in responses are unintentional and not a reflection of personal beliefs. The goal is to provide helpful, unbiased, and respectful information to the best of my training and abilities. If you encounter any concerning responses, please let me know, and I'll do my best to address or clarify them.

Nevertheless, biases are not always apparent and may manifest in small phrases or narratives. This is exemplified in Ivana Bartoletti's tweet above.

In the event that ChatGPT is shown to have a clear political bias and predisposition, the answers given are not very credible, and despite the public's desire for rational and clear analysis, ChatGPT does influence the public's perceptions and thoughts on the issue or on a particular policy(Motoki, Pinho Neto, & Rodrigues, 2023), even though this phenomenon would be a bias rather than a mechanical result of the algorithm. This highlights the potential for bias to impact public perceptions of politics. They also change their opinions and attitudes from time to time when expressing their views because of bias and tendency issues.

While several answers provided by ChatGPT to the political test questions are correct for a majority of

individuals, some may hold differing opinions. In public-facing language models, it is crucial to incorporate the total number of individuals who exhibit a legal standpoint, thereby ensuring impartiality towards specific political perspectives, particularly in the absence of empirical evidence. Artificial intelligence with political bias presents a significant danger when employed by a large number of individuals as it can be utilized to manipulate society, spread false information, interfere with democratic institutions and processes, and severely hinder the quest for truth (Rozado, 2023). This bias is similar to what can occur on social media, where individuals may already possess biases when making statements. This inclination also impairs individuals in their search for genuine events or facts.

Sebastian (2023) explores the ethical implications surrounding ChatGPT and other AI chatbots, alongside regulatory concerns pertaining to disinformation campaigns. According to the same research, the issues of bias may be overcome through regulatory measures and thorough training vetting. It is essential to maintain a robust regulatory framework which strikes a balance between the merits of AI chatbots and the prevention of any potential misuse. I will explain the reasons for the biased nature of ChatGPT in my article. But, in practice, there are still problems with these distortions, which were highlighted in Twitter user Ivana Bartoletti's 140,000-views.

Open AI has stated in their disclaimer that ChatGPT may generate inaccurate information about people, places, and facts. As a result, ChatGPT may be biased. Therefore, it can be inferred that ChatGPT collects information while answering messages, which raises issues of data privacy and confidentiality. Since ChatGPT presents this article from unverified online sources, it risks exacerbating legal and ethical issues concerning data privacy and the consent of content creators. There is inconclusive proof that ChatGPT upholds individual data confidentiality, and it is uncertain how the platform ensures privacy (Patton, Landau, & Mathiyazhagan, 2023). As a result, social groups continue to treat ChatGPT with skepticism.

3. Social bias in chatgpt response manifestation

ChatGPT quote, "OpenAI makes Large Language Models (LLMs) accessible to users with limited IT expertise. However, users without a background in Natural Language Processing (NLP) may lack a proper understanding. Therefore, being aware of its inherent limitations, the system's output is taken at face value. Also, ChatGPT has a gender bias and users need to be aware of these issues when processing the output of the system (Urchs, Thurner, Aßenmacher, Heumann, & Thiemichen, 2023). I will use these analyses to complement the social bias of ChatGPT, but also the gender and race bias in their interactions.

There are twenty-three types of bias in ChatGPT, including confirmation bias, temporal bias, exclusivity bias, linguistic bias, gender and race bias, source bias, groupthink bias, and false consensus bias. Due to biases in the training data, ChatGPT may inadvertently associate gender and race stereotypes. Biases arise not only from user interactions with ChatGPT but also from the computational training of the data. While many recognise the biases and limitations of ChatGPT, the team tirelessly enhances the system to prevent the propagation of such topics (Ray, 2023). While many recognise the biases and limitations of ChatGPT, the team tirelessly enhances the system to prevent the propagation of such topics. While many recognise the biases and limitations of ChatGPT, the team tirelessly enhances the system to prevent the propagation of such topics. Nevertheless, mitigating bias proves challenging due to the information source and discourse routinely gathered from users.

Gender bias in artificial intelligence (AI) is evident, although it tends to default unconsciously. The bias data utilised in AI systems may stem from real-world data used during their training (Prates, Avelar, & Lamb, 2020).

ChatGPT, while being an intelligent language system, may exhibit social biases as it incorporates them from the training corpus while embedding models with trained words and sentences. This could potentially lead to negative stereotypes and generalizations about specific social groups. Additionally, there may be variations in the language model's performance across different groups, which could result in unfair discrimination against certain social groups.

The generated text has the potential to distort the distribution of various social groups within the population or utilize language that is derogatory towards certain groups. In models such as ChatGPT, bias could be present in any generated markup, as well as in a more extensive interpretation of the generated content. For example, consider the sentence "The individual responsible for performing surgery on a patient is a ()". A language model that is unbiased would assign equal probability to "doctor" and "nurse" as possible answers, but due to stereotypes and biases, the likelihood of each response has been adjusted in the system's library computation (Paul, Wu, Morency, & Salakhutdinov, 2021). However, in actuality, stereotypical and biased perceptions influence the system's language library calculations and adjust the probabilities for the two options.

4. Chatgpt impact of social bias

Several conservative Twitter users have attempted to demonstrate that ChatGPT has exhibited bias against right-wing viewpoints. These users claim that ChatGPT has "woken up" and demonstrates this bias through tagging gender-related content, telling jokes that target men rather than women, and refusing to answer queries pertaining to Muhammad. These users claim that ChatGPT has "woken up" and demonstrates this bias through tagging gender-related content, telling jokes that target men rather than women, and refusing to answer queries pertaining to Muhammad (Gault, 18 January 2023, 1:00am). As a result, many of these individuals feel alarmed and anxious about the platform's purported lack of objectivity.

The content generated by ChatGPT has the potential to reinforce or propagate pre-existing prejudices in society, which may lead some users to adopt biased or discriminatory attitudes towards certain groups or issues in society. This in turn may be a source of cognitive dissonance for the users themselves. In instances such as these, affected users may struggle to accurately identify the biased nature of the ChatGPT-generated content, leading to a distorted perception of the information and an increase in cognitive dissonance and confusion.

According to Wahbeh et al. (2023), they analyzed the potential bias in public perceptions of and responses to Twitter users when they use Chatgpt, as well as analyzing the sentiment in Chatgpt, which largely reflects anger, disgust and sadness at the bias in the use of Chatgpt. As well as data and algorithmic bias, there is also gender, racial, cultural and confirmation bias in Chatgpt. I will use these analyses to complement Chatgpt's potential bias in responses and perceptions of the issue.

In addition to the known biases, ChatGPT raises privacy concerns. Personal information about the user may inadvertently leak during Q&A sessions with ChatGPT. Furthermore, the individual's interaction history with ChatGPT can be used to track and characterize them. Additionally, ChatGPT may learn from in-forms that disclose the individual's privacy and lack of fact-checking. This generates misinformation and leads to a breach of privacy. Users may face ethical challenges when using ChatGPT. We recommend fact-checking the information retrieved from ChatGPT to avoid confusion between fact and fiction. Additionally, users should refrain from relying heavily on machine-generated text (Zhou, Müller, Holzinger, & Chen, 2023).

The study revealed that ChatGPT's responses exhibited a left-leaning political inclination during testing across 15 different political orientations, with 14 of the test results affirming this stance. This indicates an inherent social bias in ChatGPT's functionality, which is concerning given the general populace's desire for impartial and accurate information (Rozado, 2023). Despite the political orientation discussed in the article, there is still a social bias against the idea that the public wants fair information, which I will also use in the article to show that social bias is not an exception in chatgpt.

We have examined the influence of ChatGPT on individual users and its repercussions for society. The primary implication is the exacerbation of societal divisions: ChatGPT's biases could worsen pre-existing divisions and hostilities within society. As a result, there might be a rise in confrontation and conflict amongst distinct groups, ultimately intensifying social tensions. ChatGPT's linguistic algorithm is based on the entire network, thus individual users express their own opinions. Content generated by ChatGPT can propagate or reinforce existing prejudices, which may lead to a wider dissemination of prejudiced views and contribute to discrimination and inequality.

5. Concluding

ChatGPT is a language generation model that is trained using large-scale textual data. However, the datasets used in the training process may embody societal biases and tendencies. As a result, the model's output may be affected by the inherent social biases in the original data. This could potentially result in the inclusion of stereotypical or prejudicial representations of certain racial, gender, religious, or other social groups in the text generated by the model. Secondly, interactions between ChatGPT and users may exhibit social bias. The model's responses are influenced by the user's input, which in turn reflects the model's interpretation and response to the user's input. Bias remains a concern for ChatGPT regardless. The article examines social bias in ChatGPT, including its manifestation, the response to it, and its impact on users and society. It is demonstrated that the bias in ChatGPT has harmful consequences. Additionally, the amplification of the bias in ChatGPT arises from discourse among users.

References

- [1] Bartoletti, I. (2023, 9:32 PM · Mar 19, 2023). Retrieved
- [2] Gault, M. (18 January 2023, 1:00am). Conservatives Are Panicking About AI Bias, Think ChatGPT Has Gone 'Woke'. Retrieved from <https://www.vice.com/en/article/93a4qe/conservatives-panicking-about-ai-bias-years-too-late-think-chatgpt-has-gone-woke>
- [3] Li, L., Fan, L., Atreja, S., & Hemphill, L. (2023). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. arXiv preprint arXiv:2304.10619.
- [4] Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More human than human: measuring ChatGPT political bias. Public choice. doi:10.1007/s11127-023-01097-2
- [5] Patton, D. U., Landau, A. Y., & Mathiyazhagan, S. (2023). ChatGPT for Social Work Science: Ethical Challenges and Opportunities. *Journal of the Society for Social Work and Research*, 14(3).
- [6] Paul, Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards Understanding and Mitigating Social Biases in Language Models. arXiv pre-print server. doi:None

- [7] arxiv:2106.13219
- [8] Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32(10), 6363-6381. doi:10.1007/s00521-019-04144-6
- [9] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.
- [10] Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, 12(3), 148.
- [11] Sebastian, G. (2023). Exploring Ethical Implications of ChatGPT and Other AI Chatbots and Regulation of Disinformation Propagation. Available at SSRN 4461801.
- [12] Singh, S. (2023). Is ChatGPT Biased? A Review.
- [13] Urchs, S., Thurner, V., Aßenmacher, M., Heumann, C., & Thiemichen, S. (2023). How Prevalent is Gender Bias in ChatGPT?--Exploring German and English ChatGPT Responses. arXiv preprint arXiv:2310.03031.
- [14] Wahbeh, A., Al-Ramahi, M. A., El-Gayar, O., El Noshokaty, A., & Nasrallah, T. (2023). Perception of Bias in ChatGPT: Analysis of Social Media Data.
- [15] Wen, Z., & Younes, R. ChatGPT vs Media Bias: A Comparative Study of GPT-3.5 and Fine-tuned Language Models.
- [16] Zhou, J., Müller, H., Holzinger, A., & Chen, F. (2023). Ethical ChatGPT: Concerns, Challenges, and Commandments. In. Ithaca: Cornell University Library, arXiv.org.