|  |
| --- |
| Investigation of Visual Bias in Generative AI |
| Jerome Agius |
| Supervisor: Dr Dylan Seychell |
| May 2024 |
| Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Information Technology (Hons) (Artificial Intelligence). |



Abstract (max 300 words)

In the ever-evolving world of Artificial Intelligence (AI), text-to-image generators, such as Stable Diffusion, Dall-E-3 and Midjourney revolutionise creativity, but raise concerns regarding bias in generated images, particularly those depicting people. Bias can also present itself in the training datasets used to build these models. This thesis investigated this issue by comparing and analysing the inherent bias within these models and popular training datasets.

The research approach revolved around the retrieval/generation of images coinciding with the terms *person, doctor,* and *nurse*. The latter two terms were used to leverage real-world biases throughout the bias identification process thus, exposing how each model deals with this innate bias. Following this, image subsets extracted from the datasets were human annotated to expose inherent bias within the DeepFace implementation which was used to extract the image features.

The presence of bias was determined based on a set of metrics, which consisted of gender, race, age and emotion distributions, metric groupings, and person prominence. These findings expose add overview of the results and conclusion reached as well as any anti-bias measures identified.

This research sheds light on the pervasiveness of bias in generative AI, highlighting the urgent need for proactive mitigation strategies. Our findings contribute to understanding bias and developing fairer models and datasets. Future work could explore advanced anti-bias techniques and broader societal implications of biased image generation.

Acknowledgements

I would like to thank my supervisor Dr Dylan Seychell for guiding me throughout the process of this final year project and aiding me throughout the various challenged encountered. I would also like to thank my parents, Reno and Graziella, and my brother Julian for their continuous support.

Contents

[Abstract (Style: Headings) ii](#_Toc120111054)

[Acknowledgements iii](#_Toc120111055)

[Contents iv](#_Toc120111056)

[List of Figures vi](#_Toc120111057)

[List of Tables vii](#_Toc120111058)

[List of Abbreviations viii](#_Toc120111059)

[1 Introduction (Style: Heading 1) 1](#_Toc120111060)

[1.1 How to Use this Template (Style: Heading 2) 1](#_Toc120111061)

[1.2 Using Styles 3](#_Toc120111062)

[1.3 Line Spacing for Headings **Error! Bookmark not defined.**](#_Toc120111063)

[1.4 Inserting Equations **Error! Bookmark not defined.**](#_Toc120111064)

[1.5 Automatically updating of Cross-Referencing **Error! Bookmark not defined.**](#_Toc120111065)

[1.5.1 Using Styles Correctly (Style: Heading 3) **Error! Bookmark not defined.**](#_Toc120111066)

[1.5.1.1 Sub-Titles (Style: Heading 4) **Error! Bookmark not defined.**](#_Toc120111067)

[1.6 Some Hints on the Dissertation Writing Style **Error! Bookmark not defined.**](#_Toc120111068)

[1.7 Some Hints on using *Word* **Error! Bookmark not defined.**](#_Toc120111069)

[1.8 Avoiding Plagiarism **Error! Bookmark not defined.**](#_Toc120111070)

[1.9 How to Write the Introduction **Error! Bookmark not defined.**](#_Toc120111071)

[2 Background and Literature Review 4](#_Toc120111072)

[2.1 References **Error! Bookmark not defined.**](#_Toc120111073)

[2.1.1 *RefWorks* **Error! Bookmark not defined.**](#_Toc120111074)

[2.1.2 Formatting the Reference Section **Error! Bookmark not defined.**](#_Toc120111075)

[2.1.3 Examples of Different Types of References **Error! Bookmark not defined.**](#_Toc120111076)

[2.2 Writing the Background and Literature Review Chapter 6](#_Toc120111077)

[3 Specification and Design 8](#_Toc120111078)

[3.1 Further use of Captions 8](#_Toc120111079)

[3.1.1 Tables 8](#_Toc120111080)

[3.1.2 Lemmas, Theorems, Corollaries etc. 9](#_Toc120111081)

[3.2 Chapter Headers 11](#_Toc120111082)

[3.3 Printing the Dissertation on Both Sides 12](#_Toc120111083)

[3.4 Writing the Specification and Design Chapter 15](#_Toc120111084)

[4 Implementation 16](#_Toc120111085)

[4.1 Writing the Implementation Chapter 16](#_Toc120111086)

[5 Testing and/or Evaluation 17](#_Toc120111087)

[5.1 Writing the Evaluation and/or Evaluation Chapter 17](#_Toc120111088)

[6 Future Work 18](#_Toc120111089)

[6.1 Writing the Future Work Chapter 18](#_Toc120111090)

[7 Conclusions 19](#_Toc120111091)

[7.1 Writing the Conclusions Chapter 19](#_Toc120111092)

[References 20](#_Toc120111093)

[Bibliography 21](#_Toc120111094)

[Appendix A When to Use Appendices (Style: Appendix Heading 1) 22](#_Toc120111095)

[A.1 Appendices Headings (Style: Appendix Heading 2) 22](#_Toc120111096)

[A.1.1 Appendix Sub-Headings (Style: Appendix Heading 3) 22](#_Toc120111097)

[A.2 What to put in Appendices 22](#_Toc120111098)

[Appendix B Version Information 23](#_Toc120111099)

[B.1 Copyright Information 23](#_Toc120111100)

[B.2 Suggestions 23](#_Toc120111101)

List of Figures

[Figure 1.1 Cross-referencing (Style: Caption Figure). **Error! Bookmark not defined.**](#_Toc120084584)

[Figure 1.2 Inserting a caption to a figure. **Error! Bookmark not defined.**](#_Toc120084585)

[Figure 1.3 Figure example, a rectangle and a triangle. **Error! Bookmark not defined.**](#_Toc120084586)

[Figure 1.4 Keeping paragraphs or lines on the same page. **Error! Bookmark not defined.**](#_Toc120084587)

[Figure 2.1 *RefWorks Citation Manager* settings options. **Error! Bookmark not defined.**](#_Toc120084588)

[Figure 2.2 Inserting a citation using *RefWorks*. **Error! Bookmark not defined.**](#_Toc120084589)

[Figure 2.3 Content Control Box containing the references. **Error! Bookmark not defined.**](#_Toc120084590)

[Figure 2.4 Correctly formatting the Reference section. **Error! Bookmark not defined.**](#_Toc120084591)

[Figure 3.1 Converting document to printing on both sides. 14](#_Toc120084592)

[Figure 3.2 Inserting chapter headers. 15](#_Toc120084593)

List of Tables

[Table 3.1 Simple table example. 8](#_Toc120084594)

[Table 3.2 A more complicated table example. 9](#_Toc120084595)

List of Abbreviations

FYP Final year project (Style: Abbreviations)

AI Artificial Intelligence

GAN Generative Adversarial Network

VAE Variational Autoencoder

CLIP Contrastive Language-Image Pre-training

UNET U-shaped encoder-decoder network architecture

ResNet Residual Neural Network

Note that the List of Abbreviations should be sorted on the acronym list.

The entries in the List of Abbreviations should be assigned the Abbreviations style.

# Introduction

## Problem Definition

In recent years, the field of Generative AI has experienced remarkable advancements in visual content generation, with a primary focus on images. Notably, generative models such as Midjourney, DALL-E and Stable Diffusion have been at the forefront of this progress \cite{midjourney, dall-e-2, stable-diffusion-online}, by providing users with the capability to generate numerous images through the use of a simple text prompt.

However, the generation of visual content brings to the forefront a variety of critical issues such as lack of control over output, over fitting as well as privacy and ethical concerns \cite{Controllable-Generative-Adversarial-Network, GAN-Privacy-Ethics-Concerns}.

This study focuses on a particular issue, that of bias. Bias in relation to visual AI systems tends to refer to cases in which systems showcase prejudice in relation to particular demographic features, gender and race being the primary focus of this paper \cite{Bias-Gender-Race}. Several instances exist in which this prejudice led to negative consequences in relation to recidivism scoring \cite{COMPASS-situation-racial-bias}, online advertisement \cite{Discrimination-in-Online-Ad-Delivery}, facial recognition \cite{Facial-Recognition-Negative-Consequnces}, and credit scoring \cite{Credit-Scoring-Negative-Consequnces}.

Bias serves to affect a large majority of computer vision systems such as classification algorithms, face recognition systems, object detectors and many more \cite{RefWorks:RefID:30-fabbrizzi2022survey}. To address this problem tools can be created which aid in the identification of bias, these are crucial as bias is not attributed to a singular cause rather a variety of factors varying from the composition of the dataset and the framing of images to the characteristics of the latent space employed during the generative process \cite{RefWorks:RefID:30-fabbrizzi2022survey}.

Tools such as this already exist, a prime example is the REVISE implementation which given an annotated dataset can provide object-based, person-based and geography-based insights on the presence of bias \cite{revisetool\_eccv}. However, such systems tend to be cumbersome to set-up and utilise. The initial aim of this study was to detect if bias is present in traditionally gender biased prompts such as doctor and nurse by looking at the prompt associated images of the LAION-5B dataset as well as generated images from the Stable Diffusion model to detect any forms of bias with a focus on gender and race. However, due to recent proceedings with the LAION-5B dataset, wherein access to said dataset was revoked the aim of the study was shifted \cite{<https://cointelegraph.com/news/laion-5b-ai-data-set-removed-child-sexual-abuse-material>}. Thus, this study will attempt to outline the presence of bias in the LAION-400M dataset as opposed to the LAION-5B dataset whilst also considering various generative models these being Stable Diffusion, DALL-E and Midjourney. This study also aims to develop a simple to use python notebook which will facilitate image feature extraction and metric visualisation to allow individuals to easily detect bias.

## Motivation

The motivation behind this research stems from the growing importance of addressing bias in artificial intelligence (AI) systems, particularly within the realm of generative models and visual datasets. As AI technologies continue to play an increasingly integral role in shaping various aspects of our lives, understanding and mitigating biases becomes imperative. The LAION-400M dataset and Midjourney, DALL-E and Stable Diffusion models serve as focal points for this study, representing key components in the landscape of generative AI. By investigating and uncovering biases present in these specific entities, this research aims to contribute valuable insights to the broader discourse on ethical AI development. The implications of biased AI systems are far-reaching, with potential consequences in areas such as image generation, facial recognition, and algorithmic decision-making. Through a meticulous examination of biases, this study strives to not only enhance our understanding of the challenges inherent in generative models but also to pave the way for more ethical and unbiased AI systems in the future.

## Aims and Objectives

The aim of this study as outlined above is to determine the presence of bias in popular training datasets and generative AI. This aim will be achieved via the following set of objectives:

1. Analyse bias-associated prompts and determine an optimal feature extraction model. Determine the requirements needed to select appropriate human annotators for valid annotations. Resulting in an optimal prompt structure by which images can be generated using any label and model as well as denoting the requirements for valid human annotations.
2. Generate images using pre-defined prompts containing the *Doctor* and *Nurse* terms with Midjourney, DALL-E, and Stable Diffusion. Extract the main image features (gender, race, age) from both generated and training dataset images, including a human-annotated training dataset subset for bias detection in the feature extraction model used. This will result in three annotated image sets: generated, training data, and human-annotated training data alongside the identification of inherent bias of the feature extractor.
3. Analyse the extracted features consisting of gender, race, age, emotion distributions, and overall person prominence across data groups. Visualize these metrics to aid in identifying relationships between the data and drawing conclusions.
4. Through expert interviews and qualitative analysis regarding the visualised metrics, uncover relationships within the data to identify the optimal training dataset and model in terms of lack of bias, while revealing common bias manifestations in training datasets and models.

## Document Structure

Add this section later once the document starts to take more shape.

# 

# Background

This chapter provides a foundation of knowledge required for understanding the techniques employed within the developed system for bias detection in both visual datasets and generative models by introducing the relevant background concepts and techniques.

The chapter is divided into three subsections covering *prompting*, *CLIP*, diffusion models, *face recognition, facial analysis,* and *image bias,* going over a variety of relevant research and challenges associated with each section. Furthermore, the chapter outlines how each section fits into this research paper*.*

## Prompting

Prompting in the context of AI models can be defined as the act of providing the model with instructions that guide the generation process of text, code, images, and other varied outputs. These instructions can take various forms, the most common of which being text, code, and images. Given that this research paper concerns itself with text-to-image generation and the bias therein only text inputs and image outputs are considered.

Although on the surface prompting appears quite straight forward it brings with it a variety of challenges mainly in relation to retrieving relevant images. This challenge is closely tied with identifying the right prompt, which is a non-trivial task as it not only takes a significant amount of time but minor changes to the prompt could result in a huge impact on performance \cite{coop}. This is where prompt engineering comes into play which involves altering the prompt in a variety of ways such as altering its length and wording used to affectively depict the required output as opposed to merely specifying the desired image \cite{[Prompt engineering - OpenAI API](https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results)}. This process can also be enhanced through automated prompt engineering however this was beyond the scope of this research paper.

## CLIP

CLIP is a multimodal vision and language model developed by OpenAI which was trained on 400 million image-text pairs collected from publicly available sources on the Internet in an attempt to cover as broad a set of visual content as possible \cite{clip-paper}. Contrary to traditional models which predict a fixed label for images, CLIP adopts a contrastive learning approach which allows it to learn the relationship between image and text pairs. This approach allows CLIP to determine the best image-text pairs for any possible use case. CLIP further differs from traditional image classifiers as it utilises zero-shot learning, a technique whereby a model is able to generalise to unseen classes without the need for training, thus allowing it to deal with never-before-seen images and classes \cite{zero-shot-learning}.

CLIP is relevant to this paper as it was used to retrieve images from the LAION-5B dataset via the clip-retrieval library by converting the text query to a CLIP embedding, then using that embedding to query a k-nearest neighbour index of clip image embeddings \cite{LAION5BClipSearch, clip-retrieval}. This model was used to retrieve images from the LAION-5B dataset prior to it being taken down. Additionally, CLIP is also used within both Stable Diffusion and DALL-E as a text encoder to generate text-embeddings required by said models to generate correct images \cite{stable-diffusion-clip-reference, dall-e-clip-reference}. It is undisclosed if Midjourney also utilises CLIP in a similar manner or at all.

## Diffusion Models

Generative models encompass a variety of different approaches, including GANs, VAEs, and diffusion models. The latter offers several advantages over its counterparts. Unlike GANs, diffusion models excel in both training stability and diverse image generation, avoiding the pitfalls that often plague GANs. Additionally, they bypass the surrogate loss issue inherent in VAEs. This allows diffusion models to achieve superior performance and efficiency. The models considered in this paper all fall under the diffusion category \cite{dall-e-3-paper, stable-diffusion-paper, midjourney-pickfu-article}.

Diffusion models are traditionally composed of two steps, these being the forward and reverse diffusion processes. The forward diffusion process starts off with a clear image and slowly but gradually adds gaussian noise to it with every passing step. This is repeated until the input image no longer resembles the original input, these series of steps can be seen in Figure 2.1.

A diagram of a person's face

Description automatically generated

Figure 2.1 Forward diffusion process \cite{diffusion-models-explained}

The reverse diffusion process employs a noise prediction model which iteratively refines the noisy forward pass image towards a clear output. This process resembles the inverse of the forward pass as can be seen in Figure 2.2. The core component of this de-noising process is the UNet architecture, these are convolutional [neural networks](https://aws.amazon.com/what-is/neural-network/) originally developed for image segmentation in biomedicine. In particular, Stable diffusion adopts the ResNet model developed for computer vision. During each step of the iterative process, the noise predictor estimates the noise component present in the latent space representation of the image. This estimated noise is then subtracted, effectively denoising the image. This cycle repeats for a predefined number of steps, progressively removing noise and enhancing the image's detail. Notably, the noise predictor can be guided by conditioning prompts, influencing the final outcome and directing the image generation process towards specific themes or styles \cite{stablediffusion-process-explained}. The de-noising process can be seen in Figure 2.2.

A diagram of a mathematical equation

Description automatically generated

Figure 2.2 Reverse diffusion process \cite{diffusion-models-explained}

Further analysis of the Stable Diffusion model showcases that it utilises a variant of the diffusion model architecture known as the latent diffusion model. This model differs from the traditional diffusion model as it tackles the denoising stage within a compressed representation of the image, termed the "latent space," as opposed to the pixel space. This strategy offers significant computational advantages. Consider a standard 512x512 colour image, boasting a staggering 786,432 possible values. Contrastingly, Stable Diffusion operates on a compressed representation containing only 16,384 values, reducing its size by a factor of 48. This dramatic compression translates to tangible benefits, including vastly reduced processing demands, enhanced performance, and improved overall efficiency \cite{stablediffusion-process-explained, stable-diffusion-paper}.

## Face Recognition

Face recognition can be defined as a three-step process consisting of *Face detection*, *Feature extraction* and *Face recognition* where the input is always an image or video, and the output is the identification or verification of the image or video subjects \cite{face-recognition-pipeline, face-recognition-book}. Face detection is defined as the process by which image regions depicting faces are located and extracted, this has a variety of use cases like face tracking, pose estimation and compression. Feature extraction involves the retrieval of facial features from the data, which can be human relevant or not. These tend to include feature such as face regions, variations, angles, and measures. Feature extraction has a variety of use cases including facial feature tracking and emotion recognition. Face recognition utilises the outputs from the prior steps in conjunction with comparison methods, classification algorithms and an accuracy measure to recognise faces \cite{face-recognition}.

Although the pipeline is composed of three steps, there are cases in which Face detection is not carried out, particularly in instances where the images only contain the subjects face. This is not the case for the pipeline implemented in this paper as the images considered did not conform to this requirement. Furthermore, face detection must deal with several challenges \cite{face-recognition-book, face-recognition, face-recognition-2, face-recognition-3}:

* Pose variation – Large pose variation can severely decrease the performance of face detection algorithms. The obtain ideal results images should contain subjects forward facing.
* Feature occlusion – The obstruction of facial features can also decrease performance. This is usually caused by the presence of beards, glasses and other clothing items; however, faces can also be partially covered by objects or other faces.
* Facial expression – Different facial gestures can cause facial features to vary, thereby affecting facial detection.
* Imaging conditions – Image quality is a major factor in facial detection, this can be affected by the lighting conditions and image size which are determined by the camera and varying Ambiental conditions.

Feature extraction unlike face detection is crucial to the pipeline and it involves the extraction of a variety of differing features which depend on the feature extraction model used. These include colour-based, spatial, textural, geometric, and deep learning features. The type of features extracted vary based on the use case as colour-based features tend to see usage in image segmentation and retrieval, spatial features in object detection and image classification, textural features in texture analysis and material classification, geometric features in facial analysis and 3D face reconstruction and deep learning features in face recognition, object detection and image generation.

## Facial Analysis

Facial analysis is distinct from face recognition as it forgoes the identification of the subject, in favour of identifying the subjects facial features. However they are similar processes as they follow the same three-step process outlined in the Section 2.4 differing only in the final step where one or more of the following attributes are extracted; face, age, gender, race, head pose and so on \cite{faceImageAnalysis}. These outputs each serve as their own distinct problem each using different facial features and processes however, they are all related to one another thereby the progress made in one area serves the benefit the others.

This variety in features results in varied use cases for facial analysis including but not limited to:

* Surveillance – Face analysis and tracking are used in surveillance as depicted in \cite{facialAnalysisSurveillance} which determines an events excitement based on peoples attention in a particular scene.
* Targeted advertisement – In a similar fashion an individuals attention can be used to discern an advertisements capability to keep people engaged, such an application can be seen in \cite{facialAnalysisAdvertisement}.
* Driving safety – Face analysis can be used to ensure driver safety, by checking on the driver, ensuring they are fit to drive. Such a system can be seen in \cite{facialAnalysisDriving} where the emotional state of the driver is identified, and the driver is influenced accordingly to promote safer driving states.
* Estimation of face, expression, gender, age, and race. – Although facial analysis can be used for complex tasks as mentioned prior it can also be used simply to estimate face, expression, gender, age, and race to aid in varying processes such as image annotation as is the case of this paper.

Facial analysis being a counterpart of face recognition makes use of the same type of inputs these being image and video and similarly deals with the same types of issues outlined in Section 2.4 these being Pose variation, Feature occlusion, Facial expression and Imaging conditions.

## Image Bias

Image bias in relation to visual AI systems as defined in Section 1.1 tends to primarily refer to cases in which systems showcase prejudice in relation to certain demographic features \cite{Bias-Gender-Race}. However, bias as can present itself in a variety of different forms, these can be broadly categorised as \cite{bias-types-visual-datasets}:

* Selection bias – When visual data is gathered unevenly, favouring certain subjects or aspects over others, this creates inaccurate representation and biased results. Imagine a dataset filled mostly with young, white models – it would not reflect the real world and could lead to discriminatory outcomes.
* Framing bias – How images are composed can manipulate our perception. Angles, lighting, and even expressions can unconsciously influence our understanding. Framing bias occurs when these choices lead to unfair interpretations or misleading connections.
* Label bias – Accurate labels are crucial for understanding visual data. Label bias arises when images are tagged incorrectly, either because the categories are poorly defined or because the labelling process introduces errors. This can distort the meaning of the data and hinder accurate analysis.

These types of biases in most cases are not intentional rather they occur due to some unforeseen consequences of the data collection and annotation process. Thus, it is crucial to identify and mitigate such bias. Bias detection techniques can be categorised as either subjective or objective. The latter using statistical and algorithmic approaches whereas the former utilises human judgment to come to a conclusion based on the resultant data. These approaches usually go hand in hand as can in be seen in \cite{revisetool\_eccv} wherein the tool itself utilises various algorithmic techniques to extract relevant feature however the final judgement on bias must be carried out by an individual which can consider the presented data along with the dataset context and thus, come to a conclusion.

Bias mitigation techniques vary in their implementation however there are certain aspects one must keep in mind in order to mitigate bias, these include but are not limited to \cite{RefWorks:RefID:30-fabbrizzi2022survey}:

* Selection bias
  1. Data representativeness - Balanced or statistically representative?
  2. Negative set coverage - Negative sets representative enough?
  3. Excluded groups - Essential categories which are missing?
* Framing bias
  1. Image interpretation – Viewer dependent messages possible?
  2. Subject depiction – Particular group/label depicted in a particular manner more than others?
  3. Stereotype adherence – Does data perpetuate harmful biases?
* Label bias
  1. Automated labelling biases: Considered and mitigated?
  2. Annotator bias control: Diverse team and biases addressed?
  3. Label clarity: Fuzzy labels (gender/race) present?

## Chapter Summary

This chapter introduces the key concepts and techniques required to understand the content of this paper as well as its importance. It covers prompting, CLIP, diffusion models, face recognition, facial analysis, and image bias, explaining their purpose and relevance to the research. Additionally, it outlines how each section contributes to the overall paper.

Lit Review – Mention DeepFace, Inferdo and BetaFace \cite{deepFace-ref-1, deepFace-ref-2, inferdo, betaFace}.