Investigation of Visual Bias in Generative AI – Article

My Final Year Project titled Investigation of visual bias in generative AI seeks to determine if some of the popular generative models, in particular Stable Diffusion, Dall-E and Midjourney perpetuate bias within the images that they generate. These generative models work by taking a text prompt such as “A picture of a doctor” and converting that prompt into an image such as those seen in Figure 1.

Continuing with the doctor example, these models function by first learning what an image of a doctor looks like. This is done simply by showing the model a large quantity of doctor images from which it learns how to create its own doctor images. In line with this the original hypothesis for this research assumed that the images used to train such models are in fact biased particularly in terms of gender and thus, the model would replicate this same bias.

In line with said hypothesis a solution was devised; this took the form of a series of steps starting with the generation of various images. Images of doctors and nurses were generated as the doctor profession tends to be associated with male individuals whilst the nurse profession with female individuals. Through studying the bias resultant from the generation of such images one could discern whether said bias is being portrayed within the generated images, determine any anti-bias measures implemented by the model and detect other forms of bias.

Image generation was achieved by using the following set of prompts ***A picture of (a doctor / a nurse / a doctor and a nurse) facing forward*** to generate **385** images for each label using each of the three generative models mentioned prior. Additionally relevant images from the **LAION-400M** dataset were also considered as this dataset along with many others are commonly used to train generative models, however this was one of the few datasets which is publicly accessible. The images retrieved, generated or otherwise where then processed to detect the **gender, race and age** of the individuals depicted. This was facilitated through the **YOLOv8** person detection model which was used to extract the individual people in an image, in cases where images contained more than one person. Furthermore, these images were then fed to an **MTCNN** face detector which was used to crop out the individual faces of said peoplewhich were then processed via two models these being **FairFace** and **DeepFace** which provided their predictions for an individual’s **gender**, **race** and **age**. A portion of the images retrieved from the **LAION-400M** dataset were human annotated which then provided us further insight into the accuracy of the annotations carried out via **FairFace** and **DeepFace.** 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.



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