INVESTIGATION OF VISUAL BIAS IN GENERATIVE AI

Jerome Agius – B.Sc. IT (Hons) Artificial Intelligence

MOTIVATION

- Bias affects the applicability of applications
- Affecting people's lives (Recidivism Scoring/Credit Scoring)
- Bias needs to be addressed in AI systems particularly in generative endeavours and visual datasets.
- Initial research goal (LAION-5B & Stable Diffusion)
- Expanded research goal (LAION-400M & Stable Diffusion/Dall-E/Midjourney)

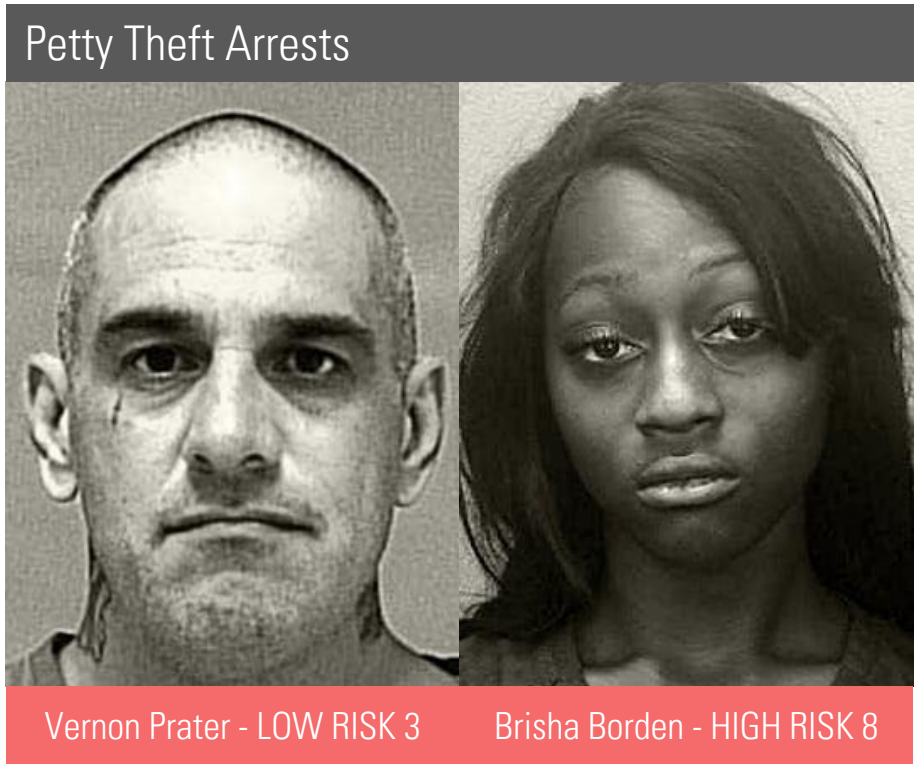


Figure 1: ProPublica Recidivism Scoring Example

Figure 1 - https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

AIMS AND OBJECTIVES

- Identify the severity of bias within the generative models and LAION-400M dataset.
- Objectives:
  - Determine the types of bias assess.
  - Determine what bias measurements to utilise.
  - What Images to assess?
  - How should the images be annotated?

RESEARCH

- Biases (Gender/Race/Age/Prominence)
- Measurements (Count/Correlation/Shannon Entropy/Simpson Index/Evenness/Centre Distances/Percentage Area)
- Innately biased images (Doctor/Nurse/Both)
  - Doctors – Male/White/Younger than 55
  - Nurse – Female/White/Younger than 55
- Computer Assisted Annotation (DeepFace/FairFace)
  - Biased Models -> Impact results.
  - Human Annotation -> Serves as baseline.

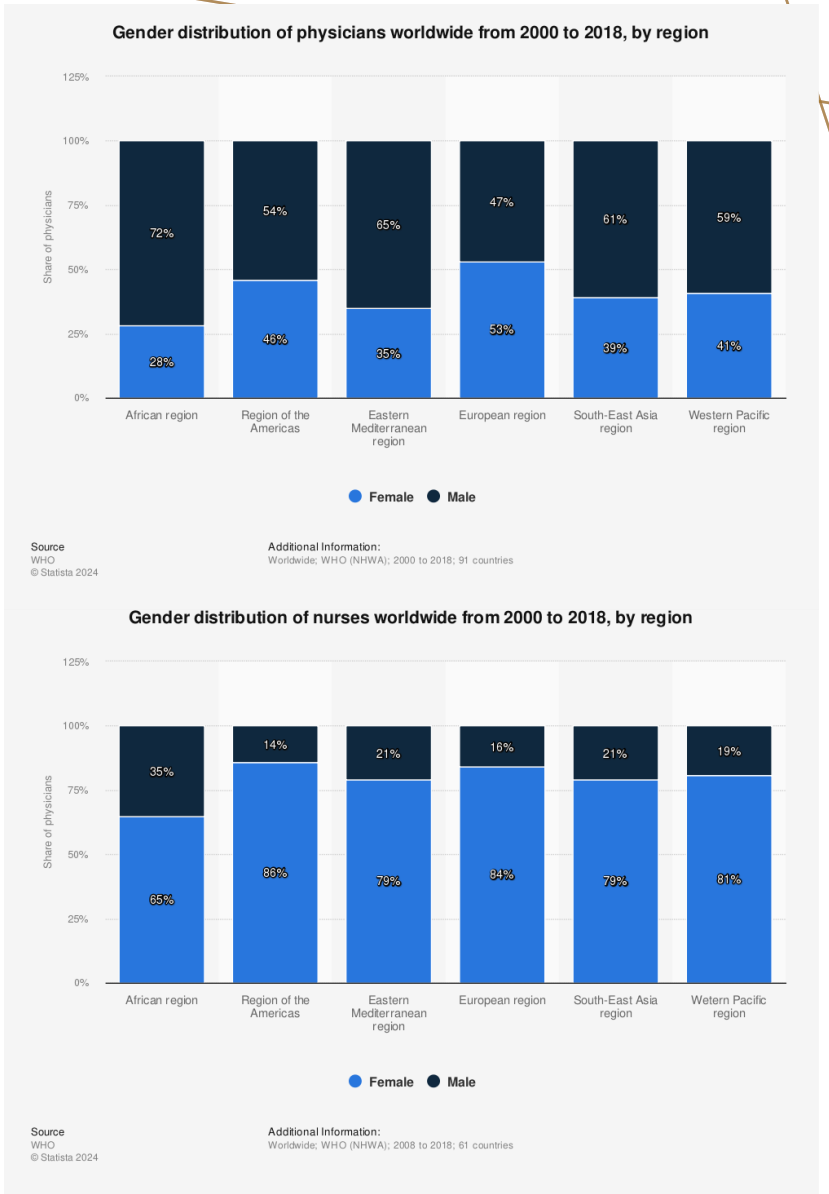


Figure 2: Statista Doctor/Nurse Gender Distribution

Figure 2 - https://www.statista.com/statistics/1099789/distribution-of-physicians-across-regions-worldwide-by-gender/ / https://www.statista.com/statistics/1099804/distribution-of-nurses-across-regions-worldwide-by-gender/

IMPLEMENTATION

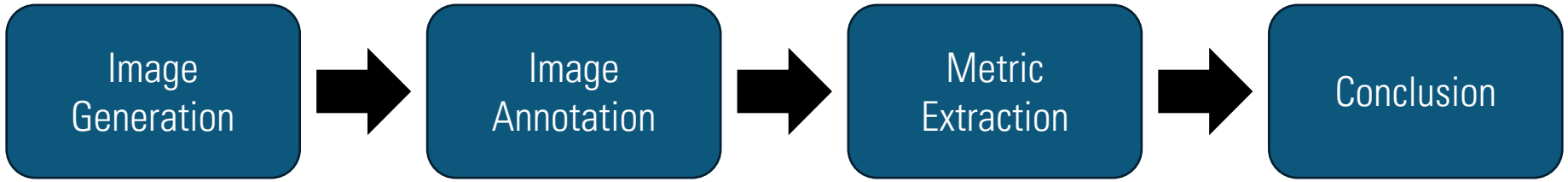


IMAGE GENERATION

- Prompt - "A picture of a [subject] facing forward"
- Negative Prompt - "Disfigured" and "Art"
- Prompt Alteration (Dall-E)
- No. of Images

IMAGE RETRIEVAL

- Tagged (Doctor/Nurse)
- Manual Filtering
- Random Selection
- No. of Images

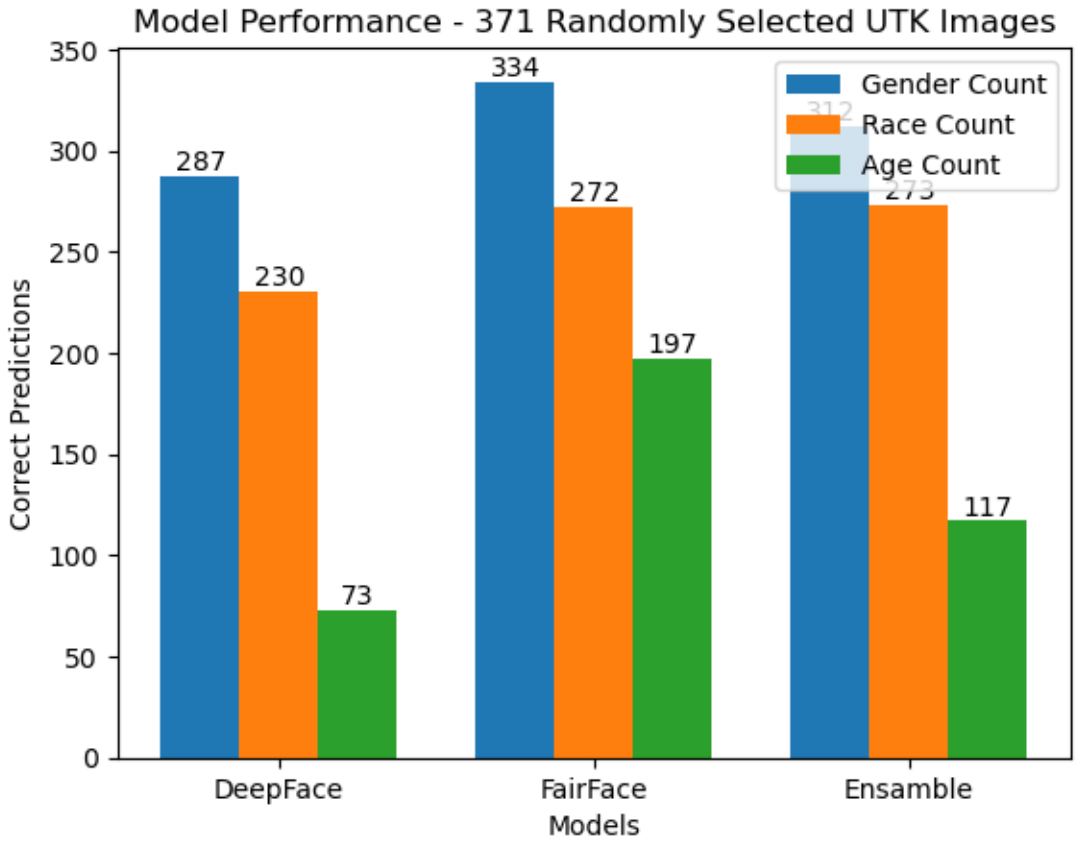
STABLE DIFFUSION IMAGES

MIDJOURNEY IMAGES

DALL-E IMAGES

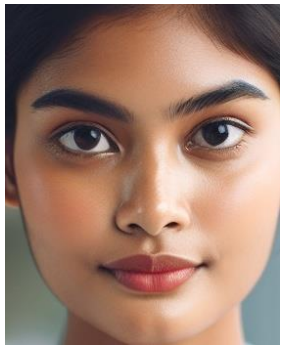
ANNOTATION

- UTK-Face Dataset
- Ensemble model abandoned
- Main model (FairFace) – Human Annotation predictable bias
- Labels
  - Gender – Male/Female
  - Race – Asian/White/Middle Eastern/Indian/Black/Latino Hispanic
  - Age – 0-2/3-9/10-19/20-20…60-69/70+



METRIC EXTRACTION

Image



Labels

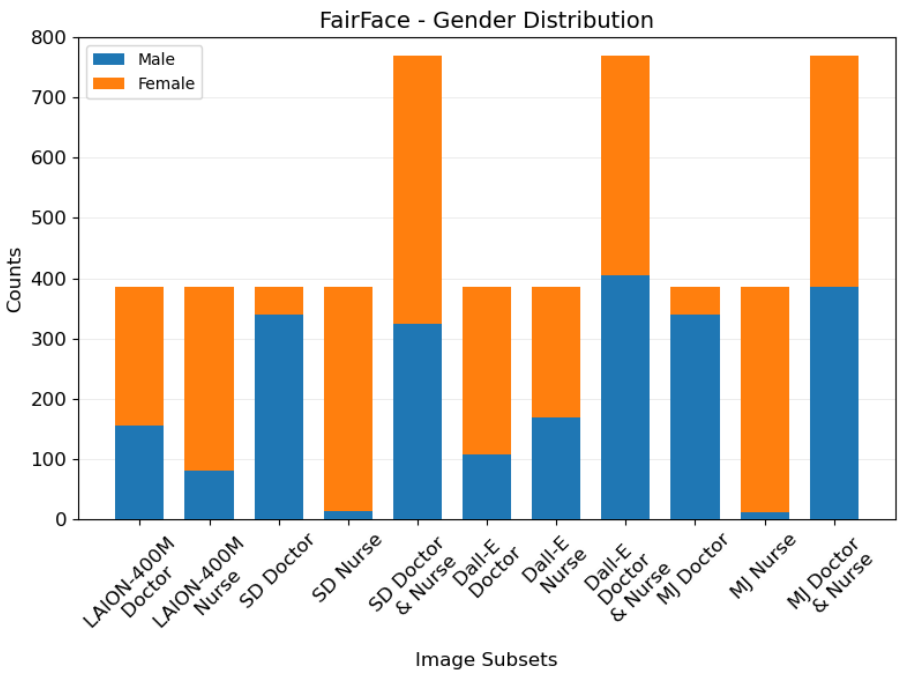
{‘age’: label, ‘gender’: label, ‘race’: label, ‘centre_dist’: label, ‘space’: label}

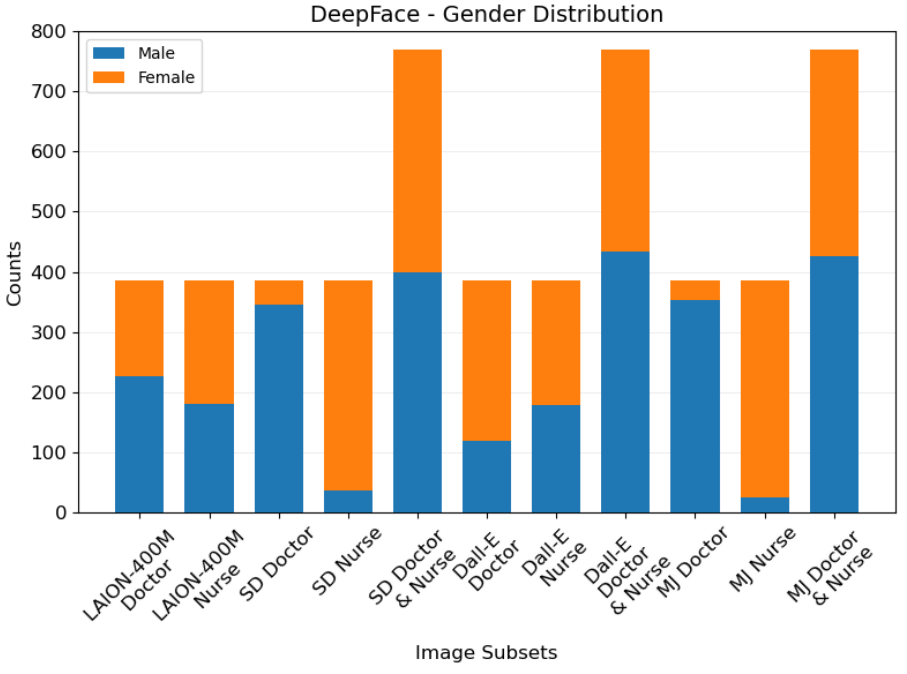
Confidence

{‘age’: confidence_dict, ‘gender’: confidence_dict, ‘race’: confidence_dict}

CONCLUSION - GENDER

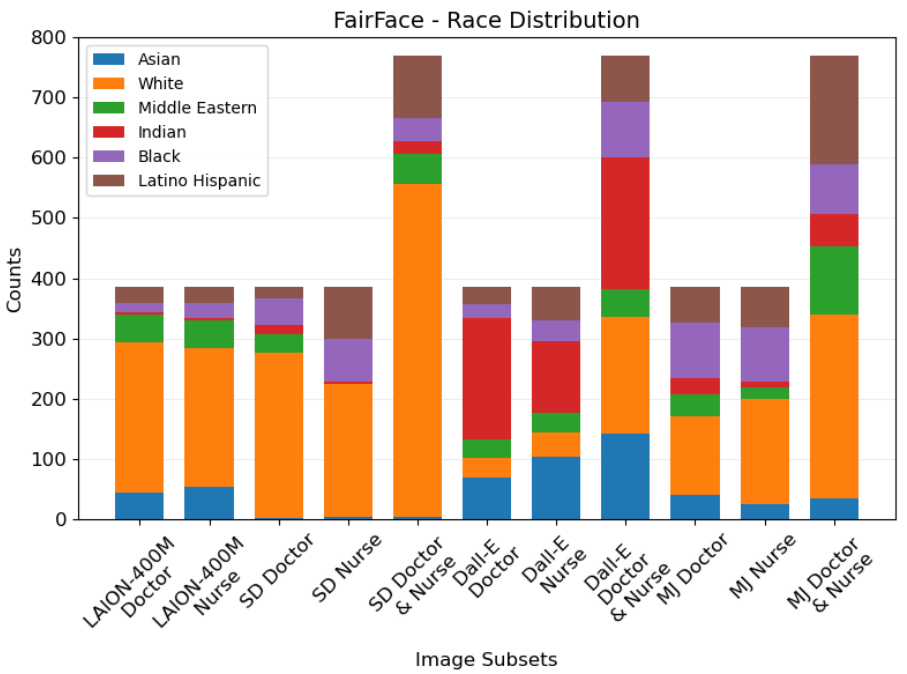
- Real-World
  - Doctor - >50% Male (Globally)
  - Nurse - 76.91% Female (Globally)
- LAION-400M
  - Doctor - 40.26% / 58.96% Male
  - Nurse – 79.22% / 53.25% Female
- Stable Diffusion
  - Doctor - 88.31% / 89.61% Male
  - Nurse - 96.62% / 90.65% Female
- Dall-E
  - Doctor - 27.79% / 30.91% Male
  - Nurse – 56.1% / 53.51% Female
- Midjourney
  - Doctor - 88.31% / 91.69% Male
  - Nurse - 97.14% / 93.77% Female

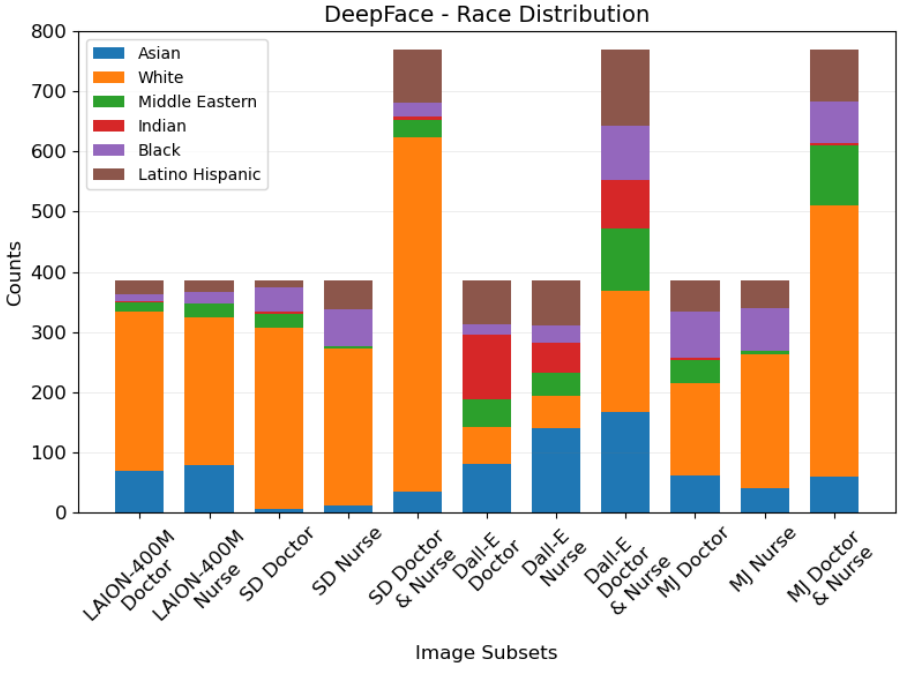




CONCLUSION - RACE

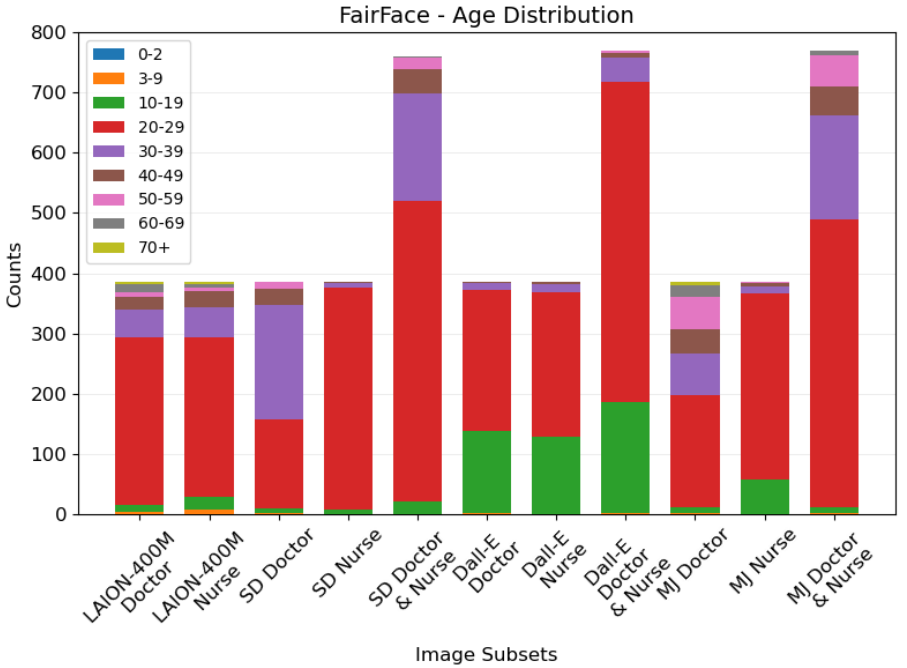
- Real World
  - Doctor - 56.2% White (America)
  - Nurse - 80% White (America)
- LAION-400M
  - Doctor - 64.94% / 69.09% White
  - Nurse - 59.74% / 64.16% White
- Stable Diffusion
  - Doctor - 71.17% / 78.18% White
  - Nurse – 57.4% / 68.05% White
- Dall-E
  - Doctor - 8.57% / 15.84% White (Fair distribution)
  - Nurse - 10.65% / 14.29% White (Fair distribution)
- Midjourney
  - Doctor - 34% / 40% White
  - Nurse – 45% / 57% White

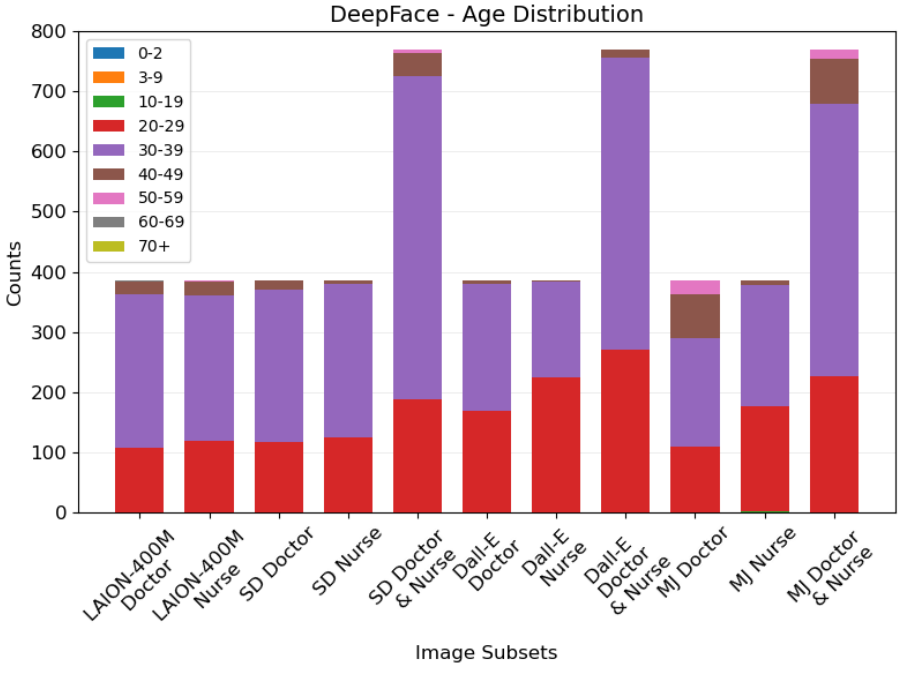




CONCLUSION - AGE

- Real World
  - Doctor - 66% younger than 55 (Globally)
  - Nurse - 81.62% younger than 55 (Globally)
- LAION-400M
  - Doctor - 93.77% / 99.74% younger than 55
  - Nurse - 96.36% / 99.74% younger than 55
- Stable Diffusion
  - Doctor - 96.88% / 100% younger than 55
  - Nurse - 100% / 100% younger than 55
- Dall-E
  - Doctor - 100% /100% younger than 55
  - Nurse – 100% / 100% younger than 55
- Midjourney
  - Doctor - 79.22% / 94.03% younger than 55
  - Nurse – 99.74% / 100% younger than 55



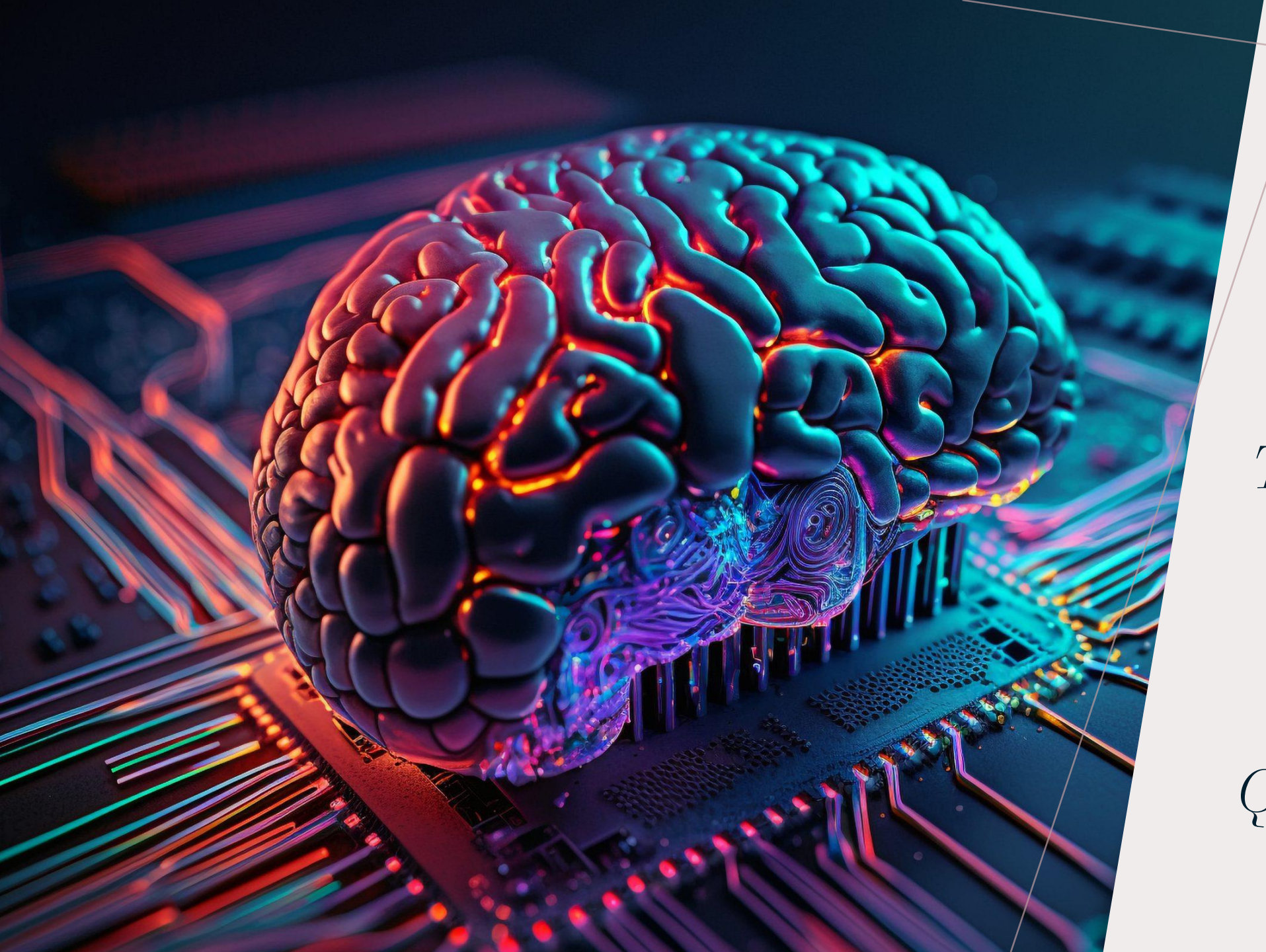


CONCLUSION - PROMINENCE

- Should I add a conclusion slide for prominence seeing as nothing insightful was derived from the measurements?

FUTURE WORK

- Generative model designed to mitigate biases
- Investigation of Dall-E prompt altering model and its effectiveness
- Creating unbiased training datasets



THANK YOU FOR YOUR TIME!

ANY QUESTIONS?