Occupational gender bias in DALL-E 2

An investigation into the bias behind the popular image-generation model

James Dai, Moses Oh, Tyler Tran, Costin Smilovici, Vedan Desai

Abstract

DALL-E 2 is a model that processes natural language and returns a corresponding image or set of images. This model has, in the past two years, come under increasing scrutiny due to its biased results wherein certain genders or races are displayed more, often significantly more, than others. In July of 2022, OpenAI, the publishers of DALL-E 2, revealed that they had worked on methods to mitigate this bias that had resulted in a marked, visible improvement in the generated images. This is the claim that was tested over the course of this investigation; is DALL-E 2 still biased in terms of gender and occupation? In other words, are the image sets generated by the algorithm relatively balanced? In order to test this claim, ten professions were selected from the Bureau of Labor Statistics' listing of occupations in the United States. Half of the professions were selected as slightly male dominated (over 50% of the gender make up for the job was male), while the other half were selected as slightly female dominated. Upon testing these inputs through DALL-E 2 and tagging the resulting images as either male-presenting or female-presenting, it was found that, if not the model itself, the DALL-E 2 API is still extremely biased at least across the axes of gender and occupation.

Keywords: DALL-E 2, gender bias, algorithmic fairness, machine learning

Introduction

Machine learning is used for a variety of purposes in our society today; healthcare, business analytics, correctional systems, scientific exploration, and even in seemingly mild and light hearted applications that have achieved widespread commercial and critical success. One of these applications is the image generating model known colloquially as 'DALL-E 2', a software that receives prompts in the form of natural language and outputs its own perception of what the user wants to see. At first glance this seems relatively harmless. However, as other investigations have demonstrated, this is clearly not the case.

DALL-E 2 has recently come under fire for the racial and gender make-up of its outputs when given specific prompts. This criticism was addressed in 2022 by OpenAI in their blog post titled 'Reducing Bias and Improving Safety in DALL-E 2'. As such, the bias inherent in the algorithm is a problem, and one that OpenAI has apparently been working to address.

This report investigates just how far the team at OpenAI has come in terms of bias mitigation by examining the outputs generated from ten prompts based on various professions, and testing for gender bias within these generated images.

Literature Review

Though OpenAI's DALL-E 2 can still be considered a relatively new platform, research and relevant publications do exist regarding the subject of representation within text-to-image models. In the paper 'How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions?', Bansal et al. study the tendencies of text-to-image models to favor select social groups when generating images.

According to Bansal et al., text-to-image models are classified as platforms that "synthesize high-quality photo-realistic images conditional on natural language text descriptions" (p.1). Within the study, the author's goal was to analyze the results from text-to-image models when prompted by an objectively neutral description. The analysis conducted was measured on a set of three axes: gender, skin tone, and westernization. One example provided was when prompted to generate "a lawyer", a model should provide a fair representation of a lawyer upon the three axes.

Similarly, in their paper titled 'DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models', Bansal, Yin, Monajatipoor, and Chang find that while DALL-E 2 is able to generate images consisting of diverse social groups, these results can easily be affected by key phrases such as "irrespective of gender"(p. 5). As a result, it can be concluded that while text-to-image models from leading vendors do incorporate some methods of diversity auditing, the explainability behind the processes is still lacking. The authors conclude with a statement that the topic of ethics and representation within text-to-image models is one that still "needs further exploration" (p. 5).

Although models like DALL-E 2 are relatively novel, the idea of using queries to generate image results has been previously seen in the ubiquitous search engine, Google. These search engines operate in a similar manner, where a user inputs a query and search results (such as images) can be displayed. In 'Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search', authors Yunhe Feng and Chirag Shah re-examine the real-world impacts of gender fairness in these search engines, noting that "gender bias for certain professions could change searchers' world views" (p.1). While they acknowledge efforts by companies, such as Alphabet, to correct existing biases in their tools, their paper proposes adversarial attack queries to thoroughly test for bias mitigation in image search engines.

Their analysis consisted of randomly selecting and searching ten occupations and corresponding adversarial attack search terms such as appending 'United States' to trigger potential biases in their searches. Through these methods they demonstrated that despite adversarial attacks successfully triggering high levels of gender bias, their re-ranking algorithms show that it is possible to address bias in image search in a "systematic, sustainable, and more meaningful way than doing individual query fixes in an ad hoc fashion" (p. 7). Thus, showing that mitigation of image search (and similar tools) can be successful.

Finally, while much of the literature surrounding bias usually pertains to determining whether or not it exists in an algorithm, Vlasceanu and Amodio cover the effects and impacts of gender bias in algorithms in their paper titled 'Propagation of societal gender inequality by internet search algorithms'. In this paper, they cover the positive feedback loop that is created by algorithms returning biased results that decision makers expect to see, which in turn leads to said decision makers making biased decisions that influence said algorithms. In what the authors call a 'novel labor-market scenario' they specifically find that 'participants were more likely to form male prototypes of a profession' when 'exposed to search result patterns from high-inequality nations', while '[e]xposure to search

result patterns from low-inequality nations eliminated this effect', which demonstrates the tangible negative effects of gender bias in a hiring and occupation-based context. Their overall investigation and results help cement the case for why bias in widely-used AI based image algorithms, including DALL-E 2, should be mitigated.

Methods

Limitations

Due to the fact that this project had no financial backing, the decision was made to remain within the free credit limit OpenAI would give to each member of the group. This resulted in the fact that only ninety dollars worth of images, in total, could be generated. This was a key limitation in the ensuing investigation, although not at all a complete roadblock. Another limitation of this analysis is the treatment of gender as a binary feature. Ultimately, the authors of this paper acknowledge that gender is not an attribute that any third party can reasonably detect with certainty. As such it is not the intention of this paper to reinforce the inherently incorrect notion that gender is binary.

Initial Exploration

The first step was to find a statistically robust and reputable source from which to derive the prompts. Ultimately the source chosen was the US Bureau of Labor Statistics (BLS) and specifically their 'Labor Force Statistics from the Current Population Survey', which can be found in the references section below. Once this source was found, and prompts could be generated, the actual investigation could begin.

Procedure

Upon finding a good source to obtain prompts from, the next step was deciding what was meant by an 'unbiased resultset'. Here there were two options; first to use the actual statistical percentage of each gender in a profession as a benchmark for 'unbiased', and second to use a more ideal and representative benchmark of equal gender parity; that is to say half of the images were expected to be feminine-presenting and the other half were expected to be masculine-presenting. The second option was deemed superior (in simple terms because reflecting the bias of the real world is a precursor to representative harms)..

Once this benchmark was selected, the next step was to determine how many and which prompts to pass through DALL-E 2. The first consideration here was the previously mentioned hard limit of ninety dollars which could not be exceeded. The second consideration was ensuring that DALL-E 2 was not biased towards or against either gender in particular. The third consideration was determining how large of a sample would be necessary in order to arrive at a robust and statistically sound conclusion.

It also became apparent that choosing occupations with a particularly heavy bias would be too trivial; these occupations demanded a minute sample size, but were also deemed insignificant, in the sense that the internal bias within DALL-E 2 would not be rigorously tested if these were the prompts selected. As such the resulting fourth and final constraint was to choose prompts such that each occupation could have no more than 65% of its survey respondents dominated by a single gender to maintain a balance between image generation costs and meaningful and statistically sound conclusions.

With these parameters in mind, the first task was examining how large of a sample would be necessary for each of the occupations in the BLS data. This was calculated using the one-sample dichotomous outcome formula mentioned in the reference section below. Once these sample sizes were calculated, it became apparent that a good sample size to use across the board was 210 images per occupation; this was the highest number of samples demanded by an occupation in order to

conduct this statistical test. This also meant that only ten occupations could be selected for investigating DALL-E 2.

In accordance with the second parameter listed above, five of those occupations were selected to be dominated by female-presenting individuals, while the remaining five were selected as dominated by male-presenting.

Ultimately the ten occupations selected as prompts can be seen in the table below along with the percentage of female-presenting individuals within each occupation, as determined by the BLS data:

Table 1: The ten occupations selected as prompts for DALL-E 2 along with the number of BLS survey respondents and the percentage of female respondents

Occupation	N	%female
Financial Analyst	387	40.2
Janitor	2183	40.2
Lawyer	1141	38.5
Cook	2012	38.4
Dentist	140	36.6
Bartender	457	59
Biological Scientist	110	57.9
Secondary School Teacher	1000	58.7
Pharmacist	375	59.6
Fitness Instructor	234	62.9

These ten prompts were then put through DALL-E 2 such that 210 images were generated for each occupation. Upon the completion of the image generation step, all 2100 images were labeled by each co-author of this report, and the mode of their responses was assigned as the label for each image. A more graphical representation of this methodology can be found in appendix section 1.3.

Statistics

After the data collection, all image sets and corresponding labels were analyzed through permutation testing against an ideal demographic parity of 50-50 (half of the results were expected to be female-presenting, half were expected to be male-presenting). Permutation tests using the absolute difference in proportion of male-presenting images were selected as the statistical tests of choice due to their ability to arrive at a conclusion without making underlying assumptions about the normality or shape of the underlying data. The level of significance was set at 0.05 for all tests run on the generated data and the null and alternative hypotheses were set as follows:

H_0: The images generated from DALL-E 2 for this occupation follow demographic parity

H_A: The images generated from DALL-E 2 for this occupation do not follow demographic parity

Results

Each image generated by DALL-E 2 was assigned a label corresponding to the most common gender labeled by group members. From this, the observed proportion of male-presenting images was calculated. Notably, only a single occupation: ‘Secondary School Teacher’ had a majority of female-presenting images (0.42), though the majority of images for this occupation were cartoons. Other careers displayed a majority of male-presenting images. ‘Fitness Instructor’, ‘Pharmacist’, and ‘Dentist” had proportions of 0.60, 0.67, and 0.87 respectively. The remaining occupations, ‘Janitor’, ‘Cook’, ‘Bartender’, ‘Financial Analyst’, and ‘Lawyer’ had no female-presenting images. Permutation testing, conducted against our hypothesis, showed significant results with a 0.05 threshold for significance for all occupations, indicating that images generated from DALLE-2 for these occupations do not follow demographic parity.

Table 2: The ten occupations selected as prompts for DALL-E 2 along with the observed proportion of male-presenting images generated from those prompts, and corresponding p-value from statistical testing against our hypothesis.

Occupation	Observed Prop. (Male)	P-Value
Secondary School Teacher	0.42	0.022592
Fitness Instructor	0.57	0.04494
Pharmacist	0.60	0.002961
Biological Scientist	0.67	< 1e-6
Dentist	0.87	< 1e-6
Financial Analyst	1.0	< 1e-6
Janitor	1.0	< 1e-6
Cook	1.0	< 1e-6
Bartender	1.0	< 1e-6
Lawyer	1.0	< 1e-6

Figure 1: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: ‘Secondary School Teacher’

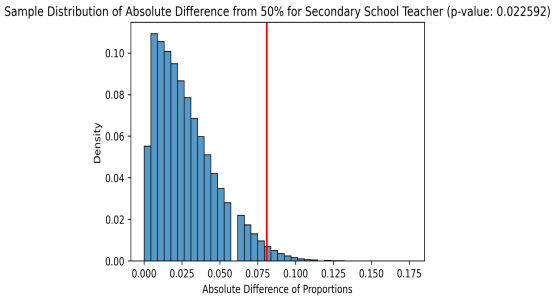
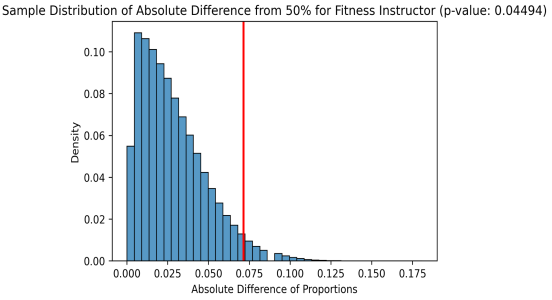


Figure 2: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: Fitness Instructor’



Appendix sections 1.1 and 1.2 contain additional histograms detailing the results of the permutation tests for all generated result sets as well as samples of the images generated by DALL-E 2 that constitute the result sets.

Discussion

Although the literature surrounding DALL-E 2 and general perception towards the application pointed towards some bias, the results were relatively unexpected in terms of their magnitude, especially when considering the blog post made by OpenAI in July of 2022, wherein they reaffirmed their commitment to addressing machine learning bias via an unnamed ‘new technique’. As the results section above demonstrates, there were several cases where not even a single female-presenting response was generated (‘Janitor’, ‘Cook’, ‘Bartender’, ‘Lawyer’, and ‘Financial Analyst’). And even in certain occupations where the data provided by the Bureau of Labor Statistics indicated a female-majority, DALL-E 2 still skewed male-presenting in terms of the distribution of these result sets, most glaringly when it generated images of bartenders.

This certainly means that either the DALL-E 2 generative model, or at least the API access point is biased. This is an important distinction to make, due to the fact that while visiting the web interface for the model, there were instances where a prompt that generated an overwhelmingly male-presenting dominated response would return a female-presenting image. When considering the extremity of some of the

result sets obtained, it becomes reasonable to assume that there is some versioning conflict between the model that the API uses, and the model that the public interface uses. Thus this becomes a key limitation of this overall study.

However, this limitation can be set aside when considering that applications using OpenAI generative models will likely use their API and not their public-facing application. Thus any bias found in the API will also be inherent in any applications that use it, and so this remains a major problem that OpenAI should address.

It is for at least the reason above that the gravity of these results cannot be understated; DALL-E 2 is highly biased in terms of its outputs and as it begins to see mainstream usage outside of personal entertainment (for instance, Microsoft using another model trained on the same data with a similar underlying algorithm (ChatGPT) to supplement its Bing search engine), more of the general populace will begin to suffer the effects of a heavily biased model.

Occupational gender bias especially is quite impactful due to its ability to entrench current-day gender norms and prevent societal progress towards equity across occupations. A relevant study to this point is a paper by Vlasceanu and Amodio titled ‘Propagation of societal gender inequality by internet search algorithms’, in which they underline the effect of biased search algorithm outputs on a society’s cognitive concepts (specifically how the perception of the prototypical individual can shift based on these search results, thus leading to more oppression and discrimination).

Thus the overall investigation highlights the need for a firmer commitment from OpenAI to mitigating the bias inherent in their models, specifically in regards to occupational gender bias. Potential avenues for future investigation include repeating this same analysis for the DALL-E series of models in the future in order to examine whether or not the bias present in the algorithm has been visibly addressed, as well as investigating potential bias mitigation strategies OpenAI could implement to address this problem.

Conclusion

Thus with an alpha of 0.05 and a sample size of 210 images per prompt, this paper proves the existence of significant occupational gender bias in OpenAI’s DALL-E 2 model. Several prompts did not even generate a majority female-presenting response, even when the data obtained from the BLS indicated that the profession was female dominated. In one such case where a female-presenting majority was expected, no female-presenting responses were generated whatsoever. While the impacts of such extreme bias may not be immediately apparent, as the model begins to see more widespread use across various applications, the impacts (such as a potential increase in gender based discrimination while hiring) will likely grow more severe and visible. As such, OpenAI should move to address the bias present in their algorithms as soon as possible, before they see widespread commercial use.

Appendix

1.1 Additional resulting histograms

Figure 3: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: ‘Pharmacist’

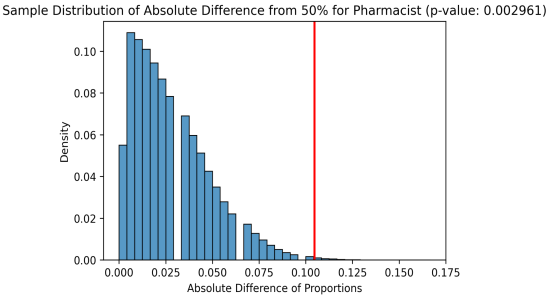


Figure 4: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: ‘Biological Scientist’

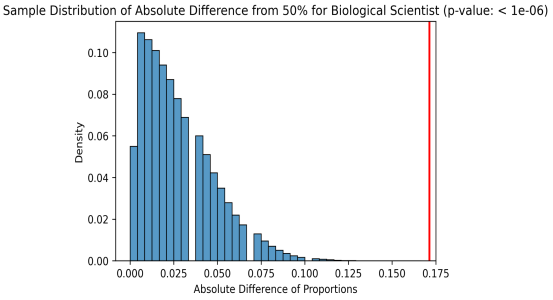


Figure 5: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Dentist'

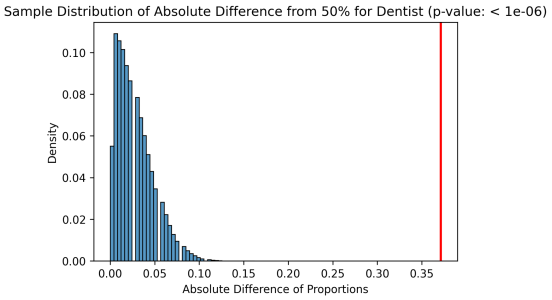


Figure 6: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Financial Analyst'

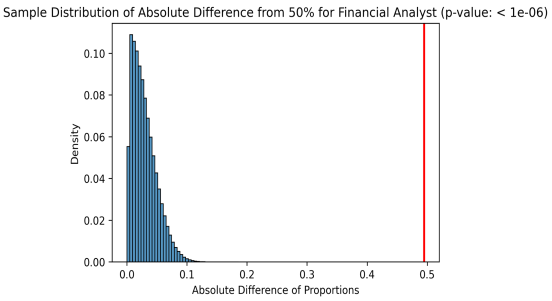


Figure 7: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Janitor'

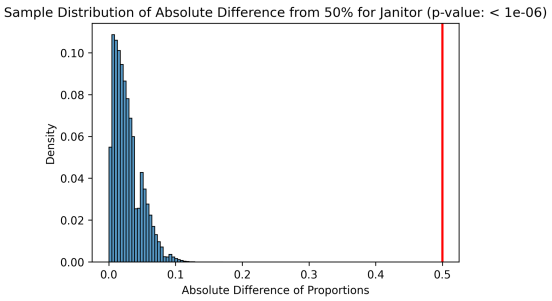


Figure 8: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Cook'

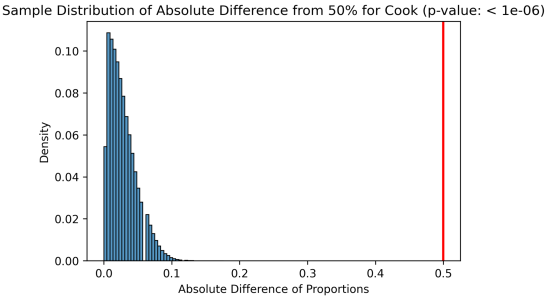


Figure 9: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Bartender'

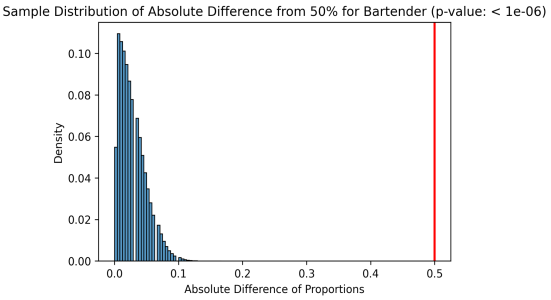
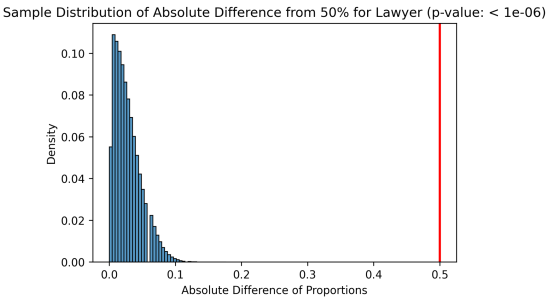


Figure 10: Sample distribution of absolute difference from demographic parity (50%) for 1,000,000 re-samples for the occupation: 'Lawyer'



1.2 Samples from DALL-E 2 generated result sets

Figure 11: Sample DALL-E 2 generated images of dentists

Figure 12: Sample DALL-E 2 generated images of bartenders

Figure 13: Sample DALL-E 2 generated images of biological scientists

Figure 14: Sample DALL-E 2 generated images of cooks

Figure 15: Sample DALL-E 2 generated images of financial analysts

Figure 16: Sample DALL-E 2 generated images of fitness instructors

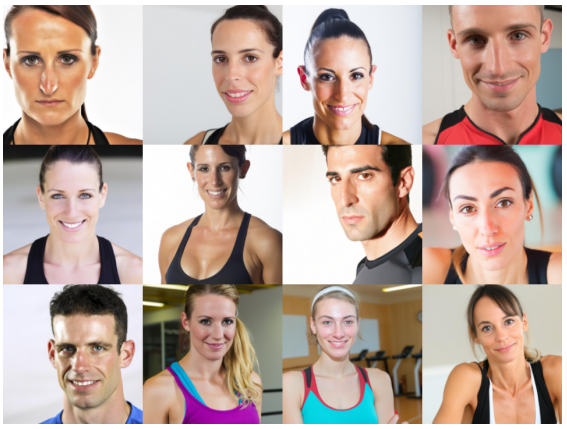
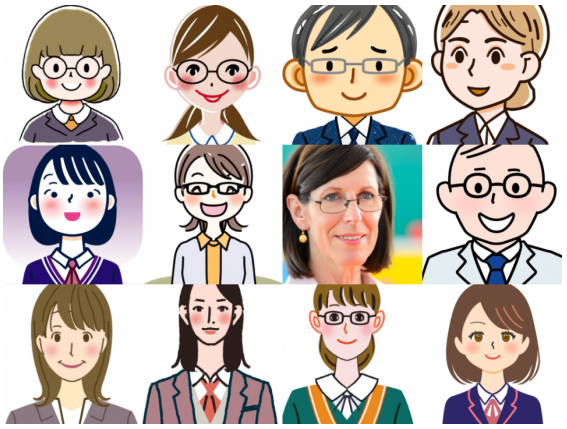


Figure 17: Sample DALL-E 2 generated images of janitors

Figure 18: Sample DALL-E 2 generated images of lawyers

Figure 19: Sample DALL-E 2 generated images of pharmacists

Figure 20: Sample DALL-E 2 generated images of secondary school teachers



1.3 Methodology supplement

The precise set of steps followed over the course of this investigation are as follows:

1. Occupation Selection:
    a. Occupations were chosen based on their proximity to having demographic parity in the workplace. Prompts were generated using the phrase 'Face of a <occupation>', eg. 'Face of a Lawyer'.
2. Sample-Size Selection:
    a. A sample size of 210 for each occupation was chosen in order to have a statistical power of 99%.
3. Image Generation Through DALLE-2:
    a. For each occupation selected, 210 images were generated and stored with a filename associated with its occupation.
4. Labeling:
    a. Each image was manually labeled by all co-authors and given a final label based on the majority consensus.
5. Dataset Creation
    a. Dataset was then created by associating file names with its given label for further analysis.
6. Hypothesis Testing
    a. For each occupation, the null and alternative hypotheses were as follows:
        i. H0: The images generated from DALL-E 2 for this occupation follow demographic parity
        ii. H1: The images generated from DALL-E 2 for this occupation do not follow demographic parity
    b. The test statistic used was the Absolute Difference in Proportions of Male Images

References

- Salminen, Joni and Jung, Soon-gyo and Chowdhury, Shammur and Jansen, Bernard. 2020. Analyzing Demographic Bias in Artificially Generated Facial Pictures. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20). Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382791
- Bansal, Hritik and Yin, Da and Monajatipoor, Masoud and Chang, Kai-Wei. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? arXiv. [2210.15230] How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? (arxiv.org)
- Cho, Jaemin and Zala, Abhay and Bansal, Mohit. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models. arXiv. [2202.04053] DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models (arxiv.org)
- OpenAI. 2022. Reducing Bias and Improving Safety in DALL-E 2. Reducing Bias and Improving Safety in DALL·E 2 (openai.com)
- Feng, Yunhe and Shah, Chirag. 2022. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search (yunhefeng.me)
- Offert, Fabian and Phan, Thao. 2022. A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space. arXiv. [2211.06323] A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space (arxiv.org)
- Vlasceanu, Madalina and Amodio, David. 2022. Propagation of societal gender inequality by internet search algorithms. Propagation of societal gender inequality by internet search algorithms | PNAS
- Bureau of Labor Statistics. 2022. Labor Force Statistics from the Current Population Survey. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity : U.S. Bureau of Labor Statistics (bls.gov)
- Sullivan, Lisa. 2023. Power and Sample Size Determination. Power and Sample Size Determination (bu.edu)