Investigation of Visual Bias in Generative AI – Article

With the increasing sophistication and widespread use of image generation techniques along with overall acceptance of their use in daily life, these is a pressing demand for fair and non-biased systems. This increased usage of such systems along with the uncertainty regarding their bias served as a main motivator for this study which delved into the possible forms of bias present in popular image generative systems in particular Stable Diffusion, Dall-E and Midjourney.

These types of models function by converting simple text prompts into images matching the description of the provided prompts. In line with this the selected approach was to initially generate various images of a set of predetermined prompts, analyse said images in terms of the gender, age, race and prominence of the person or persons depicted, and utilising various metrics conclude on the presence and severity of the detected bias.

This process was achieved by utilising the aforementioned models alongside the following set of prompts; ***A picture of (a doctor / a nurse / a doctor and a nurse) facing forward*** for which **3465** images were generated spanning across all prompts and models, a sample of these image is depicted in Figure 1. This was done to provide a sufficiently accurate bases with which to support the conclusions made. Furthermore, the images were then pre-processed extracting the individual persons from the images through the **YOLOv8** person detection model. This procedure was crucial in facilitating accurate image annotation whilst providing us with the required data to determine an individual’s prominence in an image, in line with the implementation presented in the REVISE research paper [1]. These images were then passed through the **MTCNN** face detector which extracts and realigns individual faces thereby increasing the accuracy of correct annotations. The annotation process itself was carried out through the use of the **FairFace** and **DeepFace** models which provided **gender**, **race** and **age** predictions. Finally, the resultant image attributes were processed to extract a variety of metrics which resulted in a conclusion on the presence of bias in generative models.

Furthermore, to obtain a full picture of the bias present in said models’ images from the **LAION-400M** dataset also underwent a similar process. Here images of doctors and nurses were extracted and processed in the same manner as outlined prior, save for the introduction of human-annotation. This procedure was carried out using the facilities provided via Google form and saw the annotation of a total of **194** imagesspanning across doctor and nurse. The purpose of this process was to outline the innate bias present in the annotation models used (FairFace / DeepFace) whilst also exposing the innate bias present in the training data of such models. This is as the **LAION-400M** dataset consists of a subset of the data used to train the Stable Diffusion model and was the only publicly available dataset associated with the mentioned generative models.

In conclusion this study emphasises the need for non-biased generative models whilst exposing the bias present in some of the popular publicly available generative models, whilst producing a simple tool with which other similar models can be assessed.

[1] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In European Conference on Computer Vision (ECCV), 2020.



Figure 1 Generated Images of Doctor, Nurse, and Doctor & Nurse via Midjourney, Dall-E and Stable Diffusion.