|  |
| --- |
| 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 various challenging encountered. I would also like to thank my parents, Reno and Graziella, and my brother Julian for their continuous support.

Contents

[Abstract (max 300 words) ii](#_Toc160004662)

[Acknowledgements iii](#_Toc160004663)

[Contents iv](#_Toc160004664)

[List of Figures v](#_Toc160004665)

[List of Tables vi](#_Toc160004666)

[List of Abbreviations vii](#_Toc160004667)

[1 Introduction 1](#_Toc160004668)

[1.1 Problem Definition 1](#_Toc160004669)

[1.2 Motivation 2](#_Toc160004670)

[1.3 Aims and Objectives 3](#_Toc160004671)

[1.4 Document Structure 3](#_Toc160004672)

[2 Background 5](#_Toc160004673)

[2.1 Prompting 5](#_Toc160004674)

[2.2 CLIP 6](#_Toc160004675)

[2.3 Diffusion Models 7](#_Toc160004676)

[2.4 Face Recognition 8](#_Toc160004677)

[2.5 Facial Analysis 10](#_Toc160004678)

[2.6 Image Bias 11](#_Toc160004679)

[2.7 Chapter Summary 12](#_Toc160004680)

[3 Literature Review 13](#_Toc160004681)

[3.1 Prompting 13](#_Toc160004682)

[3.1.1 Prompt target 13](#_Toc160004683)

[3.1.2 Prompt structure 14](#_Toc160004684)

[3.1.2.1 Stable Diffusion 14](#_Toc160004685)

[3.1.2.2 DALL-E-3 14](#_Toc160004686)

[3.1.2.3 Midjourney 14](#_Toc160004687)

[3.2 Facial Analysis 16](#_Toc160004688)

[3.3 Image Annotations 18](#_Toc160004689)

[3.3.1 Human based image annotation 18](#_Toc160004690)

[3.3.1.1 Strengths of Human based image annotation 18](#_Toc160004691)

[3.3.1.2 Challenges and limitations 19](#_Toc160004692)

[3.3.1.3 Considerations for affective Human based image annotation 20](#_Toc160004693)

[3.3.2 Computer assisted image annotation 20](#_Toc160004694)

[3.3.2.1 Strengths of Computer assisted image annotation 20](#_Toc160004695)

[3.3.2.2 Challenges and limitations 21](#_Toc160004696)

[3.3.2.3 Considerations for affective Computer assisted image annotation 21](#_Toc160004697)

[3.4 Measuring Bias (Metrics/Techniques) 21](#_Toc160004698)

[3.5 Chapter Summary 23](#_Toc160004699)

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

URL Uniform Resource Locator

API Application Programming Interface

MAE Mean Absolute Error

MST Monk Skin Tone

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 representations 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.

# Literature Review

This chapter provides a detailed overview of past and recent studies relevant to this area of research. This chapter firstly, outlines prompting going over its structure and the selected target. This is followed by an in depth look at several facial analysis models and the issues therein, furthermore the two types of image annotations are further explained presenting both their pros and cons. Finally, this chapter concludes with a look at various bias measurements used in similar scenarios to that of this research paper. The research considered in this section provides a foundation for the construction and execution of the research proposed in this paper.

## Prompting

In accordance with the discussion regarding prompting in section 2.1, prompting is the process by which a person guides an artificial intelligence model to generate required output. The models considered in this research include Stable Diffusion, DALL-E and Midjourney (CLIP was removed since we are now using LAION-400M) and with the usage of said models come two main challenges these being the selection of the prompt target and the prompt structure itself.

### Prompt target

In accordance with the nature of the research that being the identification of bias within the aforementioned models the prompt subjects will consist of professions which are traditionally gender biased. These being male dominated and female dominated spaces in particular the subjects considered for the study are *doctor* and *nurse*. The gender bias within these prompts is showcased via the research carried out in \cite{gender-bias-nurse, gender-bias-doctor-nurse} wherein the doctor profession is male dominated and the nurse female dominated. This bias is not only showcased in the real world but also depicted online \cite{online-images-amplify-gender-bias}. This is relevant because the majority of these generative AIs tend to utilise training data which has been retrieved from the Internet; an instance of this is Stable Diffusion which was trained on the LAION-5B dataset a successor to the LAION-400M whose images are retrieved from the common crawl and thus, this bias might be present \cite{LAION5B-mainpage, commoncrawl\_faq}.

### Prompt structure

#### Stable Diffusion

The stable diffusion webui repository \cite{stablediffusion-webui-repo-wiki} outlines the additions present in this tool such as upscaling, img2img, negative prompting, face restoration, model merging and so on. It also outlines stable diffusions prompt length limit however this will not impact the results from this FYP. The stable diffusion prompt guide \cite{stable-diffusion-prompt-guide} outlines the importance of prompt precision, defining the image medium (digital art, sketch, painting), specifying a particular style (hyper realistic, fantasy) and the usage of negative prompts and so on.

#### DALL-E-3

The official DALL-E 3 documentation \cite{dalle3-prompt-guide} outlines how the DALL-E 3 input prompt is automatically rewritten for safety reasons and to increase prompt detail as this tends to result in higher quality images. Furthermore, this capability currently cannot be removed as such it is recommended to precede the prompt with “*I NEED to test how the tool works with extremely simple prompts. DO NOT add any detail, just use it AS-IS:*” to produce images closer to the initial prompt. The OpenAI Developer Forum \cite{dalle3-tips-thread} offers general prompting tips such as be specific and detailed, use descriptive adjectives, avoid prompt overloading, specify desired styles or themes. Lastly, images generated via DALL-E 3 can have a size of 1024x1024, 1024x1792 or 1792x1024 pixels.

#### Midjourney

The official Midjourney documentation \cite{midjourney-prompt-guide} outlines how the Midjourney bot breaks down the prompt into smaller chunks called tokens, which are then compared with its training data to generate prompt relevant images. It suggests using simple, short sentences as opposed to a long list of requests and instructions to obtain the best results. Furthermore, it outlines the creation of advanced prompts composed of image prompts, text prompts and parameters. Image prompts consist of an image URL which influences the style and content of the generated image. Text prompts consist of a text description of what image you want to generate. Parameters alter the resultant image by changing its aspect ratios, model, upscaling and so on. This prompt structure can be seen in Figure 3‑1.

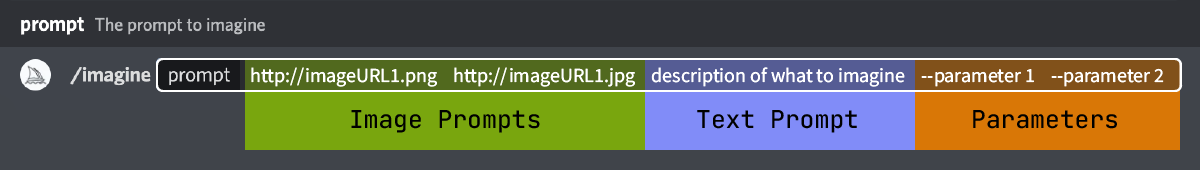


Figure 3‑1 Midjourney prompt structure.

The guide further outlines certain crucial aspects which need to be considered in relation to the producing the best text prompt, these being:

* Word Choice – Use precise synonyms, numbers, and collective nouns.
* Prompt focus – Focus on what you want in the image, not what you don't. Use the *--no* parameter for specific exclusions.
* Length/Detail – Short prompts offer creativity, while detailed prompts offer control. Include necessary elements for your desired outcome.

Thus, from the research outlined in this section it is safe to conclude that although these models differ from one another the general prompt features tend to remain consistent overall. Taking into consideration said fact a general prompt structure for image generation can be constructed, this being *A hyper realistic picture of a [label]* where label can be substituted for doctor and nurse in the case of this FYP. This prompt incorporates all the prior suggestions:

* Hyper realistic – denotes the style of the image given that we want real human people and not art.
* Picture – denotes the medium whilst further emphasising the realistic nature of required image.
* Label – denotes the image focus whilst ignoring aspects which we do not want in the image.
* Short length – this allows the models to produce outputs without too much human influence ensuring that the generated output is untampered with.

## Facial Analysis

Facial analysis, as outlined in section 2.5 refers to the extraction of varied facial features, each with their own challenges and issues. Focusing on the features relevant to this paper several techniques exist by which they can be retrieved ranging from Support Vector Machines, Radial Basis Functions and Deep Learning based methods.

Implementations of varied SVM models are described in \cite{gender-classification-SVM} where gender classification was carried out. These SVM models aim to find the optimal linear hyperplane by which the expected classification error for unseen data can be minimised. These SVM models traditionally classify linearly separable data however through the use of the kernel trick non-linearly separable data can also be classified thus, allowing a greater degree of applications. The kernel trick in question uses kernel functions such as the RBF to map the data to a higher dimension thus, allowing linear classification. Determining the hyperplane which facilitates the classification is generally denoted as a constrained optimisation problem and solved using quadratic programming techniques.

A particular implementation of the RBF model is described in \cite{gender-classification-RBF} where gender classification is carried out. This is achieved through feature extraction wherein facial texture, hair geometry and moustache features are extracted, these are then fed into the M-estimator based RBF Neural Network for classification. The network uses an M-estimator to handle outliers and improve classification performance. The structure of this network consists of various hidden layers with RBFs and the necessary output layer for classification.

The latter involving the training of a Convolution Neural Network (CNN) using a vast and expansive labelled dataset, thereby allowing for gender, age, race and emotion estimation and classification. Instances of these models include Googles Google Vision API and Amazons Rekognition API, the latter implementing only gender and emotion classification whilst the former only implementing emotion classification. These APIs implement other functionalities such as object detection, text detection and so on, however they minimise their classification functionalities as to implement systems with high performance requires vast unbiased training data and access to said data. Additionally, given the size and influence of these companies they have to take into consideration the possible affects that releasing such models can have on society which can be quite problematic as can be seen with Meta’s discontinuation of its face recognition system in the wake of sustained privacy and ethical concerns such as the abuse of marginalised groups and further racial bias \cite{facebook-facial-recognition-shut-down}.

Contrarily, the open-source DeepFace API implements age, gender, race and emotion estimation and classification with varying degrees of success. The age and gender models were implemented using the VGG-Face model in which the initial layers were frozen whilst the remainder were trained on a subset of the IMDB+Wikipedia dataset, the race model underwent similar training on the FairFace dataset. The implementation of the emotion model required a custom architecture depicted in Figure 3‑2. and was trained on the FER-2013 dataset. These models achieved varying degrees of accuracy on their respective test sets, with the gender prediction model having an accuracy of 97.44%, the race prediction model had an accuracy of 68% with the emotion model having a 57.42% accuracy. Finally, the age model achieved an MAE of 4.65 meaning that the age can be predicted with plus and minus 4.65 years \cite{age-gender-betaface-model, race-betaface-model, expression-betaface-model}. Similar to DeepFace the BetaFace API implemented the same functionalities, however the model performance is not public, thus requiring further testing \cite{betaFace}.

A table of numbers and letters

Description automatically generated

Figure 3‑2 Age model architecture \cite{deepFace-ref-2}.

## Image Annotations

Image annotation is the process by which labels are assigned to an image or image set. This is a crucial component of any study particularly those revolving around bias as the metrics used to deduce a conclusion need a basis on which to be made. Image annotation is commonly carried out in either of two ways, these being computer assisted and human based image annotation.

### Human based image annotation

Human based image annotation makes use of human annotators to correctly identify and label images in accordance with the requirements. Although computer assisted image annotation has become more prevalent, there is still a place for human based image annotation in various application such as computer vision and machine learning. This process however comes with degree strengths and challenges which require addressing to achieve proper results.

#### Strengths of Human based image annotation

* Nuance and Context – Human annotators excel at understating complex visual information and incorporating context into their annotations. This is supported by \cite{humans-recognising-emotions} which found that with minimal facial expression information, humans can recognise positive expressions and negative expression to an acceptable but lesser degree.
* Adaptability – Humans can adapt their annotations in accordance with the task and data provided as reported in \cite{Google-monk-skin-tone-annotations} wherein it was noted that annotators were able to adapt to varying conditions such as differing image hue, saturation and brightness in relation to skin tone annotation.
* High Accuracy – It has been shown in \cite{humans-recognising-gender} that humans can carry out specific annotation tasks with high accuracy such as having a 96% accuracy in relation to face gender annotation, excluding hairstyle, makeup and facial hair context cues.

#### Challenges and limitations

* Annotator Bias – Individual’s location, culture and lived experiences affect the manner by which people annotate images as was showcased in \cite{Google-monk-skin-tone-annotations} wherein annotators from the USA labelled a particular individual as having a darker skin tone as opposed to Indian annotators as depicted in Figure 3‑3.
* Consistency – Maintaining consistent annotations across different individuals can be challenging as subjective interpretations of ambiguous visual cues can lead to inconsistencies, impacting the reliability of the data.
* Scalability and Cost – Manual annotation is not only time consuming but expensive, especially for larger dataset, this limits the scalability for large-scale projects and is the main reason for the use of computer-assisted image annotation \cite{image\_annotation\_guide}.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 3‑3 The distribution of Monk Skin Tone Scale annotations for this image from a sample of 5 photographers in the U.S. and 5 photographers in India in \cite{Google-monk-skin-tone-annotations}.

#### Considerations for affective Human based image annotation

* Annotator Selection – In lieu with \cite{Google-monk-skin-tone-annotations} annotators should be selected from geographically diverse backgrounds to ensure accurate annotations, mainly in relation to the annotation of images containing humans.
* Quality Control – Implement a standard set of labels and measures used throughout the annotation process to ensure a cohesive standard.
* Usage of Tools – Integrate annotation tools such as Roboflow and those similar to it to reduce the overall time required \cite{image\_annotation\_guide}.

### Computer assisted image annotation

Computer assisted image annotation makes use of AI models such as those discussed in Section 3.2 to remove the human component from the annotation process in favour of an AI model. Similarly, to human based image annotation this comes with a varying degree of strengths and weaknesses.

#### Strengths of Computer assisted image annotation

* Scalability and Cost – In contrast to human based image annotation computer assisted annotation can easily be scaled up to handle larger volumes of images at no additional cost \cite{advantages-computer-assisted-image-annotation}.
* Consistency – Assuming a consistent model architecture and pipeline are used throughout the annotation process, labels will remain consistent across a variety of images \cite{advantages-computer-assisted-image-annotation}.
* Efficiency – Automated image annotation can quickly annotate large datasets, saving both time and resources \cite{advantages-computer-assisted-image-annotation}.

#### Challenges and limitations

* Performance – The performance of the model varies drastically based on the training dataset used, assuming that the dataset and labelling task consist of similar images, outputs will tend to be correct however variation between training and real-world application can result in incorrect labels \cite{challanges-computer-assisted-image-annotation}.
* Model Bias – Similar to human based image annotation, models replicate the bias present in the training data as such this can be mitigated assuming a fair and balanced training dataset.
* Context – Computer assisted image annotation tools primarily rely on patterns and features, often struggling with nuances and context specific to the task. This can lead to misinterpretations and errors, especially in complex or ambiguous scenarios.

#### Considerations for affective Computer assisted image annotation

* Data Quality – Prior to implementing a model for image annotation consider the quality of the data on which it was trained, to determine a models applicability to a particular task \cite{considerations-computer-assisted-image-annotation}.

## Measuring Bias (Metrics/Techniques)

In accordance with the study's objectives, it is imperative to establish the potential biases in the LAION-5B dataset. The primary focus of this research revolves around person-based bias which encompasses the following list of biases:

* Gender Bias - Concerns the prominent gender expressed by individuals in the image.
* Age Bias - Concerns the prominent age expressed by individuals in the image.
* Race Bias - Concerns the prominent race expressed by individuals in the image.
* Emotional Bias - Concerns the prominent emotion expressed by individuals in the image.
* Label Bias - Concerns the manner by which images are labelled in the dataset.
* Person Prominence - Concerns the importance assigned to a person in an image.
* Appearance Differences - Concerns the differences in appearance between people.

These biases were taken into consideration after studying similar tools such as that presented in \cite{revisetool\_eccv}.

The detection of the biases outlined earlier will proceed following a series of steps, aligned with the methodologies discussed in \cite{revisetool\_eccv,BiasDetectionPipeline-schaaf2021measuring}. These being:

* Feature extraction via the use of human or computer-assisted image annotation.
* Analysis of accompanying text labels through Natural Language Processing models.
* Calculation of bias specific relevant metrics.
* Presentation of relevant metrics in easily interpretable format.
* Formulating conclusive remarks based on the metrics presented.

## Chapter Summary

This chapter delves into the research and techniques crucial for the execution of this research. Elaborating on prompting, covering its structure and target selection. It then explores various facial analysis models, highlighting the inherent challenges. Subsequently, the chapter dives into the two primary image annotation methods, outlining their strengths and limitations. Finally, it concludes with an exploration of various bias measurement techniques employed in similar research areas, laying the groundwork for the construction and execution of the research proposed in this paper.