Examining the Fairness and Bias of ChatGPT

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Nitin Maddi
Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor
MC Forelle, Department of Engineering and Society

Introduction and Background

When we think about the fastest-growing application, a social media or video game may come to mind. But in reality, ChatGPT, a generative chat engine, has broken records by a landslide in application growth (Thormundsson, 2023). Amazed by its accuracy and human-like conversation, the public has been drawn to use ChatGPT to expedite tasks such as crafting emails, writing code, or even creating workout plans for a specific body type. No matter how detailed or tailored ChatGPT's solution might be, as a consumer, we do not know how the engine arrived at the answer it produces. So, we must ask ourselves, where did this response arise from, and why should we accept that this answer is fair?

AI and machine learning models are algorithms that make predictions based on the inputs that it has received. Specifically, a reinforcement learning model is a machine learning model that "learns" by utilizing data that has been previously collected and manipulated so that the characteristics found in the input are already labeled. When the algorithm is given new data that has been unseen up until this point, it can predict a similar output by connecting certain traits that were linked to similar inputs in the past (Whittaker & Crawford, 2018). However, there are still underlying flaws in these algorithms that need to be addressed and improved upon. One of the key flaws in the algorithms is that bias and inequality can arise from a lack of varied data or faulty algorithms, an issue that could be detrimental if left unattended, as it can lead to these models producing worse results for particular social groups. ChatGPT lies under the umbrella of a generative model, which is a form of reinforcement learning that creates new text or output, rather than classifying an input (Lund & Wang, 2023). The same issue of bias is found in these generative models, and is typically an even bigger problem, as the datasets need to be much larger than their predictive counterparts.

ChatGPT has negatively impacted minorities and underrepresented communities due to OpenAI's difficulties collecting sufficient, unbiased data and their failure to consider social effects. My literature review covers how AI, and specifically machine learning algorithms, are prone to bias, as well as show instances of these algorithms being biased in the past and the social impacts that have emerged from this. Past that, my literature review discusses the widespread impact of AI used by consumer technology giants and the black box nature of these corporate machine learning, where the consumer never understands how the model works and only sees the output. To provide the necessary background information to understand the reasons behind the analysis being conducted, I gather information regarding the bias of various algorithms and important knowledge about the datasets used by these algorithms. In addition, I identify which sectors have been most impacted by the development of these algorithms and what the consequences were. In my analysis, I find that ChatGPT has collected data through many different sources, some of which may be biased to negatively impact certain social groups. To ensure that ChatGPT is as unbiased as possible, stricter regulations should be made on what data can be collected and how data should be used to train the AI.

Literature Review

Large consumer technology companies use AI in their business practices which influences billions of people. Google is the most visited website in the US and has a massive advertisement platform, which distributes advertisements to billions of active users across the world. (Walsh, 2012). The recommendations that Google provides through its search engine are backed by AI. Additionally, they have produced large-scale, AI-integrated, computer architecture such as Kubernetes, which is used by technology companies all over the world. Google also is a Cloud Service provider, and each of its services is backed by AI for efficiency purposes. Apple

distributes over 200 million iPhones each year, which are equipped with a substantial number of machine-learning processes to make the workflow fast and smooth (Curry, 2023). AI is involved in iPhone services such as Siri, cameras, recommendations, and supplements many other features. Finally, ChatGPT, the realistic chat engine, has been in development by OpenAI for the past years and was released as a beta test to anyone who wishes to try it. Currently, ChatGPT is in service by Microsoft and is being used to improve the proficiency of their search engine Bing.

AI in corporations acts as a black box, where people are left clueless about the inner mechanisms leading to their results. When consumers interact with AI, they simply input their data and receive an output without understanding how this came to be. We are left to simply trust that the algorithm is working to our best benefit (Savage, 2022). This black box effect occurs because outsiders have no consideration in the AI implementation details. They are excluded from the design process, even if this technology has an impact on them. AI can be developed irresponsibly when details are hidden away and impacted social groups are not involved in the development conversations. (Floridi et. al., 2020).

As previously discussed, machine learning algorithms are prone to bias arising from various sources including input data and algorithm design. Having imbalanced data, where one social group holds a significantly higher percentage of training data than the other, causes greater inaccuracy/bias toward certain groups (Mehrabi et. al., 2021). Biased algorithms can have an impact on user experience, which in turn creates a feedback loop that results in the AI only improving for the social groups that were already benefiting from it (Mehrabi et. al., 2021). This is because the algorithm will collect data mainly from the groups which were initially at an advantage, heightening the disparity of data and prediction quality between the groups. Additionally, the algorithm itself can be biased towards particular social groups. An example of

this is the Line Edge Map algorithm, which has been used in several facial recognition algorithms. This algorithm does a much better job detecting facial feature edges on someone with light skin rather than dark skin due to facial edges in images being more defined between light skin tones and most backgrounds (Kaur et. al., 2020). In this case, the data involved isn't as relevant on the impact on the Line Edge Map's results, and rather the algorithm itself is flawed. These biases rooted in facial recognition led to repercussions in many different fields, such as surveillance, security, and law enforcement, where certain social groups were negatively affected due to low-accuracy algorithms.

There have been multiple instances of socially influential AI containing bias. COMPAS, an AI used to predict recidivism rates, was found to be racially biased due to its tendency to predict that black criminals would recommit crimes at a much higher frequency than other skin colors (Mehrabi et. al., 2021). Black people with similar crimes committed as white people would be unfairly assessed by the algorithm, leading to further propagation of inaccurate results against black people. Another serious occurrence of machine learning bias was when Amazon developed a hiring tool that rated job candidates based on their resumes. The AI taught itself to rate men higher than women for technical roles and would almost always reject any women candidate for a technical position (Martin, 2022). This issue arose because almost 90% of the employees holding technical roles were currently male, so the algorithm unfairly and incorrectly learned from past hired candidates that males were more qualified, without considering the achievements listed in the resume.

The framework that I use to model the interactions of ChatGPT is Latour's Actor-Network Theory. One of the main features of this theory is that human and non-human actors are treated without a difference, decentering the human from the network (Latour, 2005). This

allows for removing the labeling of social/technical relationships. Key terms include the actor, which is something that has an impact on the technological system at hand, and a script, which describes how technical objects expand or constrict human relations and the interactions between people and objects. A script essentially translates into what an object is "saying". This framework will be applied to better understand corporations and their interactions with consumers and AI. Due to the abstraction of human and non-human actors, this framework would analyze AI without distinguishing it as technology. This is very important because most AI aims to emulate human behavior and treating the same as a human may bring about interesting relationships.

Methods

To analyze the fairness of ChatGPT, I collected primary resources, including research papers written by the developers at OpenAI and datasets used by these teams. I conducted discourse analysis on these papers to understand how they utilize techniques to ensure fairness in their finished product. In addition, I analyzed secondary articles regarding the social impacts of ChatGPT in various fields such as academia and sciences. By doing this, I uncovered which areas were considered in the development process, bringing about an understanding of what the developers believed was important throughout the construction of ChatGPT. I focused on gathering sources within the last five years because the field of AI has been advancing at an extremely rapid pace and papers written before this may be drastically different than the current research. Finally, I conducted Actor-Network theory analysis to visualize which human and non-human actors were considered in ethical development to see where bias could form in the development process.

Analysis

Data Bias

Data that is collected and fed into machine learning models needs to be representative of the world's population in order to maintain the prediction quality of the system and to ensure equity for all users. This is where a problem surfaces for ChatGPT, as it is trained on real human conversations to make the speech seem as human-like as possible. Much of this data includes tweets, Reddit posts, scholarly papers, and other information pulled from the internet (OpenAI, 2023). These sources can include biased information that would be learned by the algorithm in the training process. The model may learn to associate certain groups with negative sentiments, even if they are not accurate or fair representations of those groups. Furthermore, the population of people on the internet is not a representative population of the world. Utilizing data solely from internet conversations and articles is not the same as taking samples from the entire population of the world, which leads to inaccurate results for underrepresented communities.

Although ChatGPT is taking measures to establish its reliability and high standards, their method of manually monitoring the data to prevent the algorithm from producing inappropriate content has significant opportunity for failure (OpenAI, 2023). Bias is introduced through human intervention in the process, as there is inherent bias when a human decides what is inappropriate for the AI to learn and what is not. In addition, data is also preprocessed to further filter inappropriate data, but in the same way that bias arises through monitoring, it also accumulates through preprocessing. While there is no way to completely erase human bias, systems can severely limit it by removing all human intervention from the process and instead use machine learning models to conduct this removal process. These models will still contain bias, but they will impact the system much less than individual people directly monitoring and modifying the data.

Along with its internet sources, ChatGPT is constantly collecting data from any user input it receives, which is dangerous because users can manipulate the AI to their own benefit (OpenAI, 2023). If a consumer or another entity is to input personal information into ChatGPT, the algorithm will learn the information (Lund & Wang, 2023). This allows the algorithm to possibly output personal information about other people after being accidentally or maliciously inputted by an external source. Users can intentionally sabotage the algorithm with sufficient inputs and information and lead the model to learn improper information. People may argue that the filtration process may be able to circumvent this issue, however, the pipeline to remove inappropriate data is slower than what is necessary. Additionally, the filtration process may not be perfect, leaving gaps and allowing this data to still be accessible and absorbed by the model.

When using Actor-Network Theory to analyze the situation, we see that the main actors include OpenAI and its shareholders, ChatGPT, and the consumers who use it. ChatGPT communicates with users of its own volition, having no direct control from OpenAI. However, OpenAI sets guidelines on what ChatGPT can say and provides the required training data, which is how the algorithm learns to speak. Thus, OpenAI and its employees have an immense amount of control over ChatGPT, even if it is an independent actor. This also creates a strong link between the consumer and OpenAI, regardless of the fact that there is little direct contact between the two actors. Our network shows us a relation between OpenAI and its customers where the customer feedback is mainly received is through user input to ChatGPT rather than direct discussion with company executives. When OpenAI makes decisions on how to change ChatGPT in the future, it will use this feedback from user input to guide the direction that it takes the model, leading human bias to seeps into the system.

Inequality in accessibility

Academia has been impacted heavily by ChatGPT as students can simply type in the question and receive an answer that is almost entirely correct (Lund & Wang 2023). The recent innovation of GPT-4 was even able to score within the top 10 percentile of LSAT test takers (OpenAI, 2023) and in the upper echelon of many other professional, standardized tests. Its capabilities even include being able to produce scientific writing while also being able to obtain the references for where it received the knowledge from. (Alkaissi & McFarlane, 2023). When a student or professional is given this amount of knowledge in their hands, they can learn and excel at a quicker rate than their peers who don't use the tool.

It has become evident that ChatGPT's proficiency is significantly worse for languages other than English, which poses a major issue in equality among users. When other languages such as German, Chinese, or French were used as inputs, there was a significant decrease in comprehension level (Zhuo et. al, 2023). Resource-poor languages, such as Samoan, displayed an even greater decrease in interpretation. As a result, non-native English speakers or people without English literary proficiency are unable to use this tool to the same ability. On that same note, when looking at the literacy statistics reports from the National Adult Literacy Survey, it is evident that Hispanic and Black adults tended to receive worse results on this examination (Greenberg et. al., 2001). While this can likely be attested to English not being the first language for a large portion of these groups, this doesn't take away from the fact that their English ability is worse than their white counterparts. A recent survey conducted by the Federal Reserve in 2019 showed that white families had more than four times the net worth of black and Hispanic families on average, which highlights that there is quite a disparity in resources between these two communities (Bhutta et.al., 2019). This gap in net worth is almost directly related to the education level, and communities with access to more academic resources inevitably end up

receiving higher education on average (Wolla & Sullivan, 2017). When an academic resource like ChatGPT lacks the ability to provide reasonably similar results for different languages, English becomes a main prerequisite to using the tool to its utmost ability. As a result, ChatGPT only increases this disparity between the two groups by providing much more accurate information for the English-speaking population who are on average more well off than other groups.

As noted before, ChatGPT relays useful information corresponding to a user's input, while collecting data on these inputs as well as the effectiveness of the output. When considering Actor Network Theory, we see a feedback loop form where the customers who enjoy the product and receive better results will continue to use the service and those who don't will stop. ChatGPT's data collection process will continue to collect more and more data from the social groups or in this case, English speakers, who use the chatbot more, and improve the results for these groups, but lack data from the social groups which it initially exhibited lower quality results for. This cycle will end up exacerbating the already present disparity between the groups. Furthermore, we see a very similar feedback loop form between consumers and OpenAI, where the customers who enjoy the product continue to pay OpenAI for its services, causing OpenAI to develop tools to further benefit the paying customers. Actor-Network Theory demonstrates that ChatGPT will evolve to become increasingly skewed toward the customers who already benefit the most from it if changes aren't made to improve the system.

Conclusion

ChatGPT gives people the opportunity to build entire applications in seconds. Stored in its billions of parameters and weights is enough information to be among the top professionals in many different fields. ChatGPT provides immense utility to a student trying to learn and further

their knowledge in niche topics, and at times gives full solutions and explanations which even Google doesn't find directly in its catalog. When certain communities have limited or no access to something this influential, there is a power imbalance that occurs, which can cause the already present divide between groups to expand further. While there would be no imbalance in the first place if ChatGPT was void of bias, the algorithm lacks training data in certain areas, which leads to inequality for the social groups who don't attribute to a large portion of this data. ChatGPT's data collection procedures try to improve this issue by collecting data from users directly, but this leads to another source of bias in the form of human error. In the end, this cycle of bias is extremely difficult to escape due to pressures from stockholders in OpenAI aiming to maximize profits and improve the experience for the current paying users who already enjoy the product, leaving the disadvantaged users further and further behind.

My research aims to expose the deep-rooted issues in this algorithm, which may not have been clear from the user's perspective. This will teach blind supporters to be more skeptical of artificial intelligence in general, as bias is present in every machine learning model, and hopefully keep companies accountable for their actions. Policymakers could begin to pass legislation that constricts the data that is collected, or better yet, set a breadth requirement for collected training data. A requirement like this would force AI developers to consider "enough" social groups and have sufficient training data for each of them. Implementing a policy to slow the development enough to truly understand the social consequences may be a possible avenue, and, in March 2023, 1100+ notable AI signatories signed a paper calling for the 6-month pause of developing AI stronger than GPT-4 (Loizos, 2023). Additionally, this research will hopefully encourage researchers to put additional effort into decreasing bias during the development process. Corporate AI developers in specific should understand the steep consequences of bias,

as their models will impact millions of people. This project could be furthered for future studies by broadening the scope and looking at other high-profile AIs such as Apple facial recognition or the Google search engine. Answering the question of whether this is a ChatGPT-specific issue, or one that plagues all commercial AI will provide insightful information on which algorithms are more bias prone than others. Deeper research would involve investigating and understanding how to minimize the bias in ChatGPT, past what has already been attempted. ChatGPT has greatly extended what the average human can learn or produce with limited domain-specific knowledge. While there are issues in the early stages of development and release, with further refinement, people all over the world could have access to a college-level education from just ChatGPT.

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in CHATGPT: Implications in scientific writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Bhutta, N., Chang, A. C., Dettling, L. J., Joanne W. Hsu with assistance from Julia Hewitt. (n.d.). Disparities in wealth by race and ethnicity in the 2019 survey of Consumer Finances. The Fed - Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances. Retrieved April 1, 2023, from https://www.federalreserve.gov/econres/notes/feds-notes/disparities-in-wealth-by-race-and-ethnicity-in-the-2019-survey-of-consumer-finances-20200928.html

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review*, *34*(2), 257–268. https://doi.org/10.1016/j.clsr.2018.01.004

Callon, M. (1987). Society in the Making: The Study of Technology as a Tool for Sociological Analysis. In W. Bijker, T. Hughes, & T. Pinch, *Social Construction of Technological Systems: New Directions in the Sociology and History of Technology* (pp. 83-103). Cambridge, MA: The MIT Press.

Cressman, D. (2009). *A Brief Overview of Actor-Network Theory: Punctualization, Heterogeneous Engineering & Translation*. Simon Fraser University. https://summit.sfu.ca/item/13593

Curry, D. (2023, January 11). *Apple Statistics (2023)*. Business of Apps. Retrieved February 15, 2023, from https://www.businessofapps.com/data/apple-statistics/

CHATGPT general FAQ. OpenAI Help Center. (n.d.). Retrieved March 11, 2023, from https://help.openai.com/en/articles/6783457-chatgpt-general-faq

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fosso Wamba, S., Bawack, R., Guthrie, C., Queiroz, M., & Carillo, K. (2020). Are we preparing for a good ai society? A Bibliometric Review and research agenda. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3735322

GPT-4 Technical Report. OpenAI (n.d.). Retrieved March 11, 2023, from https://cdn.openai.com/papers/gpt-4.pdf

Greenberg, E., Chan, T., Rhodes, D., & Macías, R. F. (2001). Literacy and Language Minorities English Literacy and Language Minorities in the United States (4th ed., Vol. 3, Winter 2001, pp. 73-75, Rep.). National Center for Education Statistics.

Kaur, P., Krishan, K., Sharma, S. K., & Kanchan, T. (2020). Facial-recognition algorithms: A literature review. *Medicine, Science and the Law*, *60*(2), 131–139. https://doi.org/10.1177/0025802419893168

Latour, Bruno (2005). Reassembling the Social: An Introduction to Actor-Network

Theory. New York: Oxford University Press.

Loizos, C. (2023, March 29). *1,100+ notable signatories just signed an open letter asking 'all AI Labs to immediately pause for at least 6 months'*. TechCrunch. Retrieved April 4, 2023, from https://techcrunch.com/2023/03/28/1100-notable-signatories-just-signed-an-open-letter-asking-all-ai-labs-to-immediately-pause-for-at-least-6-months/

Lund, B. D., & Wang, T. (2023). Chatting about chatgpt: How may AI and GPT Impact Academia and libraries? *Library Hi Tech News*. https://doi.org/10.1108/lhtn-01-2023-0009

Martin, K. (2022). *Ethics of data and analytics: Concepts and cases*. CRC Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Savage, N. (2022, March 29). *Breaking into the black box of Artificial Intelligence*. Nature News. Retrieved February 15, 2023, from https://www.nature.com/articles/d41586-022-00858-1

Sirur, S., Nurse, J. R. C., & Webb, H. (2018). Are we there yet? *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. https://doi.org/10.1145/3267357.3267368

Thormundsson, B. (n.d.). *Topic: Chatgpt*. Statista. Retrieved March 11, 2023, from https://www.statista.com/topics/10446/chatgpt/

Walsh, S. (n.d.). *50 Google Search Statistics & Facts*. Semrush Blog. Retrieved February 15, 2023, from https://www.semrush.com/blog/google-search-statistics/

Whittaker, M., & Crawford, K. (n.d.). *AI Now Report 2018*. Retrieved March 10, 2023, from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Wolla, S. A., & Sullivan, J. (n.d.). *Education, income, and wealth*. Economic Research - Federal Reserve Bank of St. Louis. Retrieved April 4, 2023, from https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. *ArXiv*.