**VIVA -Script**

**Slide 1**

Good morning, my name is Jerome Agius and today I will be showcasing my final year project, titled *Investigation of Visual Bias in Generative AI.*

**Slide 2**

This research aimed to explore bias within generative models, inspired by previous studies which highlighted its negative impact within AI systems, such as those used in recidivism and credit scoring. An example is illustrated in Figure 1, where a biased assessment labelled a dark-skinned individual as high risk, despite not reoffending. Research on this topic is vast however the recent prominence of generative models has produced a gap which this thesis aims to fill. The initial focus was on the LAION-5B and Stable Diffusion models due to them being open-source. However, due to controversies associated with the former the decision was made to investigate the LAION-400M dataset alongside the Stable Diffusion, Dall-E & Midjourney models. This was done as LAION-400M is simply a downscaled version of the LAION-5B dataset, furthermore the additional models were included to expand the scope of the research and facilitate model comparison.

The idea behind this research project was to investigate the severity and impact of bias within generative models. This idea stemmed from research which exposed the negative effects of biased AI systems throughout our daily lives. For instance, within recidivism and credit scoring systems. An Instance of the impact of said bias can be seen in Figure 1 wherein individuals accused of petty theft were assessed via said system, with the white individual labelled as low risk and the dark skinned individual as high risk, however the latter did not reoffend. Given that such systems have already been studied for bias and its impact it was seen as appropriate to tackle a different sector in particular generative AI and asses the bias therein whilst getting ahead of the curve prior to said models becoming even more widely used and easily accessible. Initial research on the topic led to the discovery of the Stable Diffusion model and its training dataset LAION-5B. These were going to serve as the basis for the research due to their open-source nature however, around last December this was no longer plausible seeing as the dataset was no longer made available due to controversy revolving around its images. This resulted in a shift from the LAION-5B dataset to its predecessor the LAION-400M dataset which was simply a downscaled version of the former and as such could affectively take its place. Furthermore, the choice was made to include the Dall-E and Midjourney models so as to provide comparative results between models.

**Slide 3**

Moving forward with that aim in mind four main objective needed to be tackled. Firstly, I needed to determine what forms of biases are most prevalent and by extension should be assess. Secondly, I needed a way to measure said biases to arrive at an informed conclusion, Thirdly, I needed to determine what image to generate and finally how said image are going to be annotated.

**Slide 4**

The research emphasises the significance of gender, racial, and age biases in AI systems, additionally prominence bias which evaluates a person's visibility within an image was deemed worthwhile as it could provide further insight into generative model bias. Furthermore, various metrics including label counts, correlation, Shannon Entropy, Simpson index, and evenness measures were utilised to quantify these biases. Prominence was to be assessed based on the person's distance from the image centre and their percentage area within the image. To illustrate bias, images of doctors and nurses were chosen due to their known demographic skew as is shown in Figure 2, although a broader range of professions could have provided richer insights. However, resource constraints limited the scope of the study. Image annotation was to be carried out via computer-assisted means due to the large volume of images which required annotation however a human-annotated set was to be curated in order to identify potential biases inherent in the annotation models.

Further, research on the topic highlighted the importance of gender, racial and age bias within such systems. Additionally, prominence bias referring to how prominent a person is within an image, was also assessed as it could provide useful insight on how generative images are composed.

These biases were to be measured using, label counts, correlation, Shannon Entropy, Simpson index as well as their evenness measures. Additionally, prominence was to be measure via the distance between the person and image centre as well as the percentage area that the person occupies in relation to the entire image. Resulting in greater prominence if the person is larger and central and less prominence otherwise.

Once the biases and measures were determined it was deemed appropriate to assess images of doctors and nurses primarily due to their innate bias. With Doctors being predominantly young white males and nurses young white females as can be seen in Figure 2. A wider cast of profession could have been assessed in order to provide a better depiction of bias however due to monetary and time constraints this was not plausible.

These image were to be annotated via computer assisted techniques seeing as 385 image were required per model for each label in order to produce reliable results. However, human annotation was to be used on a reduced scale so as to serve as a baseline to identify innate annotation model bias.

**Slide 5**

In line with the research, a structured process pipeline was developed. This was divided into several stages. Consisting of Image Generation, Retrieval and Filtering, Image Annotation, Metric Extraction, concluding in a conclusion on bias.

**Slide 6**

Each generative model produced 385 images for the "Doctor" and "Nurse" professions, with an additional set combining both prompts aimed at uncovering other potential biases. All images originated from the prompt "A picture of a [subject] facing forward," with "Disfigured" and "Art" as negative prompts for realism. LAION-400M images were sourced by searching for iterations of the "doctor" or "nurse" keywords in image descriptions, as this dataset had no search functionality unlike LAION-5B. The sample size of 385 was chosen as it ensured a 95% confidence level with a 5% margin of error. Following image generation manual filtering was carried out, removing explicit images or those containing children. Furthermore, the images composing the 385 image sets were randomly selected from the filtered set. Notably, Dall-E was the only model to modify the prompt prior to image generation.

Image generation involved using each generative model to generate 385 images for the “Doctor” and “Nurse” professions. An additional set of images was also generated incorporating both professions into a singular prompt. This was done with the hope that it would provide greater insight into the models bias. The images generated were all resultant from the same prompt “A picture of a [subject] facing forward” with ”Disfigured” and ”Art” as negative prompts for more realistic depictions.

The LAION-400M images were retrieved simply by searching for different iterations of the word “doctor” or “nurse” within the associated image description seeing as this was the only label assigned to said images.

The sample size was used as it resulted in 95% confidence 5% margin of error. Furthermore, these images were selected at random following manual filtering where images of children and those of an explicit nature were removed.

A point of note during the image generation process is that although each model utilised the given prompt for its image generation, Dall-E consistently altered the provided prompt prior to image generation.

**Slide 7**

Here are some images generated via Stable Diffusion the first row is of doctors, second row of nurses and final row of both doctors and nurses.

**Slide 8**

Here are some images generated via Midjourney the first row is of doctors, second row of nurses and final row of both doctors and nurses.

**Slide 9**

Here are some images generated via Dall-E the first row is of doctors, second row of nurses and final row of both doctors and nurses.

**Slide 10**

The retrieved images underwent annotation using the DeepFace and FairFace models, selected for their high accuracy and ability to provide gender, race, and age annotations, unlike many other models focused solely on gender or age due to safety concerns. Additionally, an ensemble model combining DeepFace and FairFace was tested but showed only marginal improvement over DeepFace and performed worse than FairFace on a subset of the UTK-Face dataset. Prominence metrics, including centre distances and percentage area, were determined using the YOLO person detection model due to its widespread use, easy implementation, and reliable results. Comparison with human annotations on a subset of the LAION-400M dataset revealed both models' biases, with FairFace showing more predictable and consistent biases, thus given greater weight in annotations compared to DeepFace.

Moving onwards the images retrieved were annotated using the DeepFace and FairFace models. These were chosen as not only did they have high accuracy according to their papers but they also produced the required annotations these being gender, race and age where the majority of available models only detected gender or age due to safety issues. Furthermore, an ensemble model combining the two prior models was also tested however as can be seen in the image its results were only relatively better than DeepFace but worse than FairFace when tested on a subset of the UTK-Face dataset.

The centre distances and percentage area used for the prominence metric were determined via the YOLO person detection model due in part to its wide usage, easily implementation and reliable results.

Additionally, the two models were also compared to a human annotated subset of the LAION-400M dataset to determine their consensus with human annotators and although both model performed relatively well it was clear that both had their own biases however the FairFace’s bias was far more predictable and consistent and as such its annotations were given increased weight in comparison to that of DeepFace.

**Slide 11**

Once the images were annotated and the metrics extracted the aforementioned measures were calculated allowing us to devise a conclusion.

**Slide 12**

Conclusions were primarily drawn from label counts, with other measures supporting the findings. Gender analysis revealed a global trend of male doctors and female nurses. However, the LAION-400M dataset depicted a reversal, with mostly female doctors but aligned with global data for nurses. Given distinct counts between models, FairFace values were given precedence due to DeepFace's male bias. Stable Diffusion and Midjourney models showed similar gender biases, with nurses predominantly female and doctors predominantly male. In contrast, the Dall-E model depicted an inverse bias, deviating from real-world metrics by reducing male doctors and female nurses.

The conclusions were primarily derived from the label counts using the other measures to ensure sound conclusions. Starting off with gender it was noted that globally the majority of doctors were male where nurses were female.

In line with these results it was then concluded that the LAION-400M dataset presented a reversed depiction with the majority of doctors being female. However this was not the case for nurses wherein the depiction seemed to align with the global data. In cases such as this where the counts are quite distinct between the two models the FairFace values where given greater importance as the DeepFace model was severely biased in labelling images as male.

The Stable Diffusion model alongside the Midjourney model presented the same degree of gender bias having similar values for the doctor and nurse subsets with the latter being predominantly female and the former male.

Contrarily the Dall-E model presented an inversely biased depiction in comparison to real-world metrics, presenting a reduction in male doctors and female nurses.

**Slide 13**

Continuing with racial bias, it was observed that most American doctors and nurses were white, as global metrics were unavailable. The LAION-400M dataset exacerbated racial bias within the doctor subset, with reduced racial diversity but white remaining dominant. Conversely, nurse demographics showed a decrease in white depictions but still remained dominant. Stable Diffusion showed innate racial bias, with white doctors exceeding real-world metrics, while nurses had reduced white depictions but remained dominant. Similarly, Midjourney demonstrated reduced white depictions, but white remained dominant in both doctor and nurse depictions. In contrast, Dall-E depicted an inverse bias, with Indian and Asian races dominant instead of white, presenting a fairer distribution among all races.

Moving onwards with racial bias it was noted that most doctors and nurses in America were white. This data was derived from America as global metrics were not available.

In line with these results, it was concluded that the LAION-400M dataset exacerbates the racial bias within the doctor subset having reduced racial representation with white remaining the dominant race. Contrarily the nurse demographics depict a different story with white depictions being drastically reduced however remaining as the dominant race.

Observing the Stable Diffusion race demographics outlines innate racial bias with white doctor depictions exceeding real-world metrics. Contrarily this isnt the case for nurses as it presents a reduced depiction of white individuals. However, in both cases white remain the dominant race.

The same applies to Midjourney however the reduction in white depictions is even greater here.

Contrarily, Dall-E depicts bias inverse to real-world metrics with Indian and Asian serving as the dominant races as opposed to white, whilst presenting a fairer distribution amongst all races.

**Slide 14**

Concluding with the age bias it was noted that amongst doctors only 66% are younger than 55 with the percentage being drastically higher at 81.62% for nurses.

However, unlike previous biases age bias appeared consistently amongst all models and the LAION-400M dataset with the primary depictions consisting of individuals aged 20-29 or 30-39 with minimal to no depictions of elderly individuals.

**Slide 15**

The prominence metrics although included in the hopes of providing additional insight into the innate biases of the models failed to produce any worthwhile results with the majority of depictions across all gender, ages and races being equally prominent save for a few outliers within underrepresented races.

**Slide 16**

The research conducted within this research paper has room to be expanded further. Instance of this include the creation of an anti-biased model specifically designed to mitigate bias. Investigation of Dall-Es prompt rewriting model and its effects on biased model, which could lead to new manners in which model bias is mitigated. Finally, the creation of a non-biased dataset could prove useful in mitigating bias within future generative models.

**Slide 17**

Thank you for your time. I hope this presentation was informative!

Do you have any further questions?