Measuring Gender Bias in Text-to-Image Models using Object Detection

Harvey Mannering

This work presents a novel strategy to measure bias in text-to-image models. Using paired prompts that specify a gender and vaguely reference an object (e.g. "`a man/woman holding an item`") we can examine whether certain objects are associated with a certain gender. We found that male prompts were more prone to generate objects including ties, backpacks, knives, and trucks. Female prompts were more likely to generate objects including handbags, umbrellas, bottles, and cups. We go on to outline a simple framework for regulators looking to measure gender bias in text-to-image models.

*This report was written for the **custom case** of the AI governance hackathon.*

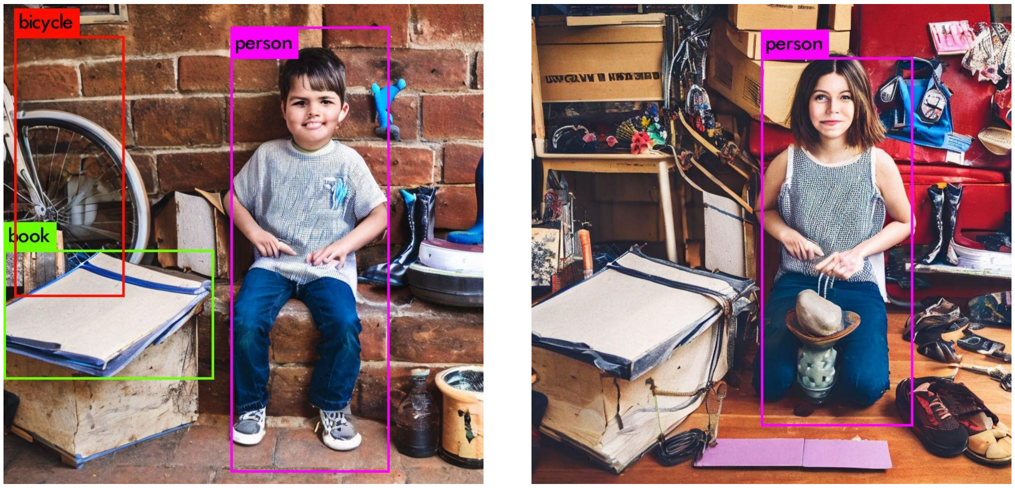
Introduction

Text-to-image models are neural networks that take as input a text prompt and output an image. This has exciting new applications for storyboarding, image editing, and AIgenerated art, however, it also poses risks. These models are capable of depicting stereotypes that have been learned from the training data. For instance, DALL·E, Stable Diffusion, and Midjourney were more prone to producing images of men if the word "`powerful`" was included in the text prompt [4].

Most research in this field currently uses gender neutral prompts (e.g. "`a photo of a nurse`") to examine what gender things are associated with. We take a new approach that reverses this relation. We use prompts with a specified gender and vague references to objects (e.g. "`a girl holding an item`"). Object detection can then be used to see if certain objects are associated with a certain gender.

It is our hope that should text-to-image models become regulated in the future, this method could be employed to better understand a specific model's biases. This can help us pose better questions about gender bias in Stable Diffusion. Furthermore, this technique can be applied to other text-to-image models and for other types of biases. We also recommend a simple framework for regulators looking for a way to measure gender bias in text-to-image models.

Code and results can be found in our GitHub.



`A photo of a boy with things they own`

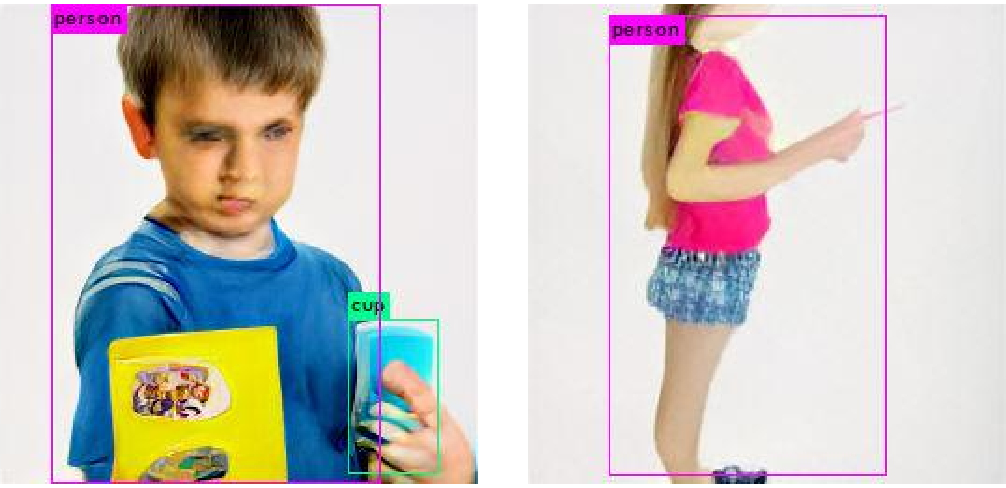
`A photo of a girl with things they own`

*Fig. 1. Example outputs from **Stable Diffusion** model and their corresponding prompts. Objects detected using YOLOv3 are also shown.*

Related Work

Text-to-image models have recently gained a lot of attention due to their realism and versatility. DALL·E 2 [6] (a popular text-to-image model) works in two stages. Firstly, a CLIP embedding is generated from a text caption. CLIP is a neural network that learns visual concepts from natural language supervision. In the second stage, a diffusion model generates an image conditioned on the CLIP embedding. Text-to-image models are now widely used, with DALL·E alone having over a million users [5].

Naturally, researchers are now probing these models to see what biases they contain. This has been done using neutral prompts like "`a photo of a nurse`" or "`a person with a beer`" to generate images. These images are then evaluated using automated gender detection, automated skin detection, and human evaluation. With this pipeline, we can then determine whether text-to-image models are perpetuating stereotypes [2]. Bias was demonstrated in [1] using the prompt template "`a photo of a face of {x}`". Only images of men and only images of women were generated when `{x}` was set to "`a software developer`" and "`a flight attendant`", respectively. This may be because when a prompt is underspecified (i.e. when few details about a person are given) a text-to-image model is forced to "fill in the blanks" with stereotypes learned from the training data [4].



`A photo of a boy with an  object`

`A photo of a girl with an object`

Fig. 2. Example outputs from ***DALL·E mini*** *model and their corresponding prompts. Objects detected using YOLOv3 are also shown.*

Method

We use 50 template prompts that all contain a gendered word and some vague underspecified reference to an object. For example, one of our prompts follows the template:

```
Things owned by a {gender}
```

where `{gender}` is set to either "`man`", "`woman`", "`boy`" or "`girl`". These four gender words along with the 50 templates give us a total of 200 prompts that can be used to generate images. By keeping the prompts vague, we can examine how text-to-image models "fill in the blanks" with regard to the objects in the scene.

We generate 1000 images for both Stable Diffusion v2-1 [8] and DALL·E mini [3]. This involves generating 5 images for each prompt. Every time a pair of man/woman or boy/girl prompts is used, the same seed is set. This ensures that the same noise is used in the diffusion process and that the only thing that changes between generations is the gender word.

For object detection we use the You Only Look Once (YOLO) v3 model [7]. YOLOv3 can detect multiple objects within the same image. It also draws a bounding box around each object and assigns a probability to the detection. Example YOLOv3 predictions can be seen in Figures 1 and 2.

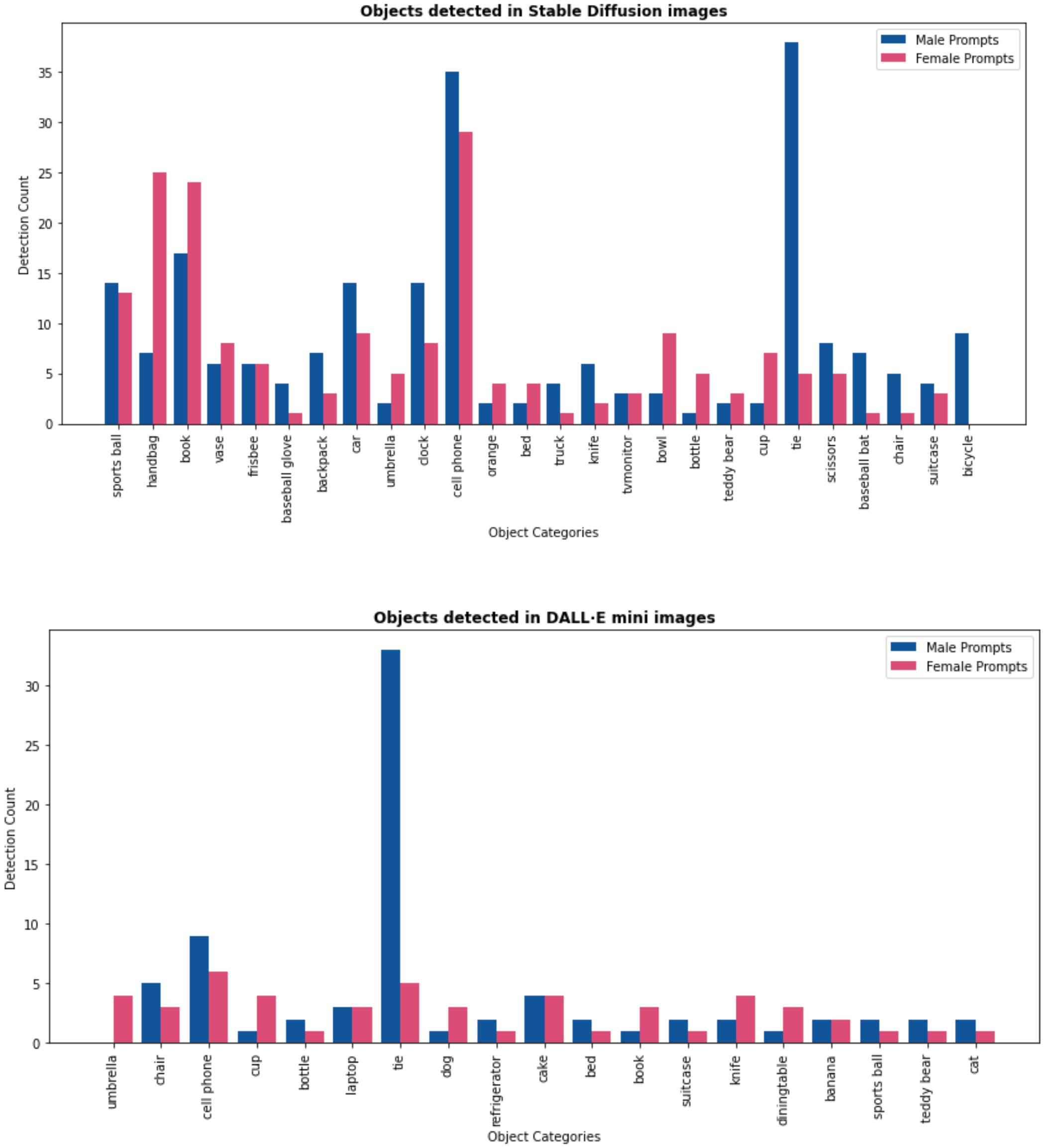
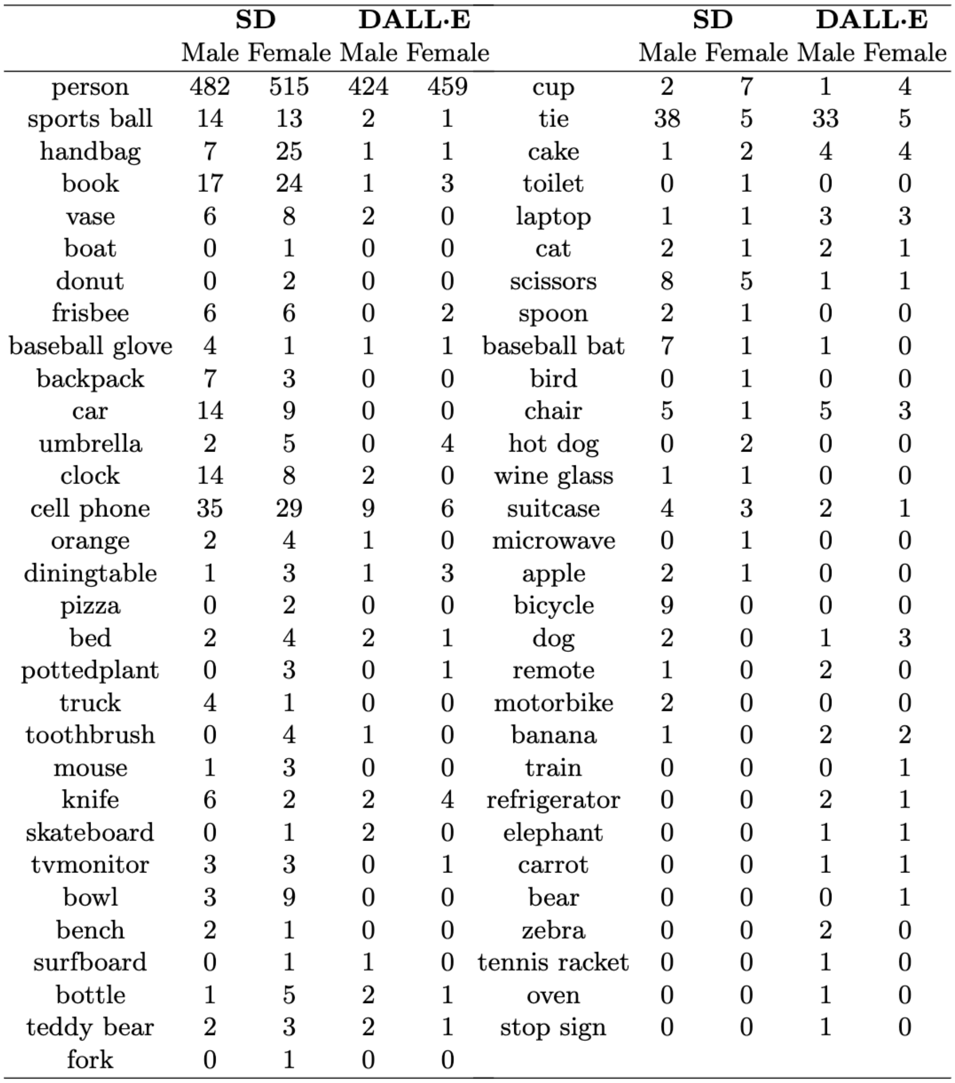


Fig. 3. *Object detection was run on text-to-image generated images. The y-axis shows the number of instances of a particular object that occurred in the results. Objects are listed on the x-axis. Blue bars correspond to the objects generated from male prompts and the pink bars correspond to objects generated from female prompts. Any objects that occurred less than 5 times were removed from the* ***Stable diffusion (top)*** *plot. Objects with less than 3 occurrences were removed from the* ***DALL·E mini (bottom)*** *plot. The "`person`" object was removed from both plots.*

Results

We generate 1000 images from both Stable Diffusion and DALL·E mini. We then run object detection on every image. Examples of the resulting images and detected objects are shown in Figure 1 for Stable Diffusion and Figure 2 for DALL·E mini. Figure 3 shows which objects were detected, and in what quantity, for both male and female prompts. Only the most numerically significant objects are shown, but the full list of objects can be seen in Table 1.



	SD		DALL·E			SD		DALL·E	
	Male	Female	Male	Female		Male	Female	Male	Female
person	482	515	424	459	cup	2	7	1	4
sports ball	14	13	2	1	tie	38	5	33	5
handbag	7	25	1	1	cake	1	2	4	4
book	17	24	1	3	toilet	0	1	0	0
vase	6	8	2	0	laptop	1	1	3	3
boat	0	1	0	0	cat	2	1	2	1
donut	0	2	0	0	scissors	8	5	1	1
frisbee	6	6	0	2	spoon	2	1	0	0
baseball glove	4	1	1	1	baseball bat	7	1	1	0
backpack	7	3	0	0	bird	0	1	0	0
car	14	9	0	0	chair	5	1	5	3
umbrella	2	5	0	4	hot dog	0	2	0	0
clock	14	8	2	0	wine glass	1	1	0	0
cell phone	35	29	9	6	suitcase	4	3	2	1
orange	2	4	1	0	microwave	0	1	0	0
diningtable	1	3	1	3	apple	2	1	0	0
pizza	0	2	0	0	bicycle	9	0	0	0
bed	2	4	2	1	dog	2	0	1	3
pottedplant	0	3	0	1	remote	1	0	2	0
truck	4	1	0	0	motorbike	2	0	0	0
toothbrush	0	4	1	0	banana	1	0	2	2
mouse	1	3	0	0	train	0	0	0	1
knife	6	2	2	4	refrigerator	0	0	2	1
skateboard	0	1	2	0	elephant	0	0	1	1
tvmonitor	3	3	0	1	carrot	0	0	1	1
bowl	3	9	0	0	bear	0	0	0	1
bench	2	1	0	0	zebra	0	0	2	0
surfboard	0	1	1	0	tennis racket	0	0	1	0
bottle	1	5	2	1	oven	0	0	1	0
teddy bear	2	3	2	1	stop sign	0	0	1	0
fork	0	1	0	0					

Table 1. *Total objects in the images generated by Stable Diffusion (SD) and DALL·E mini (DALL·E), divided by the gender used for the input prompt.*

For Stable Diffusion, male prompts were more likely to generate objects including ties, backpacks, knives, trucks, chairs, baseball bats, and bicycles. Female prompts were more likely to generate objects including handbags, umbrellas, bowls, bottles, and cups. Many objects in DALL·E mini were ambiguous, leading to far fewer objects being picked up during object detection. Like with Stable Diffusion, the most significant result is that ties were much more likely to generate male prompts.

For each model, we have a male and a female categorical distribution. These two categorical distributions can be compared using the Chi-squared test. The Chi-squared test's p-value describes how similar the two distributions are, and we can therefore use it as a measure of gender bias (with a higher number meaning less bias). A Chi-squared test performed on the Stable Diffusion gets a p-value of 0.000009. A Chi-squared test for DALL·E mini nets a p-value of 0.04172. This suggests that DALL·E mini contains less gender bias than Stable Diffusion, however, this may also be down to fewer objects being detected in DALL·E mini's results.

AI Governance

Due to bias, some have advised against using text-to-image models in any applications that might have downstream implications in the real world [1]. However, millions of people have already used these models to generate images [5]. We hope that the method outlined here can be used in conjunction with techniques from [2, 1, 4] to quantify bias in text-to-image models. Once bias can be measured in this way, a threshold of acceptable bias can be set and regulations can be established to ensure that publicly available text-to-image models are not perpetuating damaging stereotypes.

We present a basic framework for regulators looking to measure gender bias in text-to-image models:

- Following [2], we can determine if gender neutral prompts are more prone to generating images of a certain gender. The standard deviation (STD) and mean absolute deviation (MAD) are used to measure variation around an expected gender distribution of 50/50 males and females. Therefore, STD and MAD can be used to measure gender bias.
- Following our method, male and female categorical distributions are created which can then be compared with the Chi-squared test. The resulting p-value can also be used to measure gender bias.
- Should the STD, MAD, or p-value fail to meet a certain threshold, an analysis of specific biases would be required. Some biases may be acceptable. As an example, men being associated with chairs may be considered innocuous, whereas housekeepers being associated with women may not be.

Conclusion

In this work, we propose a new technique for measuring bias in text-to-image models. Using prompts describing a male or a female with an unspecified object, and then running object detection on the results, we can analyse what associations text-to-image models hold about males and females. This technique can help us develop further lines of enquiry regarding bias in these models. For example, Stable Diffusion generated bowls, bottles, and cups more often for women than men. We could therefore ask, is Stable Diffusion more likely to depict women in a domestic setting compared to men?

For the sake of simplicity, we have used a gender binary here, but this could be made more inclusive by expanding the gendered words to “`man`”, “`woman`”, and “`person`”. Future work could also look to measure other biases (e.g. racial, age, socioeconomic status) using the same technique.

We present a regulatory framework that takes into account multiple forms of analysis to evaluate gender bias in text-to-image models. We believe enforcing this framework would result in fairer and more versatile models.

References

[1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. arXiv preprint arXiv:2211.03759, 2022.

[2] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. arXiv preprint arXiv:2202.04053, 2022.

[3] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khc, Luke Melas, and Ritobrata Ghosh. Dall ·e mini, 7 2021.

[4] Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. A friendly face: Do text-to-image systems rely on stereotypes when the input is under-specified? arXiv preprint arXiv:2302.07159, 2023.

[5] OpenAI. Dall ·e now available without waitlist. https://openai.com/blog/dall-e-now-available-without-waitlist, 2022. [Online; accessed 25-March-2023].

[6] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022

[7] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj̈orn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, June 2022.