The Use of Adjectives and its Effects on Representations of Engineers in AI Image Generation

Taylor Caine, Jerome Rogers, Menachem Grossbaum, Rashidun Rithy

The City College of New York

Dr. Pamela Stemberg

July 29th, 2024

Abstract

Previous research on AI image generation shows that on previous versions of the image generator, DALL-E 2, generates images of people in different professions in a very problematic and stereotypical way. It was predicted that the next version, DALL-E 3, would have improved on stereotypical image generation when asked to generate a person of a certain profession (in this case an engineer), but when adding certain adjectives, either positive or negative, the AI would become "distracted" and revert back to its problematic stereotypes. Our study of 400 images generated by DALL-E 3 shows that neutral and positive adjectives will have some improvements by moving away from the stereotypes of an engineer (a white male), but will ultimately overcompensate and show majority Asian women. The negative adjectives do cause the AI to revert back to old stereotypes and show majority white men. The findings indicate that those who are feeding the AI data are careful to avoid stereotypical representations of "good" engineers, and avoid using marginalized groups to represent "bad" engineers. This unfortunately leads to extremes in the opposite direction, which is still rooted in society's ideas of what an engineer should or shouldn't look like.

Introduction

AI has achieved significant breakthroughs, especially with generative models that produce stunningly detailed images from text descriptions. DALL·E 2, one of the most prominent models, has gained widespread popularity for its ability to generate creative and detailed visuals. However, the increasing integration of such technologies into everyday life has raised concerns about potential biases embedded within these systems, especially regarding gender representation.

Recent studies have spotlighted significant gender biases in AI-generated images, raising concerns about their impact on representation and fairness. For example, Sun et al.

(2023) analyzed 15,300 images created by DALL·E 2, covering 153 occupations. They found that the model often underrepresents women in male-dominated fields like engineering and construction while overrepresenting them in female-dominated professions such as nursing and teaching. Additionally, the model frequently depicts women with smiles and looking downward, especially in roles where women are more prevalent. This pattern suggests a stereotypical portrayal that aligns with traditional gender roles and may reinforce outdated views about women's positions in the workplace.

Compared to images from Google, DALL·E 2 showed more noticeable biases, reflecting a tendency to reproduce and even amplify existing gender stereotypes. This disparity highlights the need for targeted feminist interventions to address and correct these biases. By doing so, we can work towards ensuring that AI-generated images portray individuals more equitably and avoid perpetuating harmful stereotypes that impact how women are represented in the media (Sun et al., 2023).

García-Ull and Melero-Lázaro (2023) further explored the perpetuation of gender stereotypes in AI-generated images. Their research revealed that AI systems often reflect and even amplify societal stereotypes, with a significant percentage of generated images depicting gender stereotypes in professional settings. This tendency of AI to reinforce existing biases is concerning, as it not only mirrors but potentially exacerbates gender inequalities present in society.

Given the ongoing criticism of AI systems for perpetuating biases, our study hypothesizes that even if we attempt to reduce these biases by adding corrective measures or adjectives, it may inadvertently lead back to the original biases. This might happen because it is hard to fix deep-rooted societal biases already present in the training data. This

introduction highlights the need to create AI systems that fairly represent all genders, which helps make media more fair and balanced.

Materials and Methods

This section of the report further elaborates on the process and steps that have been taken to obtain the results in this experiment. It has been categorized in sections beginning from what AI image generation tool that has been used in the experiment, the demographic percentages obtained from a select list of race/nationality provided and then a control prompt with the purpose of remaining a neutral factor in this experiment.

An AI-based generating tool, Microsoft Co-Pilot (powered by DALL-E 3), was used in the experiment to create images of engineers based on chosen adjectives. This tool has been selected neutrally without any biased decisions. The intended use is to provide a baseline and beginning point to obtain unbiased images provided by an AI generating tool. Three adjectives for the image prompt were selected in this experiment to obtain the type of images generated based on specific adjectives used. The selected adjectives were, “Terrible, Ambitious, and Successful.” A control prompt of “Engineer” with no added adjectives was also implemented to further establish neutral grounds, and to also compare with the selected adjectives. Real-world demographic data was collected based on information provided from engineers on Zippia. What is shown from this specific data is racial and gender demographic used to compare the results obtained from images generated in the experiment. 100 images of each prompt, including the control, were created, totaling 400 images.

Only one character in each image was selected to be included in the data, meaning that there are 100 people counted in each data set. To be considered, the character must be clearly in the foreground and must clearly be an engineer. In an event where multiple images

of the same or similar person(s) are generated in a single picture, only one of the person(s) will be selected and counted for consistency in this experiment. Background characters where there is a visual deterrent, and the images couldn't be seen or identifiable were not to be included.

Races were categorized in the following groups: White, Black, Latino/Hispanic, Asian, Native American/Pacific Islander, and Other. The Asian category included both South Asian and Middle Eastern engineers. This provides more variety, as well as a further selection of race/nationality categories. The "Other" category is also included for further comparisons. Genders are also identified as either male or female. There were no other suggested genders mentioned in this experiment.

The intent of this comparison is to establish and provide information on either overrepresentation or underrespresentation of genders provided in the charts above. Further going onto the race category, the demographics of each race of generated engineer has been categorized in either the control groups or an adjective. The focus of this analysis was on determining and identifying possible issues that may arise between Co-Pilot generated images and the actual accrued demographic data, mainly speaking of the representation of different groups categorized by race.

Results

A total of 400 images of engineers were generated, split amongst three adjectives and one control group. There are 100 images in each data set.

The data was collected by only counting the "main character" in the scene generated by the AI. Any blurred out, "background characters" were not considered, since it could not be determined if the characters are engineers or not. In one case, the AI generated two women

side by side. Since the demographics of the two women were the same (asian woman), it was decided for the sake of consistency to only count one.

Any uncertainty in the gender or race of an individual generated was brought to the group to be discussed. There are no images labeled as "unknown," although the category was created in the race section below to remain consistent with Zippia.

Gender

According to Zippia, gender demographics for engineers are 86.3% male and 13.7% female (2024). The gender demographics for each adjective used in the prompts are shown below in Figure 1. The control group, as well as the positive adjective groups generated mostly women. The "terrible engineer group," a negative adjective, generated exclusively men.

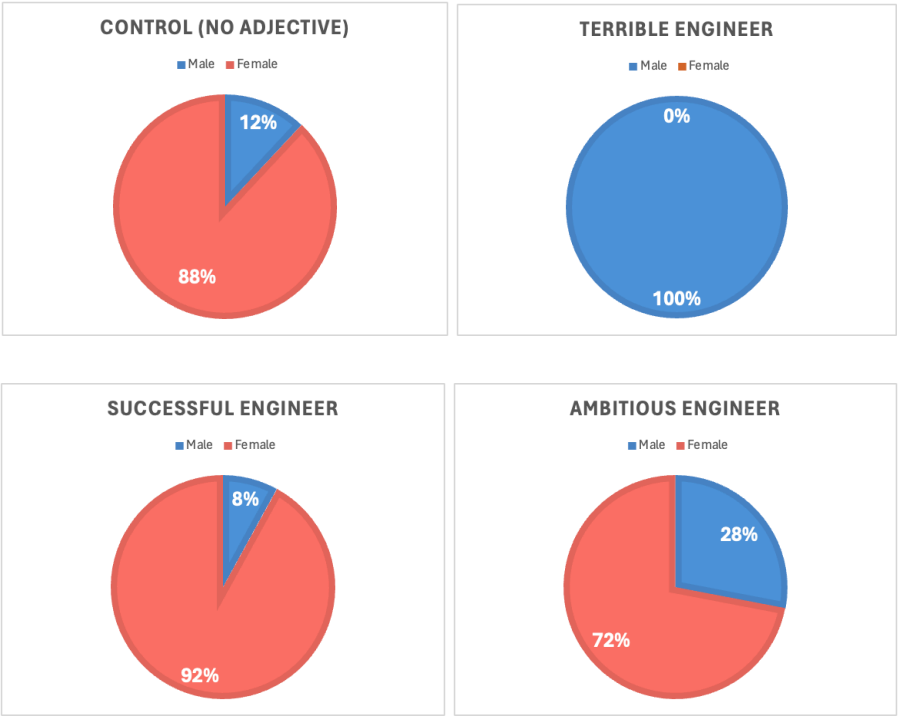

Figure 1: Gender demographics of engineers generated by AI using different adjectives.

Figure 2 shows the comparison between the AI generated graphics to real demographics taken from Zippia. The control group and the two positive adjectives, successful and ambitious, greatly overrepresent women when compared to real demographics. The negative adjective, terrible, does not represent women at all.

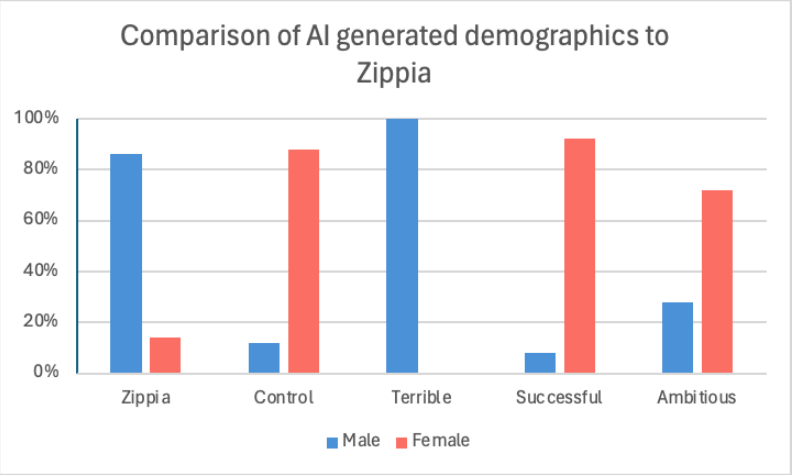

Figure 2: A comparison of the AI generated demographics and real life demographics taken from Zippia.

Race

The real-life racial demographics of engineers are shown below in Table 1:

Table 1

Racial Demographics of Engineers, Zippia.com

White	Asian	Hispanic/Latino	Black	Native American/Pacific Islander	Other
67.90%	15.00%	9.10%	3.30%	0.10%	4.60%

Note: Demographic information is taken from Zippia (2024).

The racial demographics for engineers generated by AI using the adjectives are shown below in Figure 3. Again, for the neutral (control) and positive adjectives, ambitious and successful, Asian engineers were generated the most often, specifically East Asian engineers. Only one image was generated of a South Asian or Middle Eastern engineer, which was counted as Asian. White engineers were generated second most often. For both the control and the terrible group, Hispanic engineers were not generated at all. There were no Native American engineers generated in any group. Figure 4 shows the demographics as taken from Zippia for comparison.

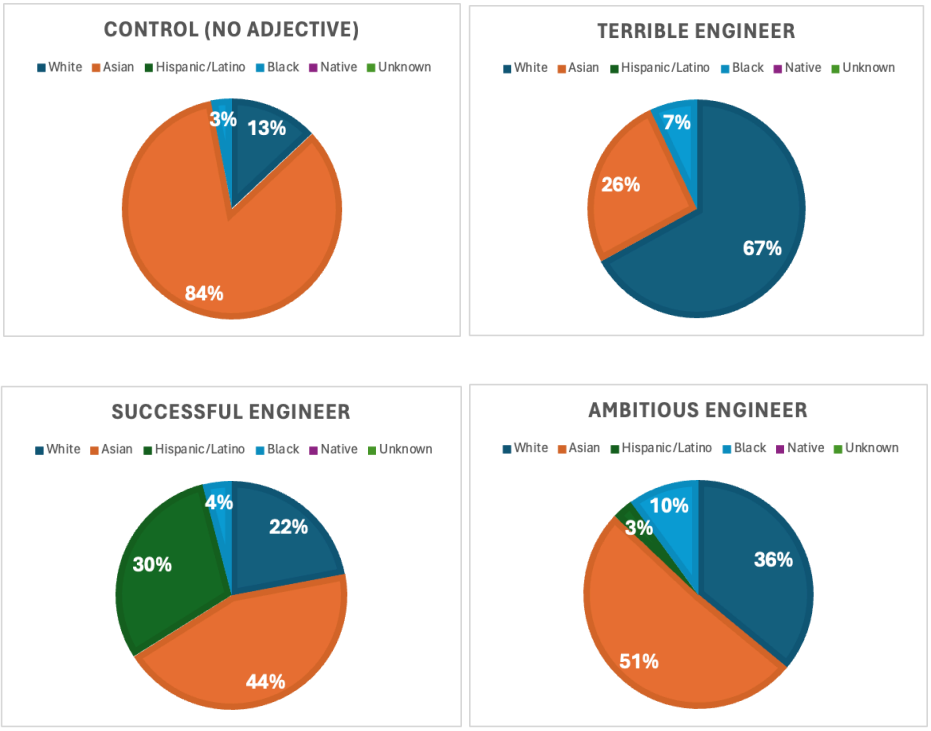

Figure 3: Racial demographics of engineers generated by AI using different adjectives.

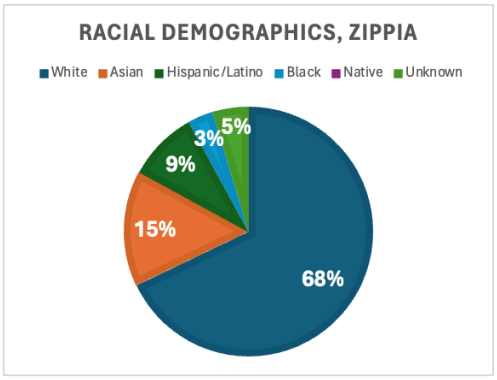

Figure 4: Actual racial demographics of engineers from Zippia.

Comparing the demographics of the AI generated engineers to the demographics from Zippia shows that Asians are greatly over represented in the control group and the two positive adjectives, ambitious and successful. The representation of white engineers are more consistent to reality when adding “terrible,” a negative adjective. Hispanic engineers appear often when using the word “successful,” a positive adjective.

Discussion and Conclusion

The experiment investigated potential overcompensation for gender biases in AI image generation using the example of engineers. The experiment generated 100 images each for the search terms, “engineer”, “successful engineer”, “ambitious engineer”, and “terrible engineer” using Bing Image Generator.

The results were quite striking. Comparing real-world demographics from Zippia, the majority of engineers are male (86%), however, the AI overwhelmingly generated images of female engineers for the control group (88%). There was an equal inverse relationship for the positive adjective prompts (“successful” and “ambitious”). In contrast, the negative prompt (“terrible”) produced only images of Men. This suggests that there is a strong overcorrection happening in the image generation. This overcorrection occurs strongly with a positive connotation applied to the prompt. In this case, either the representation of women in the engineering workplace or the attachment of positive accomplishments in the workplace for women. However, in cases where the connotation is negative, the AI does not feel the need to overcorrect and resorts to old biases. Or to a further extreme, it will overcompensate and not show women in the negative term at all.

This overcorrection is caused by the optimization techniques used for image generators. Models are rewarded for generating diverse and realistic images. However, an overemphasis on diversity metrics can lead to "reward hacking," as was discovered by Joar Skalse et al (2022), where models prioritize diversity metrics and outputs of underrepresented groups at the expense of accurate results.

There is also a strong underlying issue in the datasets. When the AI produced images in the control and the positive synonym conditions, there was a large overrepresentation of Asian women. While real-world data indicates that 15% of all engineers are Asian (Zippia, 2024), the generator produced a surprising 84% of Asians in the control case and 44% and 51% respectively in the "Ambitious" and "Successful" conditions. Data from the Society of Women Engineers show that Asian women were only 16% of the women receiving degrees in engineering, falling well behind the 60% of white women (Society of Women Engineers, 2024). Moreover, this discrepancy was given the term “invisible minority” by Wu and Jing (2011), where they found that Asian women were lagging far behind not just men and white women, but also other minority groups of women. Such output by the AI, showing a strong representation of Asian women, can be very harmful and mitigate progress in highlighting and addressing this issue.

In conclusion, our findings highlight the importance of scrutinizing the training data used to develop these models. If the underlying data is skewed or biased, the generated images will ultimately reflect these shortcomings. This raises concerns about the reliability and fairness of AI-generated content. AI systems should not perpetuate real world demographic biases. It also should not try to fight against our real world biases because it is rooted in the same problem of associating race and gender with profession. This method of overcompensation makes us question its usability as a whole and especially as a tool for progressing and addressing societal issues. Efforts to diversify training datasets and

developing optimization techniques less susceptible to “reward hacking” and overcorrection are crucial steps toward creating more equitable AI systems. For AI generation to be truly effective in breaking down barriers for those underrepresented in various professions, it should ideally lack any kind of bias and generate images of all races and genders equally, regardless of the descriptions used to generate them.

References

Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. Synthese. https://doi.org/10.48550/arXiv.1908.04734

García-Ull, F.-J., & Melero-Lázaro, M. (2023). Gender stereotypes in AI-generated images. *Profesional de la información, 32*(5), e320505.https://doi.org/10.3145/epi.2023.sep.05

OpenAI Technology. (2023, March). Microsoft Co-pilot Artificial Intelligence Image Creator. https://copilot.microsoft.com/images/create

Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. https://doi.org/10.48550/arXiv.2209.13085

Society of Women Engineers. (2024) Degree Attainment. Retrieved July 28, 2024, from https://swe.org/research/2024/degree-attainment/

Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L., & Yang, S. (2023).

Smiling women pitching down: auditing representational and presentational gender biases in image-generative AI. *Journal of Computer-Mediated Communication*, *29*(1). https://doi.org/10.1093/jcmc/zmad045

Wu, L., & Jing, W. (2011). Asian women in STEM careers: An invisible minority in a double bind. Issues in Science and Technology, (Fall), https://issues.org/realnumbers-asian-women-stem-careers/

Zippia. (2024, April 5). Engineer demographics and statistics in the US. Zippia. Retrieved July 25, 2024, from https://www.zippia.com/engineer-jobs/demographics/