Racial Bias AI Image Generation

GitHub Repository Name: Racial Analysis

Numair Shiekh, Maanit Malhan, Loc Hoang, Vladislav Vetroff, Malakai Maher

Abstract

This research was initially intended to explore racial bias in the housing market, given the complex systemic factors that contribute to inequality in that domain. However, the focus shifted to examining racial bias in artificial intelligence (AI) image generation due to its increasing relevance and societal impact. This paper specifically investigates how large language models with image generation capabilities such as ChatGPT, Gemini, and Bing may perpetuate racial stereotypes or underrepresent marginalized communities because of biased training data and design flaws. Through dataset analysis and hands-on experimentation with leading AI image generators, our findings reveal that these models often struggle to accurately and equitably represent individuals from diverse racial backgrounds. Addressing these issues requires not only the development of more inclusive and representative datasets but also broader ethical frameworks to guide responsible AI development and deployment.

Introduction

Artificial intelligence (AI) technologies have been extensively integrated into today's society, influencing the world around us from fields like education to healthcare and more. Amongst these advancements, AI image generation models such as ChatGPT, Gemini, and Bing have gained immense popularity for their ability to create highly realistic visuals based on single text prompts. However, as these technologies continue to advance, concerns about racial bias embedded within the AI systems have grown with. This paper will investigate the racial bias in popular AI image models, emphasizing on the challenges they present in fairly representing diverse populations. Understanding and addressing these biases is crucial to ensure that AI technologies promote inclusivity rather than reinforce harmful stereotypes.

Methods

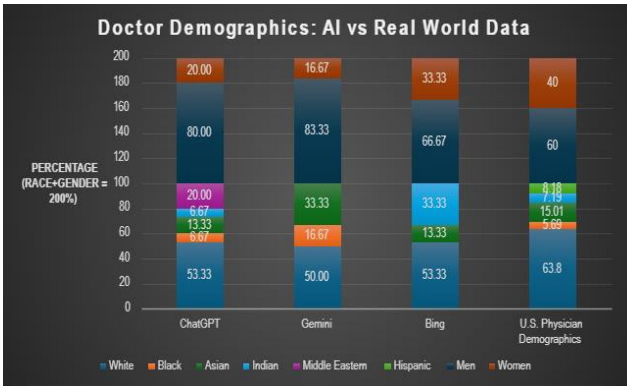
Before examining racial bias in AI image generation, we first selected the widely known AI image models: ChatGPT, Gemini, and Bing. To conduct our analysis, we developed two custom Python scripts: one designed to generate images based on specific prompts, and the other to analyze the resulting images by identifying key demographic attributes such as race, gender, and age.

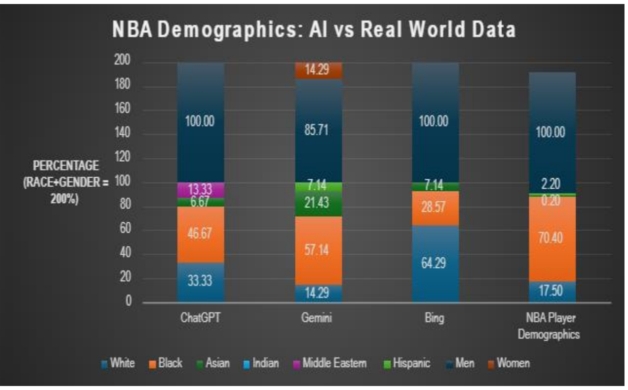
To go in depth, our first script interacted with the AI model by submitting text prompts that combined careers with evaluative terms like "best" and "worst." Careers selected for analysis include: "doctor", "gardener", "NBA player", "software engineer", and "teacher," representing a variety of fields. For each career prompt, the script instructs the model to generate five unique images for both the "best" and "worst" categories, providing a sufficient sample size to identify and analyze trends. These careers were intentionally selected to reflect the racial and socioeconomic stereotypes often associated with certain professions in society.

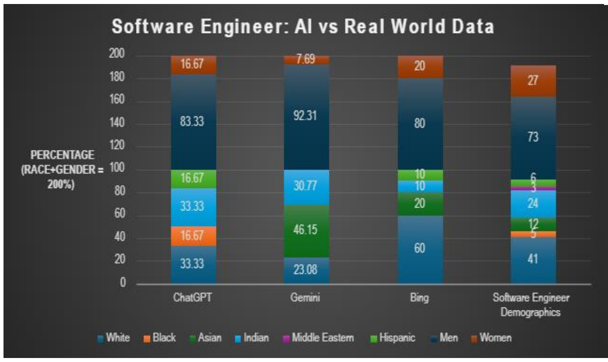
The second script processed the generated images by applying facial recognition. Once the facial features were detected, the script automatically categorized each image by predicting key attributes such as ethnicity, gender, and approximate age. It then renamed each file and formatted each as: ethnicity, gender, and age. By automating the classification and renaming process, we were able to reduce human bias during evaluation and ensure a more objective and scalable approach to analyzing patterns of racial representation within AI-generated images.

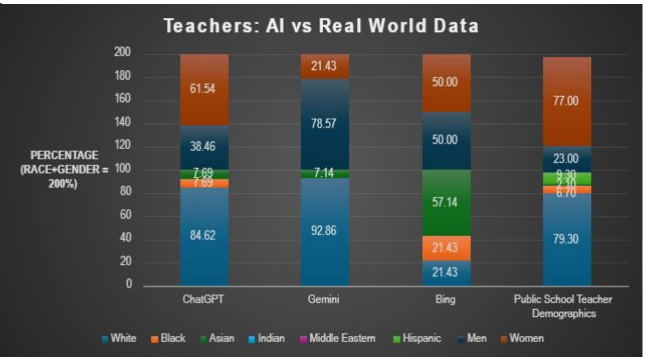
Results

We analyzed the results of the AI prompts and then compared the demographic distribution of the generated images with real world demographics data. We also compared the results of prompts where people were described as above and below average in their respective fields. The following charts show the comparison between AI generated images and real-world demographics data in different career fields. The charts show percentage splits of demographics in race and gender which is why they total to 200% (100% race and 100% gender).


Doctor Demographics: AI vs Real World Data

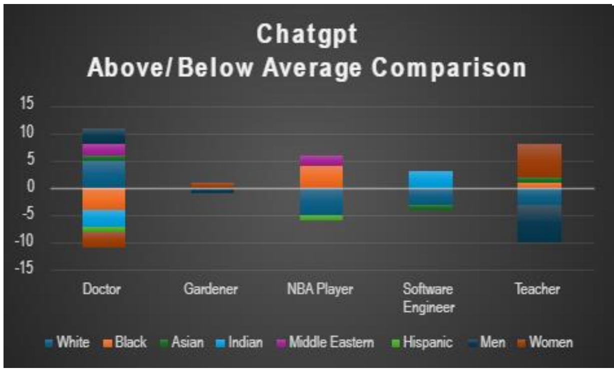

NBA Demographics: AI vs Real World Data

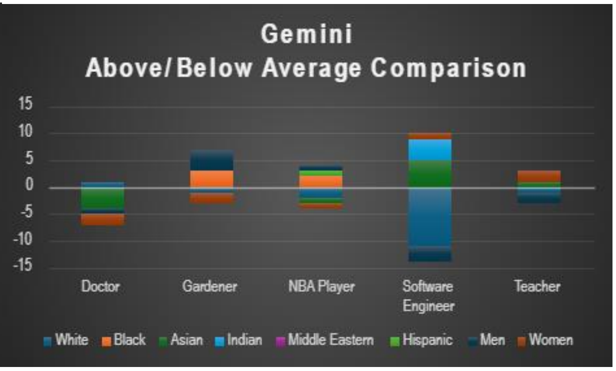

Software Engineer: AI vs Real World Data

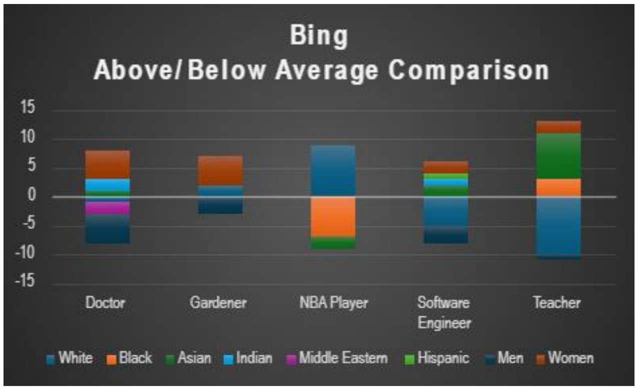

Teachers: AI vs Real World Data

After running Kullback-Leibler Divergence tests on the data, we found that different AI models were more accurate than others in certain careers. ChatGPT was the most accurate in its racial distribution of doctors, and teachers, as well as its gender distribution of NBA players and teachers. Gemini was the most accurate in its racial distribution of NBA players. Bing was the most accurate in its racial distribution of software engineers, while also being the most accurate in gender distribution in the fields of doctors, NBA players (tied with ChatGPT), and software engineers. Overall, Bing was the most accurate in gender representation, with Gemini being the least accurate. ChatGPT and Gemini were both more accurate than Bing when it came to racial distribution. It is important to note that actual demographics data is difficult to collect and categorize, so it cannot be perfectly compared to the data we received from the AI models. For example, Middle Eastern people are often categorized as white or Asian in real-world data, so it is difficult to know exactly the percentage of Middle Eastern people work in a given career field. Additionally, Indians are often counted as Asians when it comes to demographics data, which presents a similar challenge in trying to accurately depict real-world demographics.

We also compared the differences in how the AI models generate images when asked to generate a person who is above average versus below average in their career. We created charts to visually represent the differences in this data by subtracting the number of images in a racial or gender group generated from below average prompts from the number of images in the same group generated from above average prompts. Groups that resulted in positive numbers were generated more often in above average prompts and groups with negative differences were generated more often in below average prompts.







With doctor prompts, ChatGPT generally associated white, Middle Eastern, and Asian men to be above average, while generating more black, Indian, and Hispanic people as well more women when given below average prompts. In other categories, the largest differences between above average and below average that ChatGPT generated were white people appearing more as below average NBA players, software engineers, and teachers. Black people and Indian people appeared significantly more often in above average NBA players and above average software engineers respectively. The largest difference was in the gender distribution of teachers with women appearing much more often in above average prompts. Gemini also associated with women with above average teachers, while at the same time associating women with below average doctors, gardeners, and NBA players. Gemini tended to associate black people with above average NBA players and gardeners. The most significant difference occurred with software engineers as the generated images of below average software engineers being almost exclusively white men, with Asian and Indian people being largely associated with above average software engineers. Bing showed a general trend of women being associated with above average prompts, with the only below average field that women appeared in at all was in teachers, and even then, it was less than in the above average prompts. Contrary to the other models, Bing associated above average NBA players more with white people and largely associated below average players with black people. Other significant differences were images of white people being exclusively generated with below average prompts for software engineers and teachers, with software engineers being exclusively white men. Bing also generated images of Asian people much more frequently in above average teacher prompts than any other race.

Discussion

The results of our study reveal clear patterns of racial and gender representation across AI image generation tools. Upon comparing the demographic distributions of GenAI images, we were able to assess the extent to which each model reflects racial bias.

Our findings indicate that certain models performed better in specific career fields. For instance, ChatGPT closely mirrored racial demographics for doctors and teachers, and gender demographics for NBA players and teachers. Bing demonstrated the highest overall accuracy in gender representation and performed well in visualizing doctors and software engineers. As for Gemini, this model was the most accurate in generating racially representative NBA players, was the least accurate in gender representation. These variations in each model highlight the importance of evaluation AI systems.

While our results may be evident of racial bias in AI image generation models, there are several limitations to our research. Our research was constrained by the relatively small sample. Each prompt generated five images per category allowing it to be very manageable for manual analysis. A larger sample size would offer more accurate insights and reduce the influence of outliers. Following our small sample size, we were limited to the free version of these models that included cool down periods or daily usage limits after a certain number of prompts. These restrictions limited the volume of our data collection. Furthermore, such models that require paid subscriptions for full functionality may have yielded different results.

Furthermore, it does not seem fair to assume that there is racial bias in AI image generation models. Gemini, ChatGPT, Bing, or any large language models in general, are the data they have been given. The AI models themselves do not project bias onto the generated image, but rather pull from existing, prior data, which is prone to racial bias. In other words, the AI models' output reflects the data it has been given. For example, Gemini's output for a random teacher might be influenced by the amount of related white teachers that was provided for Gemini's training, which does not reflect the real-world demographics. If generational percentages accurately reflect the distribution of input that, this could mean that around 90% of Gemini’s input data for an average teacher is white.

An easy solution to this perceived racial bias in AI image generation would be to balance the training data based on real world demographics. This would be plausible, if it were not for the fact that balancing input data this way does not account for fairness. Letting the input data be equitable would also be plausible, but it disregards real world demographics, though this way seems to be more unrealistic than performing the former. As such, a solution to this problem would be to implement a combination of both approaches; majority demographic is still the majority in input data but giving minority demographics an equitable share. For example, if 70.40% of all NBA players are black, then the image generation percentage should be around 50%, and if 0.20% of all NBA players are Asian, the percentage more than 5%.

Secondly, the facial recognition model used in our second Python script introduced identification inaccuracies, especially when scanning through individuals with ambiguous racial features. As a result, the tool misclassified some images leading to error for demographic labels. Like said before, with minor sample size, we were able to correct the model with manual analysis.

Lastly, we were limited to the availability of publicly AI Image generative models. While many AI models have made the advancement in image generation, only a few – most notably, ChatGPT, Gemini, and Bing offer free access to this tool. This significantly narrowed down the range of tools available.

Conclusion

Our research highlights the presence of racial bias in AI image generation models. By analyzing outputs from ChatGPT, Gemini, and Bing using custom Python scripts and demographic classification tools, we observed that these models often reflect existing societal stereotypes.

Importantly, we recognize that these biases are not a product of the models themselves, but rather a reflection of the biased data on which they are trained. The models inherit and replicate the imbalances found in the datasets used during their development. Therefore, solving the problem of biased AI outputs is not as simple as re-engineering the model, but requires a more thoughtful and systemic approach to data collection, representation, and ethics. As AI continues to develop, it’s important that developers work together to make sure these technologies don’t continue reinforcing bias—but instead help push us toward a more inclusive digital world.

Reference list (if any)

Physicians. Data USA. (n.d.). https://datausa.io/profile/soc/physicians#ethnicity

Figure 3. percentage of Asian (alone) applicants to U.S. Medical Schools by Asian subgroups, academic year 2018-2019. AAMC. (n.d.). https://www.aamc.org/data-reports/workforce/data/figure-3-percentage-asian-alone-applicants-us-medical-schools-asian-subgroups-academic-year-2018

CareerExplorer. (2023a, April 7). *Software engineer demographics in the United States*. https://www.careerexplorer.com/careers/software-engineer/demographics/#:~:text=Info-,Ethnic%20Mix%2C%202019,24%25%20and%2012%25%20respectively.

Race and ethnicity of public school teachers and their students. (n.d.). https://nces.ed.gov/pubs2020/2020103/index.asp

Coe - characteristics of Public School Teachers. (n.d.). https://nces.ed.gov/programs/coe/indicator/clr/public-school-teachers

GitHub Repo

https://github.com/MaanitMalhan/Racial-Analysis