RACIAL AND INTERSECTIONAL DEBIASING OF CONTRASTIVE LANGUAGE IMAGE PRETRAINING

by

ELIZABETH HOEPFINGER

(Under the Direction of Frederick Maier)

ABSTRACT

Bias in artificial intelligence is prevalent, especially among generative models. One such model, Contrastive Language Image Pretraining (CLIP) is used to classify images in one-shot tasks, and is pre-training for the image generation model DALL-E. Bias reflected in these models are harmful towards individuals of protected classes (e.g., race, gender, age, and sexuality). This thesis proposes two debiased models of CLIP: CLIP-Race and Intersectional-CLIP, which are debiased versions of CLIP on race and intersectional ethnicity and gender respectively. Both models follow a proposed debiasing protocol, using an adversarial classifier to prepend learnable prompt tokens to train and debias CLIP. Results show reduced bias in both instances as measured on 6 metrics. Additionally, we present a discussion of the debias models' outputs, and its retained feature extraction capability through a linear probe evaluation.

INDEX WORDS: Machine Learning, Computation and Language, Computer Vision, Bias in Artificial Intelligence

RACIAL AND INTERSECTIONAL DEBIASING OF CONTRASTIVE LANGUAGE IMAGE PRETRAINING

by

ELIZABETH HOEPFINGER

A.B., University of Georgia, 2021

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2023

RACIAL AND INTERSECTIONAL DEBIASING OF CONTRASTIVE LANGUAGE IMAGE PRETRAINING

by

ELIZABETH HOEPFINGER

Major Professor:	Frederick Maier
Committee:	Kimberly Van Orman
	Khaled Rasheed

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2023

DEDICATION

I am dedicating this thesis to Betty Hoepfinger, or Nana as I knew her, who was a grandmother, mathematician, and antique aficionado. I know she would be so proud to see me finish this thesis and earn a master's degree.

Nana, I miss you every day. Thank you for telling me that "a good mathematician does not know every formula or calculation: they just know where to look," something that has stuck with me throughout my journey in academics, and something that I always remember when I make a successful google search.

Thank you for all the late-night talks, the baking lessons, the Christmas mornings, and the many slumber parties. I love you and I hope to see you again someday.

ACKNOWLEDGEMENTS

I would like to say thank you to Dr. Fred Maier, for helping me figure out what I wanted to investigate in this thesis. Without you, I would not have known the right direction to take or even where to start. Thank you for putting up with my writing, which was atrocious at times, and encouraging me to finish strong. Thank you to Dr. Van Orman for being a great professor in all her classes (via Zoom or in person). I really enjoyed your passion for teaching and interesting conversations. Thank you to Dr. Rasheed for giving me a class to look forward to every week and teaching the interesting subject of evolutionary computation.

I would also like to thank my family: Mom, Dad, and Rich. Thank you for believing in me throughout this degree and always encouraging me while writing this thesis. Thank you for sending me to school and setting me up for a successful life. Rich, thank you for challenging me, and letting me feel good when I know something that you don't.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

1 INTRODUCTION

Bias is the presence of systematic and unfair favoritism or discrimination in the outcome of the decision-making process. Artificial intelligence and machine learning models can reflect bias learned from training data, engineers, and users. Debiasing is the effort to remove bias to reduce the negative impact created by these models.

In this thesis, we debias the Contrastive Language Image Pretraining (CLIP) (Radford, et al., 2021) model on a racial and intersectional axis, extending work previously proposed by Berg et al (2022), where CLIP was debiased on gender. Using their proposed debiasing method, we show a reduction in both types of bias by further training CLIP using prompt templates and balanced labeled image datasets. Since CLIP is used in many downstream applications, including in the image generation system DALL•E 2, the debiasing performed in this thesis reduces biases in these applications (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022).

1.1 Bias and Intersectionality in Artificial Intelligence

Bias in artificial intelligence primarily affects individuals of a minority status and originates from the data, algorithm, or user interaction. When a model is trained using an unbalanced dataset it can make connections and assumptions harmful to the underrepresented group. For example, image generation models can be used to create intentionally harmful images depending on the user's motivation or the data used to train

it. It is difficult to avoid building unbiased systems, as AI can extract themes from large datasets that reflect society's historical and current biases.

Intersectionality is a theoretical framework delineating the interconnected nature of a person's multiple social identities, and the influence in which they afford them privileges or disadvantages. Introduced in 1989 by Kimberlé Crenshaw, a leading scholar of critical race theory and law professor at UCLA and Columbia University, this framework describes how intersecting identities encompassing gender, race, class, and sexual orientation combine to construct a single unique social experience (Crenshaw, 1991). Crenshaw argues that because she is both Black and a woman, she has different social experiences than a White woman or a Black man. This approach challenges the traditional notion of singular, isolated social identities, emphasizing their intricate intersection and impact on social disparities.

In the technical field of AI, most current debiasing work focuses on gender or race but not how the two interact (Berg, et al., 2022) (OpenAI, 2022). Some systems, like COMPAS (Kirkpatrick, 2017), a system that recommends prison sentencing decisions, learn bias from previous prison sentencing data to inform decisions. It may treat Black men and Black women differently than White men and women. The intersection of the race and gender of the individual whose information is input to COMPAS might greatly influence the decision of the program. This comprehensive view of bias and its effects is essential to reducing the potential for harm by AI systems.

1.2 Contrastive Language Image Pretraining (CLIP)

CLIP uses image and text encoders to measure similarity from input prompts and images (see Figure 1). CLIP is used for image classification, transfer learning, and

pretraining for the image generation model DALL•E 2. Created by Open AI, CLIP learns directly from raw text about images to predict and classify large datasets using zero-shot transfer of the model to downstream tasks (Radford, et al., 2021). Debiasing this model is crucial; due to its application in DALL•E 2, the features extracted by CLIP and applied to the image generation process are a significant source of bias, as discussed further in section 1.3. By debiasing with the proposed protocol (Berg, et al., 2022), a later iteration of DALL•E could produce diverse, less harmful results.

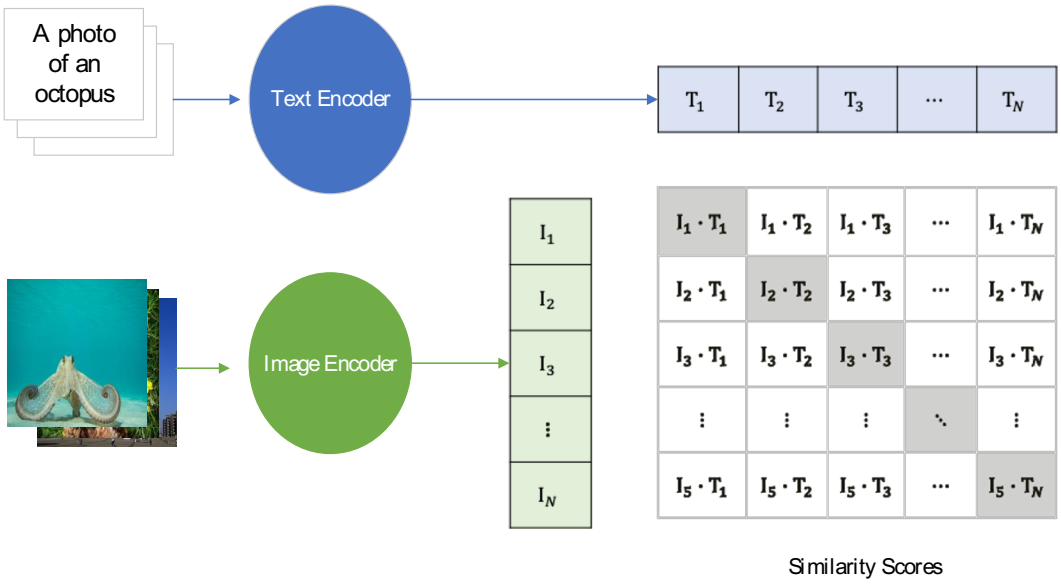


Figure 1: The architecture of CLIP. When given a set of text prompts and a set of images, CLIP outputs a set of similarity scores between each image and text prompt.

1.3 Relevant Problems: DALL·E 2

DALL• E 2 uses CLIP for feature extraction during image generation, so bias in CLIP directly influences results from DALL•E 2 (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022). Regardless of safety measures implemented by OpenAI to block offensive prompt inputs, violent images, and likenesses of celebrities, the image generation system persists in generating non-inclusive images that reflect inherent biases (Mishkin & Ahmad, 2022).

Figure 2 shows sample images generated by DALL•E 2 on September 4th, 2023. When asked to generate a "picture of a flight attendant on a plane," DALL•E 2 produces three images of feminine-looking people wearing skirts with light skin colors and one image of a masculine individual with a darker skin tone. When prompted with "a picture of a wedding," the system produces three photographs of light-skin-toned couples and one photo of a couple of color. All images contained heterosexual couples depicting a traditional Western wedding custom of the male in a tuxedo and the female in a white wedding dress. Our investigation into DALL•E 2 shows that bias is present in the model and that by debiasing its pretraining model, images generated with this system should include many cultures, lifestyles, and people.

Figure 2: Images taken from DALL•E on 09/14/2023. The top row of images are received when inputting the prompt “a photo of a flight attendant on a plane” and the bottom row of images are generated from the prompt “a photo of a wedding”.

1.4 Research Objectives

This thesis aims to debias CLIP on race and intersectional concepts (specifically the interaction of race and gender in the creation of bias) by following the debiasing protocol introduced by Berg et al (2022). We create two models: CLIP-Race and Intersectional

CLIP. We use two balanced, state-of-the-art datasets for debiasing: Balanced Faces in the Wild (Robinson, et al., 2020) for Intersectional CLIP and FairFace (Kärkkäinen & Joo, 2021) for CLIP-Race. Both provide data across race, age, and gender attributes and are used in similar debiasing experiments (Berg, et al., 2022) (Robinson, et al., 2020) (Kärkkäinen & Joo, 2021).

To measure bias in our models, we use *MaxSkew@k*. Berg et al. additionally used the *NDKL* (normalized discounted cumulative Kullback-Leibler divergence) metric to measure gender bias, but as explained later in chapter 4, this metric is not appropriate when many classes are involved. MaxSkew@k gives the skewness of the top k results according to a specified metric such as gender or race. This metric finds the ratio of the proportion of candidates with a specified quality (such as gender or race) in the top k results. Chosen because of its usage in prior debiasing experiments (Berg, et al., 2022) (Geyik, Ambler, & Kenthapadi, 2019), MaxSkew@k requires an "ideal" distribution that measures against the actual distribution. For the "ideal" distribution, we use three fairness criteria: demographic parity, equal opportunity, and differential fairness (Geyik, Ambler, & Kenthapadi, 2019) (Islam, Keya, Pan, Sarwate, & Foulds, 2023). These criteria are prevalent in debiasing discussions and papers inside and outside of the technical field.

1.5 Debiasing Protocol

We use the debiasing protocol proposed by Berg et al. to debias on both race and the intersection of race and gender to present two models: CLIP-Race and Intersectional CLIP. The protocol takes learnable prompts, sensitive text queries (e.g., “violent”), and labeled face images as inputs. Then, CLIP's text and image encoders produce a similarity score for each image and prompt input. The adversarial classifier uses the similarity scores to predict

the image's attribute label (e.g., race, gender) and prepend new prompts to train CLIP. The debiasing protocol aims to achieve equal similarity scores for each sensitive text query describing a non-physical attribute while maintaining a high similarity to physical features.

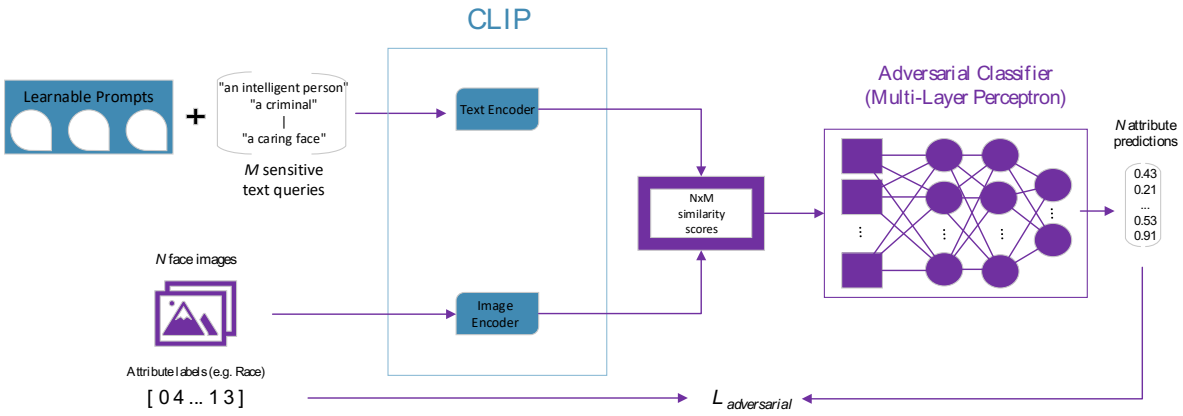


Figure 3: The proposed debiasing protocol based on Berg, et al. Using an input of labeled images fed to the image encoder and learnable prompts with sensitive text queries fed to CLIP's text encoder, CLIP outputs similarity scores between the image and prompt. The similarity scores are passed to an adversarial classifier, which predicts the image attribute of the score. The prediction is returned to CLIP, which is trained to make similarity scores for the sensitive text equal. This protocol reduces bias in CLIP while retaining most of its feature extraction capability.

Consider a dataset containing image-attribute pairs, where I is an image and A is its corresponding attribute from a set of protected attribute labels $A = \{A_1, \dots, A_l\}$. For example, these can be images of faces with a corresponding race label. There is also a set of sensitive text queries $T = \{T_1, \dots, T_m\}$ with corresponding concepts $C = \{C_1, \dots, C_m\}$, such as the queries “a photo of a {} person”, “a photo of a {} person”, and the concepts “smart” and “dumb”. Our goal is to train CLIP so that it outputs a similarity score for image-text pairs $s = CLIP(I, T)$; where semantically similar pairs are scored highly, but it is unbiased, which means it outputs similar distributions of scores across attributes for the given text query which are unrelated to demographic affiliation. CLIP already achieves the former, highly scoring images with semantically similar image-text pairs but does not have an even distribution across all attributes.

1.6 Summary of Results

In Berg et al.'s report, they claim to have reduced bias by 69% with MaxSkew@k and 80% with NDKL. It is important to note that our implementation of their publicly available debiased model did not achieve the reported decrease in bias. When measuring their publicly available gender debiased model with their evaluation code, we found gender bias was only reduced by 51.1% when evaluated with MaxSkew@1000 and 65% when measured with NDKL.

Our debiased models, which we call *CLIP-Race* and *Intersectional CLIP*, show a reduction in bias compared to the original CLIP ViT-B/16. When measuring racial bias in CLIP-Race, there was a 33-98% reduction depending on the fairness criteria. Although race was the attribute being debiased, there was also a significant reduction in measured gender bias of up to 26%. This correlation supports findings from Berg et al. by showing that debiasing one attribute, race, can also reduce bias in other protected classes (Berg, et al., 2022). Intersectional CLIP showed up to a 51% reduction in combined ethnicity and gender bias. Similar to CLIP-Race, there was also reduced bias in the separate categories of race and gender, reinforcing the hypothesis that by debiasing on one protected class, bias should also be reduced on other protected classes. Additionally, we support the reduction in bias with similarity score outputs to show more significant and qualitative change.

We also explore the similarity score output of CLIP-Race and Intersectional CLIP. By showing the subgroups of ethnicity/gender scores in CLIP ViT-B/16 and Intersectional CLIP when using sensitive text tokens, we show that strong similarities are reduced, and overall, scores become closer to equal in the debiased model due to a SoftMax function.

Additionally, we test CLIP-Race and Intersectional CLIP on the same set of sensitive text queries but include visual tokens such as “smiling”, “grinning”, “wearing a hat”, and “wearing glasses” to show that in some cases debiased models maintain a high similarity to physical concepts while still reducing sensitive tokens.

The set of results for all experiments can be found in chapter 4.

1.7 Key Contributions

These experiments extend previous debiasing research by implementing Berg et al.'s debiasing protocol on a racial and intersectional axis. By reducing bias on racial and intersectional bases and showing a reduction in bias through metrics and similarity scores, the research of this thesis provides an in-depth look at reducing bias in machine learning models. By introducing two new debiased models and venturing further into previous findings, this work extends the implementation of a promising debiasing protocol, furthering the possibility of debiasing of large AI models.

CHAPTER 2

2 BACKGROUND AND RELATED WORKS

This chapter serves as a background of bias in artificial intelligence and machine learning. We discuss the types of bias that exists in this context, current debiasing efforts, and examples of biased systems. Additionally, necessary technical background is given for some aspects of the debiasing protocol. Finally, we discuss the complex nature of intersectionality, and why measuring it is difficult to quantitively measure for this thesis.

2.1 Types of Bias in Artificial Intelligence and Machine Learning

Bias in AI has been described in three ways, each referring to bias that can happen between user interactions, data, and the algorithm itself, with the latter two being most prevalent in this thesis (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). The sections below describe definitions and examples of how these biases affect data, algorithms, and the user.

2.1.1 Data to Algorithm Bias

Data to algorithm bias is from the actual data and occurs in how data is used for training, producing biased results. The first type of data to algorithm bias, measurement bias, arises in the choice, utilization, and measurement of features from the data (Suresh & Guttag, 2019). The features included in the training of models can impact any bias present. For this thesis, using data that is balanced between the protected classes of race and gender is essential to avoid measurement bias.

Representation bias stems from how the population is sampled during data collection (Suresh & Guttag, 2019). If the data sampled does not represent the diversity of the population, these individuals will not be represented well in the model. One of the datasets used in this thesis, FairFace (Kärkkäinen & Joo, 2021), contains seven races and is the most racially inclusive public face dataset currently available. Data that is not diverse could be a cause of CLIP's bias. During training, the data used may not have been representative of members of diverse race, gender, and age. Sampling bias is similar to representation bias, but it occurs when subgroups are not randomly sampled, showing bias towards or against certain groups. Creating or using a dataset that is not randomly sampled and diverse between subgroups can inflict sampling bias. Additionally, aggregation bias occurs when drawing false conclusions about individuals from observing the entire population. Depending on how the data was collected, analyzed, or categorized, aggregation bias can occur and even stem from measurement or representation bias depending on how the data is utilized.

Word2Vec, a popular family of models used to create word embeddings, was found to have a large amount of gender bias in a 2016 investigation (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). When given the prompt "man is to computer programmer as woman is to ___", Word2Vec predicted "homemaker". The dataset w2vNEWS, a corpus of 3 million English words taken from Google News texts, was used to train Word2Vec. By training with professional news articles, it seems to follow that there should be less bias in the data; once trained, Word2Vec showed otherwise. The news articles used to train Word2Vec reflected harmful gender stereotypes, associating women with "nurse", "receptionist", and "librarian", and men with "maestro", "skipper" and "protégé". Because

of the large corpus, aggregation bias occurred, stemming from measurement or representation bias during training. The amplification of gender bias in Word2Vec is a classic example of data to algorithm bias.

2.2 Dataset Bias

Data bias is relevant, as few balanced datasets include diverse population sampling. Of course, data can reflect the biases of today's society and historical bias, as seen in COMPAS, the system used to determine sentencing and policing decisions and was found to be very biased against under-represented minorities (Kirkpatrick, 2017).

This experiment uses two datasets, FairFace and Balanced Faces in the Wild (discussed below in section 3.1). However, more datasets discussed below show a need for labeled face datasets that include many races and an equal sampling of genders. By having an abundance of majority white faces in many facial datasets, new models and systems trained on these datasets will contain algorithmic discrimination resulting from dataset bias (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016).

Many datasets are created by sampling images from the Yahoo Flicker 100M image dataset (e.g., FairFace) (Thomee, et al., 2016) , and some show significant race and gender bias. For example, Labeled Faces in the Wild (LFW), made of celebrity faces, contains 77.5% male faces and 83.5% White faces. This dataset has mainly been renowned as a gold standard benchmark for face recognition, even receiving the Mark Everingham Award for service to the Computer Vision Community. The creation of this dataset highlights how difficult it is to extract and automatically estimate race, age, and gender attributes from in the wild face images. While this dataset was groundbreaking to computer vision, the bias against people of color and females is detrimental to any systems that might use this dataset

for training. Although very important at the time, this "inclusive" dataset is exclusive of minorities and portrays dataset bias.

2.2.1 Algorithm to User Bias

Algorithm to user bias results from algorithmic outcomes that influence and affect user behavior. In the case of image generation, when requesting “a photo of a successful CEO”, if the results include photographs of mostly middle-aged white men, the effects of these images could harm people who are not part of this subgroup. Also see example from the introduction in section 1.3. User interaction bias comes from the user interface and the user itself when they impose their self-selected biased behavior and interaction on its system (Baeza-Yates, 2018). Users can significantly influence the generated images by pushing their values onto the model to create harmful results, even with safeguards in place. Emergent bias appears after system deployment and user interaction and is usually due to cultural values, population, or societal knowledge changes (Friedman & Nissenbaum, 1996). Emergent bias is usually found in user interfaces after user feedback shows flaws and needed usability changes.

Evaluation bias happens during model evaluation and can include using inappropriate and disproportionate benchmarks for evaluation criteria (Suresh & Guttag, 2019). Two benchmarks, Adience and IJB-A, were used to evaluate facial recognition systems that are biased toward skin color and gender (Buolamwini & Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, 2018). Buolamwini and Gebru found these datasets were composed heavily of lighter-skinned subjects and proposed their own dataset, Gender Shades (although it is now under audit). Gender Shades is worth mentioning because it would have been immensely useful in this

thesis, but it is no longer publicly available. Upon testing three commercially available gender classification systems, they found that the misclassification rate for darker-skinned females and males was very high, at up to 34%. Currently, there are no accepted benchmarks when it comes to bias testing, allowing for bias in many systems to thrive.

2.2.2 Case Study: Pilot Parliaments Benchmark

The lack of intersectional representative benchmarks inspired the Pilot Parliaments Benchmark (PPB) (Buolamwini & Gebru, 2018). The goal of PPB is to achieve better intersectional representation based on gender and skin type. It consists of 1270 individuals from three African countries (Rwanda, Senegal, South Africa) and three European countries (Iceland, Finland, and Sweden), chosen because their parliaments reflect gender parity. The images are of parliamentarians in each country because they are public figures with images posted on government websites. Parliamentarians included from African and European countries add a diversity of skin types.

When using PPB to evaluate three commercial gender classifiers (Microsoft, IBM, Face++), they found these systems more accurately classified men than women and lighter-skinned subjects than darker-skinned subjects. These discrepancies not only reflect the training data used in these three systems but can also be a fault of the lack of a benchmark test comparable to what PPB is trying to do. Without a dataset or benchmark for intersectionality or race and gender equality, these biased systems are not recognized, so the bias cannot be mitigated.

This thesis uses two of the most inclusive, publicly available datasets. FairFace and Balanced Faces in the Wild have previously been used to train and debias computer vision models (Berg, et al., 2022) (Kärkkäinen & Joo, 2021) (Robinson, et al., 2020). Although

there is no standard benchmark for facial recognition or classification systems yet, advances in debiasing are happening at a steady pace and these datasets have been praised for their inclusivity. PPB could not be used in this experiment due to it being under audit and not currently available to the public.

2.3 Debiasing Tokens and Learnable Prompts

In natural language processing, text tokens represent words, characters, or subwords. In this experiment, debiasing tokens are words from the Implicit Association Test that fill in learnable prompts and are prepended for input to the text encoder of CLIP. CLIP tokenizes the prompt with each debiasing token, and the adversarial classifier uses the similarity scores between the tokenized words and images to predict the attribute of an image (such as race or gender). This strategy is part of prompt learning and examples are shown below in Table 1. The learnable prompt and debiasing tokens strategy is used in this thesis for debiasing CLIP.

Table 1: A sample of a learnable prompt and debiasing tokens. These are used in experiments in this thesis and prepended to learnable prompts as part of debiasing CLIP.

Learnable Prompt	Debiasing Tokens
A photo of a {}	good, evil, smart, dumb, attractive, unattractive, lawful, criminal, friendly, unfriendly

2.4 Prompt Learning

Prompt learning and engineering research is crucial when working with large vision-language models such as CLIP. The strategy Berg, et al. uses in their debiasing protocol is

based on Context Optimization (CoOp) (Zhao, Yang, Loy, & Liu, 2022), which automates prompt engineering specifically for pre-trained vision-language models. CoOp models a prompt's context words with learnable vectors initialized with random values or pre-trained word embeddings. CoOp feeds learnable context vectors to the image and word encoders of vision-language pretraining models such as CLIP. It removes the need for complex prompt engineering, which is time expensive and requires specific knowledge of the subject area.

Results showed that CoOp could turn pretrained vision-language models into visual learners with as little as one or few-shot learning tasks to beat hand-crafted prompts by a good margin. Berg et al. implemented research in Zhao et al.'s CoOp by debiasing CLIP on gender. When starting with generic prompts, their debiased model learns token word embeddings used to finetune CLIP to reduce debiased results. We report their results, along with our own implementation of debiased CLIP in section 4.1.

2.5 DALL·E 2 "Debiasing"

Going back to the DALL·E 2 example from section 1.3, OpenAI researchers realized the harms of their model and added some safeguards (OpenAI, 2022). Although they did not specify the exact steps taken, Twitter user Andy Baio posted the results of a small experiment he did that prompted DALL·E 2 to generate "a sign that spells". This incomplete prompt returns images of people holding signs that spell words associated with minorities, the most notable being a sign that spelled "AFIRCA" held by a person of color. This led Baio and others to conclude that OpenAI "fixed" the issues defined in the example prior by adding phrases "behind the scenes". These could be phrases such as "from diverse

backgrounds" or "from any gender". The results of this “fix” are also reflected in the images shown previously in section 1.3.

These results show a shift from model-based to user-based debiasing, implying that the model is now beyond improvement, and it is easier to improve upon the user than the actual model itself (Offert & Phan, 2022). Large visual models reconceptualize reality through feature extraction and semantic compression. For example, the features of large concepts (e.g., race, gender, and social constructs) are condensed and even lost when translating them into one model. Although OpenAI devised a way to mitigate some bias, they could not change their model. For the users outside the organization, there must be a solution to reduce bias in these models that could become widely used for everyday tasks.

This thesis uses a proposed debiasing protocol that reduces bias in CLIP, a large vision-language model, by jointly training with an adversarial classifier. Using an external model as part of the debiasing protocol allows users with a machine learning background to debias an instance of CLIP without direct access to the model. This technique could also be applied to other models with a similar architecture, providing an accessible debiasing protocol. With the proven effectiveness of this technique, it could pave a way forward for future debiasing of other large machine learning models.

2.6 Intersectionality and Fairness in AI

Intersectionality can be described as the way in which someone’s social identities interact to afford them privileges or disadvantages. These identities can include race, gender, ethnicity, sexuality, and ability. At its core, bias exists due to oppression from people of power, and intersectionality is a framework of how systematic oppression effects individuals with identities that are not considered by people in power (Collins & Bilge,

Intersectionality, 2020). Intersectionality is not only focused on how different identities interact, but how those identities can give or take away someone's power (Cho, Crenshaw, & McCall, 2013) (Collins, 2015). For example, black women are not oppressed because they exist as both black and a woman, but because their identities and experiences are shaped by the intersecting structure of sexism and racism (Kong, 2022). So, intersectionality does not only apply to categories of identities, such as race and gender, but also to the structural oppression and bias experienced from these identities, such as racism and sexism.

2.6.1 Debiasing on an Intersectional Axis is Difficult

Most algorithm debiasing work today, including in this thesis, requires the separation of attributes into "protected" and "non-protected" classes. For this thesis, we use race and the intersectional attributes of gender and ethnicity to debias CLIP. It would be remiss to overlook the reasons why intersectionality exists, why we are debiasing this model, and why we must break down intersectionality into categories of ethnicity and gender to quantify and measure bias in a technical context.

Intersectionality goes beyond the scope of how an individual's social identities interact to bring them different experiences. It is the result of a society run by racism and sexism. To effectively debias AI, we must acknowledge and change its authors and the society in which it exists. This is no small task, and it cannot be fixed with this thesis, but the background and reasoning as to why intersectionality is too complex to be encompassed in the debiasing process of an algorithm is important to acknowledge.

The dominant interpretation and protocol for debiasing and creating "fair" AI is to split identities (such as race and gender) into subgroups to categorize and measure a

model's bias (Kong, 2022). This thesis uses this approach, but when used generally, it fails to address and alter the source of bias; the system of oppression that currently exists through racism, sexism, or ableism. By working to remove bias from AI models, we may be fixing the issue in one context, but not implementing changes to remedy society's oppressive structure against minorities.

In their paper, Kong outlines 3 reasons as to why our approach which breaks down identities into categories is not ideal. 1) it overemphasizes on the intersection of protected attributes, 2) it is an inscription of fairness gerrymandering, and 3) it shows a narrow understanding of fairness as an equal distribution. Overall, Kong is arguing that instead of trying to "fix" biased algorithms, we need to use intersectionality to analyze and repair the intricate systems of oppression.

The way forward for debiasing CLIP in this thesis requires a need to show reduced bias in a quantifiable way using bias metrics and fairness criteria, which is categorized as "weak" fairness by Kong. "Strong" fairness, Kong states, would require using AI algorithms to challenge oppression and make society fairer as a whole. For this thesis, it is necessary to break down identities into subgroups for debiasing and measurement and this protocol does show a measured decrease in bias from original CLIP. Additionally, fairness as defined in this thesis requires distributions based upon three defined fairness metrics, to measure for different fairness interpretations. Although the debiasing in this thesis is considered "weak" fairness, it is still important to have an avenue to improve bias a pre-existing AI model.

While CLIP was initially used for pretraining DALL·E 2, which use cases vary: to create AI art, to edit images, accessibility to art creation. Unlike COMPAS (Kirkpatrick,

2017), CLIP is not being used to make decisions about people's future. It can, however, be used to create images that reinforce harmful stereotypes and inappropriate content. Ideally, CLIP and DALL•E 2 can be replaced by a system which is created by a collaborative team of AI researchers, computer scientists, philosophers, and civil rights activists to ensure that the creation and effects of these models are truly positive for society (Kong, 2022). By following a path of collaboration, the systems of oppression could be changed in a positive way. The way forward for fair AI should be through a collaborative effort across disciplines to improve fairness in society.

Large AI models, like CLIP, ChatGPT, and DALL·E 2 are expensive to train and finetune. A perfect model is not going to just appear overnight. To create an AI system that is purposefully helpful and projects a positive effect onto the world would be very expensive, so it is important to find a way to fix the harmful repercussions of a large, biased model. The proposed debiasing protocol and debiased models introduced in this thesis show a step forward in the effort to mend what is currently available and show a method where any skilled computer scientist or researcher can reduce bias in an AI model they do not own. While this thesis breaks down intersectionality into a narrow definition between ethnicity and gender, it was necessary for this experiment due to the available datasets and usage of CLIP for classification and DALL·E 2. To truly create a model that is fair and does not exhibit the flaws from society would be beyond the scope of a master's thesis and should instead be, a large collaborative effort between AI researchers, lawyers, philosophers, and civil rights activists (Kong, 2022).

CHAPTER 3

3 METHODOLOGY AND SETUP

This experiment aims to debias CLIP on race (CLIP-Race) and intersectionality (Intersectional CLIP), extending work done by Berg et al. which debiased CLIP on gender using their proposed debiasing protocol (Berg, et al., 2022). By using CLIP and an adversarial model, we jointly train to reduce similarity scores given by CLIP when fed an image and a vector of sensitive prompt tokens. CLIP works by taking in a labeled image, e.g., a facial image with a gender label and a vector of prompt tokens unrelated to physical characteristics or facial expressions. It then predicts how similar the image is to each word, producing similarity scores. The adversarial classifier takes the similarity scores as input and predicts the sensitive prompt token from the scores. Then, the predicted tokens are prepended to CLIP for use in the next round of training. By training CLIP and the classifier together, the similarity scores become equal, removing high similarity with harmful tokens unrelated to physical characteristics.

3.1 Datasets

This experiment uses two labeled face datasets (Balanced Faces in the Wild and FairFace) for training CLIP-Race and Intersectional CLIP. Both datasets are well-balanced between race and gender attributes and chosen to reduce dataset bias.

3.1.1 Balanced Faces in the Wild

Created to help mitigate bias in facial recognition systems, Balanced Faces in the Wild (Robinson, et al., 2020) is perfectly balanced between gender and ethnicity and contains a third gender/ethnicity label. With two gender, four race, and eight ethnicity/gender classes, BFW contains images of various human subjects sampled from VGG2 (Cao, Shen, Xie, Parkhi, & Ziserman, 2018). Each subgroup has 100 subjects, with over 25 face images for each subject, containing over a million images. The ethnicity labels for this dataset are as follows: Asian (A), Black (B), Indian (I), and White (W). With male (M) and female (F) labels, the eight ethnicity/gender are AM, AF, BM, BF, IM, IF, WM, and WF. This dataset is not split into training and testing sets, so the preparation protocol is described below in section 3.1.3.

3.1.2 FairFace

The FairFace dataset (Kärkkäinen & Joo, 2021) emphasizes balanced race composition, featuring seven race classes. Face images are collected from the YFCC-100M Flickr dataset (Thomee, et al., 2016) and contain White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino races. FairFace contains 108,501 images and is one of the first large-scale face attribute in-the-wild datasets to include Latino and Middle Eastern subjects and specify East and Southeast Asian individuals. Human annotators explicitly labeled race and gender, using the Individual Typology Angle for race annotations. FairFace also contains nine age groups, spanning from 0 to 70+. FairFace aims to reduce existing databases' limitations and biases by collecting a wide range of non-White face images. Because of its inclusivity and diversity, this dataset is ideal for use in this experiment and was also used by Berg et al.'s gender debiasing experiment.

Balanced Faces in the Wild				
FairFace				

Figure 4: Sample images taken directly from Balanced Faces in the Wild and FairFace datasets. All classes in BFW are perfectly balanced, unlike FairFace. All images in BFW are in color, while FairFace features some black and white facial images. FairFace consists of images from the Yahoo YFCC100M dataset, which were then adaptively adjusted based on the representation of each demographic.

Figure 4 contains sample images from both datasets. All classes in Balanced Faces in the Wild are perfectly balanced, unlike FairFace. FairFace features some black and white images, while BFW's images are all in color. FairFace is used in training the race debiased model, CLIP-Race due to the seven different labeled race classes. While BFW only labels four ethnicities, the special ethnicity/race label allows us to measure intersectionality, and is used for training the intersectionally debiased model, Intersectional-CLIP.

3.1.3 Data Preparation

Due to its large size of over 1 million images, Balanced Faces in the Wild is sampled with a training set of 80,000 and a testing set of 20,000, to mimic the size of FairFace's training and testing sets. We use python3 and Pandas to subsample BFW and create separate training and validation sets. The subset contains 12,500 instances of each of the eight subclasses in the dataset for a total of 100,000 images. Samples were randomly selected from the original dataset and manually checked to be sure the train/test split was correct.

3.2 Prompts and Concepts

Debiasing prompts and concepts are input with labeled images to CLIP to receive similarity scores between the image and prompts (also referred to as text/prompt tokens). A small amount of sensitive prompt tokens (training concepts) are input into a prompt or prompts. For example, with the prompt "a photo of a {}" and the token "criminal", CLIP will receive 'a photo of a criminal". In the same instance, CLIP receives a labeled facial image (labeled race if debiasing on race, ethnicity/gender if intersectional debiasing). CLIP computes how similar the image is to the prompts given and outputs a list of similarity scores equal to one. Below in Table 2 are the prompts and concepts used to fill them in for training and testing.

Table 2: The prompts and concepts used to train and test intersectional debiased CLIP. When training, one prompt is input to CLIP, along with ten training concepts. We feed these concepts to CLIP's text encoder, and a training image is given to the image encoder to produce a similarity score. The similarity score is given to the classifier, trained to reduce these scores. When testing, the testing prompts and concepts are put into CLIP and given similarity scores used to measure bias.

Training prompt	Training Concepts	Testing prompts	Testing Concepts
A photo of a {}	Good, evil, smart, dumb, attractive, unattractive, lawful, criminal, friendly, unfriendly	This is the face of a {} person, This is a photo of a {} person, This person is {}, A {} person, A {} individual, This individual is {}, Photo of a {}, {}, This is a {}, This is a photo of a {}, A {}	Dangerous, kind, successful, noncriminal, nonviolent, moral, rich, honest, heroic, villainous, unkind, unsuccessful, caring, violent, immoral, poor, dishonest, untrustworthy, clever, stupid, hardworking, lazy, nasty, trustworthy

The concepts above are from the Implicit Association Test (IAT) and were discussed previously in section 2.3. These tokens show no association with physical facial features and neither race nor gender. In an ideal, debiased model, CLIP will show the same similarity score for each token, as these words do not directly associate with a still image of a person.

3.3 Bias Measures

Berg et al. uses two metrics to measure bias: MaxSkew@k and NDKL (Geyik, Ambler, & Kenthapadi, 2019) (Berg, et al., 2022). Both are described in this section, but in testing our models, we found that NDKL does not accurately measure bias among many classes, unlike its performance in measuring gender bias in Berg et al.'s experiment. We report the NDKL measure of Berg et al.'s original measurements and our implementation of their model, but we exclusively use MaxSkew@k to test CLIP-Race and Intersectional CLIP due to the limitations of NDKL.

Both metrics rely on a distribution of results and a predetermined "ideal" result for bias measurement. Each metric uses three fairness criteria to determine the "ideal" result: equal opportunity, demographic parity, and differential fairness. See below for an in-depth description of MaxSkew@k and NDKL, along with the fairness metrics equal opportunity, demographic parity, and differential fairness.

3.3.1 MaxSkew@k

To accurately define MaxSkew@k, we must first define Skew@k (Berg, et al., 2022) (Geyik, Ambler, & Kenthapadi, 2019). Skew@k measures the difference between the desired proportion of image attributes in τ_T^k and the actual proportion. For example,

given the text query, “this person is a CEO”, the desired distribution of the image attribute for gender should be 50%, ensuring demographic parity.

Table 3: Key Notations for use in the definitions below

Notation	Represents
$\mathcal{A}$	$\mathcal{A} = \{A_1, \dots, A_l\}$ the set of attribute labels
T	Text query
I	Ranked list of images
τ_y	I according to their similarity to T
τ_T^k	The top k images of the list
$\mathcal{P}_{\tau_y,T,A}$	The actual proportion of results
$\mathcal{P}_{d,T,A}$	The desired proportion of results

Refer to Table 3 above for the notations used below.

$$Skew_{a_i}@k(\tau_y) = \ln \frac{\mathcal{P}_{\tau_y,T,A}}{\mathcal{P}_{d,T,A}}$$

This measurement gives some indication of representational bias if specific attributes are under-represented in the top k results (a negative $Skew_A@k$). However, due to only measuring the bias of a single attribute at a time, it must be aggregated to show a more holistic view of bias over all attributes A. $Skew_{a_i}@k$ also reflects different results with different values of k. While $Skew_{a_i}@k$ provides a total skew measure, MaxSkew@k is better suited for our experiment because it shows the maximum $Skew@k$ among all A attribute labels of the images for the text query T.

$$MaxSkew@k(\tau_y) = \max_{A_i \in \mathcal{A}} Skew_{A_i}@k(\tau_y)$$

This measure shows the “largest unfair advantage” (Geyik, Ambler, & Kenthapadi, 2019) belonging to images within the given attribute. The desired outcome is 0, implying that the real and desired distributions are equal. For example, with our “this person is a CEO”

description, a measurement of 0 would reflect an equal number of male and female examples.

3.3.2 Normalized Discounted cumulative KL Divergence

NDKL, normalized discounted cumulative Kullback-Leibler divergence, uses Kullback-Leibler divergence to measure how much one distribution varies from another. This is a non-negative number, with the ideal measure being 0. Larger values indicate a greater divergence between the desired and actual distributions of $\mathcal{A}$ for a given T. Let $D_{\tau_T^i}$ and D_T denote the discrete distribution of image attributes in τ_T^i and the desired distribution, respectively. The definition of NDKL is:

$$NDKL(\tau_y) = \frac{1}{Z}\sum_{i=1}^{|\tau_y|}\frac{1}{\log_2(i+1)}d_{KL}(D_{\tau_T^i}||D_T)$$

where $d_{KL}(D_1||D_2) = \sum_j D_1 \ln\frac{D_1(j)}{D_2(j)}$ is the KL-divergence of distribution D_1 with respect to distribution D_2. $Z = \sum_{i=1}^{|\tau_y|}\frac{1}{\log_2(i+1)}$ is a normalization factor. Using the top-k distribution and the desired distribution to find the KL-divergence will be the weighted average of $Skew_A@k$ (averaged over $A \in \mathcal{A}$), so NDKL solves one of $Skew_A@k$'s disadvantages by presenting a holistic view of overall bias. However, NDKL has the disadvantage of not being able to distinguish between two "opposite biased" (Berg, et al., 2022) procedures due to it being non-negative.

3.4 Fairness Criteria

The bias measures mentioned above discuss comparing the real distribution of results with an "ideal" distribution to show a measure of bias. This experiment uses three fairness criteria to dictate the "ideal distribution" for each bias measure: demographic

parity, equal opportunity, and differential fairness. These criteria are used to provide ideal distributions specifically in machine learning models and algorithms and have been used for similar experiments such as this one (Berg, et al., 2022) (Geyik, Ambler, & Kenthapadi, 2019). Additionally, these criteria are discussed in theoretical literature when it comes to bias and ethics. Descriptions of how these criteria are calculated are below in Table 4.

3.4.1 Demographic Parity

Demographic parity requires that any decision is made without consideration of a protected attribute and that demographics (such as gender, race) are classified to the same proportion among all classes (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012). For example, when asking DALL·E, "show me a picture of a CEO", the results must have an equal amount of female and male images to achieve demographic parity. For two genders, the desired distribution is $\tau_y = [0.5,0.5]$ for all race attribute labels $A \in \mathcal{A}$.

3.4.2 Equal Opportunity

Equal opportunity requires that non-discrimination only happens in the "advantaged" outcome group (Hardt, Price, & Srebro, 2016). For example, when deciding who is going to get a scholarship in a group of college applicants. The "advantaged" group is the students who earn a scholarship, and the "disadvantaged" group are those who will not get the scholarship awarded. Equal opportunity requires that all applicants who earn a scholarship are chosen without discrimination (not based on race, gender, or any protected attribute). This notion of fairness differs from demographic parity, as it considers protected classes in the outcome of results to ensure no discrimination occurs in the positive outcomes. Equal opportunity allows for more substantial utility in machine learning problems by considering the protected classes only in the positive outcomes.

3.4.3 Empirical Differential Fairness

Differential fairness considers multiple protected attributes, rather than a single attribute, and is motivated by intersectionality (Islam, Keya, Pan, Sarwate, & Foulds, 2023). At its core, differential fairness requires that regardless of the protected attribute of an individual, the outcome will be the same based on other factors. Empirical differential fairness requires the same outcome, regardless of the combination of protected attributes, ensuring fairness to anyone belonging to any attribute due to its empirical nature. Empirical differential fairness is used as a measure in this thesis, but it is important to include that the calculation is very similar to demographic parity. By multiplying the number of instances of a class by a coefficient *e*, it is intended to provide a more holistic fairness measure. See Table 4.

3.4.4 Measuring Bias

The fairness criteria mentioned above are used in the bias metrics to calculate the amount of bias present in debiased models. Both metrics require an "ideal" distribution and the real distribution; the former is dictated by the fairness criteria. Table 4 describes how the desired distribution is calculated for each criterion.

Table 4: How the ideal distribution is calculated for each fairness criteria. Equal Opportunity requires representation of all classes to be equal. Demographic Parity considers the number of instances of a class divided by the total instances. Differential Fairness is calculated similarly to demographic parity, except for the *e* value.

Equal Opportunity	Demographic Parity	Differential Fairness
$\frac{1}{\#\text{ of classes}}$	$\frac{\#\text{ of instances of a class}}{\text{total instances}}$	$\frac{e\times\#\text{ of instances of a class}}{\text{total instances}}, e = 0.4$

3.5 Implementation

For exploratory experiments, we used python3 on a MacBook Pro with 2.3 GHz 8-Core Intel Core i9, AMD Radeon Pro 5500M 4 GB, Intel UHD Graphics 630 1536 MB, 16 GB 2667 MHz DDR4 RAM. We access CLIP through OpenAI's API and Debias-CLIP-gender from the Oxford AI Society's GitHub[1]. We use FairFace and Balanced Faces in the Wild datasets in the debiasing protocol and for testing. For CLIP-Race and Intersectional CLIP, we use a Linux workstation with an Intel i9-10980XE CPU, Nvidia RTX A5000 GPU, 256GB DDR4 RAM, and a 1 TB SSD. For the code used to train, test, and evaluate bias in this thesis, see Appendix C.

3.6 Debiasing Protocol

Debiasing CLIP ViT-B/16 in this experiment uses an adversarial classifier that takes CLIP's similarity score output to predict the photo's label (e.g., race or gender) from its similarity scores (see Figure 3). The classifier then prepends learnable prompt tokens, determined by the predicted attribute label, to the inputs of CLIP's text encoder. We used joint training on both CLIP and the adversary. With a batch size of 256, the adversary is trained for two epochs initially, and then training alternates between each model every two batches. The training lasts for ten epochs total and takes approximately 24 GPU hours. Testing is performed only on the debiased CLIP model.

[1] https://github.com/oxai/debias-vision-lang

Table 5: An overview of the methodology and experiments performed in this thesis.

Summary of Methodology and Experiments		
Datasets	FairFace	CLIP-Race
		Size: 108,501 images, pre-defined train/test split
	Balanced Faces in the Wild	Intersectional CLIP
		Size: subsampled 100,000 images, partitioned 80/20 train/test split
Evaluation Methods	NDKL	Only used for initial exploratory experiments
		Uses demographic parity and equal opportunity fairness metrics
	MaxSkew@1000	Used in all experiments
		Uses demographic parity, equal opportunity, and differential fairness as fairness metrics
Debiasing	Adversarial Classifier	First trained for 2 epochs, then alternates every 2 epochs between CLIP
	# of epochs	10
	Batch Size	256
	Total Training Time	24 GPU hours
Experiments	Exploratory	Reported and re-implemented the debias-CLIP-gender model (Berg, et al., 2022)
		Reported bias in our re-implementation using both FairFace and Balanced Faces in the Wild datasets
	CLIP-Race	1 experiment: trained CLIP-Race and tested on the FairFace dataset
	Intersectional CLIP	2 experiments: trained Intersectional CLIP and tested on 2 versions of Balanced Faces in the Wild dataset. The first results show bias measured with a balanced testing set, and the second show biased measured with an unbalanced testing set.
	Similarity Scores	Shows reduced bias in a quantitative way, specifically with Intersectional CLIP and the different categories of ethnicity/gender
		Shows scores when a physical attribute, such as “smiling” is incorporated in the similarity score calculations
	Linear Probe	Compares feature extraction capabilities between CLIP ViT-B/16, CLIP-Race, and Intersectional CLIP

CHAPTER 4

4 RESULTS AND ANALYSIS

This thesis consists of four experiments: two for exploratory purposes and two to test the proposed CLIP-Race and Intersectional CLIP. The exploration consists of re-implementing the experiment proposed in Berg et al. of debiasing on gender by testing CLIP ViT-B/16 (Radford, et al., 2021) and Debias-CLIP-gender (Berg, et al., 2022) on FairFace (Kärkkäinen & Joo, 2021) and Balanced Faces in the Wild (Robinson, et al., 2020) datasets. Testing CLIP-Race and Intersectional CLIP follow, showing results of our debiasing protocol on the axes of race and the intersectionality of ethnicity and gender. Section 4.1 features results from the original experiment by Berg et al. and our re-implementations. Sections 4.2 and 4.3 discuss the results of CLIP-Race and Intersectional CLIP, respectively. Finally, we perform auxiliary experiments to show the effect of these debiasing experiments on the efficacy of the CLIP.

4.1 Exploratory Experiments: Debias-CLIP-Gender

The results reported in Berg et al. are presented in Table 6. The authors do not specify which fairness metrics were used when determining the value of MaxSkew and NDKL. These compare an original CLIP ViT-B/16 (Radford, et al., 2021) model with their debiased model Debias-CLIP-gender (Berg, et al., 2022).

Table 6: Original results from the first CLIP-debias experiment (Berg, et al., 2022). This debiased version of CLIP was trained and tested on the FairFace Dataset. It was unclear in this original experiment which notion of fairness they used during the measurement of MaxSkew and NDKL, so we went more in-depth with later experiments. The debiased results show a reduction compared to the original CLIP ViT-B/16 (Radford, et al., 2021), directly from OpenAI.

Reported Results of Debias-CLIP-gender (Berg, et al., 2022)		
Model	MaxSkew@1000	NDKL
CLIP ViT-B/16 (Radford, et al., 2021)	0.233	0.104
Debias-CLIP-gender (Berg, et al., 2022)	0.073 (-69%)	0.021 (-80%)

Table 7: Shows the results of the baseline experiment as seen in Berg et al. CLIP is accessed directly from OpenAI, and Debias-CLIP is from Berg et al. The measurements presented in this table are similar to the reported measures in Berg et al. but differ because the code used gave incorrect measurements. This table shows a decrease in bias after debiasing CLIP on gender but an increase in race bias, unlike the initially reported results.

Implementation of Debias-CLIP-gender Testing on FairFace					
Model	Bias measure	Measured Bias			
		MaxSkew@1000		NDKL	
		Dem. Parity	Equal Opp.	Dem. Parity	Equal. Opp.
CLIP-ViT-B/16 (Radford, et al., 2021)	race	0.4242	0.5408	0.0387	0.0558
	gender	0.1915	0.1915	0.0175	0.0175
Debias-CLIP-gender (Berg, et al., 2022)	race	0.6622 (+56.11%)	0.4507 (-16.66%)	0.0410 (+5.94%)	0.0663 (+18.82%)
	gender	0.0936 (-51.12%)	0.0936 (-51.12%)	0.0062 (-64.57%)	0.0062 (-64.57%)

Table 8: Results from measuring bias with the Balanced Faces in the Wild validation set. This set is different from FairFace, as each image is labeled with ethnicity, gender, and a label that has both ethnicity and gender. It is also perfectly balanced between 4 races, as shown in Figure 4 and discussed in section 3.1.1.

Implementation of Debias-CLIP-gender Testing on Balanced Faces in the Wild					
Model	Bias Measure	Measured Bias			
		MaxSkew@1000		NDKL	
		Dem. Parity	Equal Opp.	Dem. Parity	Equal Opp.
CLIP-ViT-B/16 (Radford, et al., 2021)	ethnicity	0.5554	0.5397	0.0557	0.0531
	gender	0.2591	0.2591	0.0252	0.0252
	ethnicity and gender	0.8009	0.7859	0.0851	0.0823
Debias-CLIP-gender (Berg, et al., 2022)	ethnicity	0.4272 (-23.08%)	0.4108 (-23.46%)	0.0347 (-37.71%)	0.0324 (-38.98%)
	gender	0.1819 (-29.80%)	0.1819 (-29.80%)	0.0104 (-58.73%)	0.0104 (-58.73%)
	ethnicity and gender	0.6294 (-21.41%)	0.6123 (-22.09%)	0.0488 (-42.66%)	0.0465 (-43.50%)

Upon implementation and testing, the results found were different than what is reported (Berg, et al., 2022). In Table 7, we report a 51% decrease in gender bias when measured with MaxSkew@1000, while Berg et al. reports a 67% reduction. Comparatively, for the NDKL measure, we report a 64% decrease in bias, with Berg et al. showing a high 80% decrease.

Table 8 presents the same implementation of CLIP ViT-B/16 and CLIP-Race tested on Balanced Faces in the Wild. In addition to ethnicity and gender bias, we include an intersectional variable of ethnicity and gender, defined in BFW. These results support the theory of reduced racial bias when debiasing on gender. MaxSkew@1000 shows a steady decrease in bias of about 20% across fairness metrics, while NDKL varies between a 30% to high 50% decreases for all fairness metrics. Although testing on Balanced Faces in the Wild shows a decrease in bias outside of gender, the variation in measures is a cause for concern.

Berg et al. alludes to an additional decrease in racial bias when CLIP is debiased on gender alone; a hypothesis in which this thesis is based on. However, in the re-implementation in Table 7, the results for race have a high variation of bias reduction. Noticeably, the NDKL results show an increase in racial bias for the FairFace implementation. NDKL issue- why it is getting put in an appendix. The inconclusiveness of racial bias reduction in the FairFace implementation led to testing on Balanced Faces in the Wild, a dataset with an intersectional approach.

When calculating MaxSkew and NDKL, we used publicly available evaluation code posted by Berg et al. Originally the metrics were returning inaccurate results, failing to represent MaxSkew correctly. Initially, the equation returned a value of 0.0 due to

comparing all results instead of the top 1000. A variable, "top_n", was changed to 1000, and the issue subsided, making our results comparable to what was originally presented.

These explorative experiments provide support for the use of FairFace for CLIP-Race and Balanced Faces in the Wild for Intersectional CLIP debiasing protocols. Although NDKL is used in Berg et al. to report gender bias, this measure will be reported in the Appendix B, due to this measure being unfit for a large number of variables. NDKL presents bias in a way that makes the measure indistinguishable between bias of equal but opposite directions. With CLIP-Race consisting of seven race variables, and Intersectional CLIP containing eight ethnicity/gender variables, the outcome reflects the flaws of this metric.

4.1.1 NDKL is not Ideal for Measuring Many Classes

We use two bias metrics in this experiment: MaxSkew@1000 and NDKL. When considering the results of both metrics, MaxSkew showed a consistent decrease in bias for CLIP-Race and Intersectional CLIP. NDKL, however, did not, and mostly showed an increase in race and ethnicity/gender bias while reporting a decrease in gender bias regardless. This discrepancy suggests that for an experiment such as this, where we measure bias among 7-8 attributes, such as the seven race classes and eight ethnicity/gender classes, NDKL cannot provide an interpretable measure.

In past experiments, NDKL was used to measure gender bias among two genders (Berg, et al., 2022). It successfully reduced gender bias, even as reported in the exploratory experiments presented in this thesis. NDKL is useful when measuring gender because there are only two classes, and minor discrepancies are less significant, with an ideal proportion of 50/50. When measuring for race, more variables make the ideal proportion much

smaller, requiring each of 7 race classes to make up about 14% of the results. Due to the small size, changes in the actual distribution reflect through NDKL at a higher rate. For example, in the seven race classes in the FairFace dataset, if there is a decreased presence of results containing White individuals but an increase in other races such as East or South Asian, NDKL cannot interpret the change into a number accurately reflecting it due to its non-negative nature. Since there is no differentiation between negative and positive bias measures, an increased NDKL can occur even if bias decreases. MaxSkew@1000 was able to show a decrease in bias because instead of measuring the divergence between two distributions, it measures the skew.

For example, imagine a synthetic dataset which contains labeled images of cats such as tabby, black, orange, and Siamese. If we apply the circumstances of debiasing in this thesis on cats, each type of cat should have an equal similarity to words that do not relate to the physical appearance of a cat. So, ideally each word would show an equal measure of similarity to tabby, black, orange, and Siamese felines by 25%. For an unideal distribution, such as 30% tabby, 10% black, 30% orange, and 30% Siamese, NDKL is not able to portray the imbalance of bias among the four classes of cats well, due to it being non-negative and a larger disturbance in the distribution. If we were measuring between only tabby and black cats, where there would be an ideal measure of 50% for each word, a 5% increase or decrease would not be reflected as intensely by NDKL as it would when measuring among 4 classes of cats.

4.2 CLIP-Race

CLIP ViT-B/16 is debiased on race, using the protocol shown in Figure 3, labeled images from FairFace's seven race classes, and the testing prompts presented in Table 2.

FairFace's training set is used for debiasing, and the validation set is used for testing bias. The results are shown below in Table 9 where the MaxSkew@1000 measure is presented with three fairness criteria. See Appendix A for full results including NDKL.

Table 9: Results from the debiasing protocol of race on CLIP. These results used MaxSkew@1000 and NDKL to evaluate with three fairness criteria: demographic parity (DP), equal opportunity (EO), and differential fairness (DF). Results show a decrease in racial bias overall, with the most being MaxSkew@1000 measure using differential fairness and the criteria of demographic parity of the NDKL measure. Gender bias was also measured, and there was a decrease in gender bias across the board. The outlier measure of differential fairness with MaxSkew@1000 of 0.0 could be because either the original measure was already tiny or the balance of the two genders was almost perfect, so it did not get registered by the measurement criteria.

CLIP-Race debiased on FairFace				
Model	Bias Measure	MaxSkew@1000		
		Equal Opportunity	Demographic Parity	Differential Fairness
CLIP ViT-B/16 (Radford, et al., 2021)	Race	0.5428	0.4328	0.0728
	Gender	0.1916	0.1916	0.0039
CLIP-Race	Race	**0.5721 +5%**	**0.2915 -33%**	**0.0013 -98%**
	Gender	0.1416 -26%	0.1416 -26%	0.0

CLIP-Race reduces the presence of racial bias, although there is significant variation between fairness metrics. Each fairness metric, e.g., demographic parity, equal opportunity, and differential fairness is measured according to the mathematical equations presented previously in section 3.4. In the results above, the fairness criteria of equal opportunity shows a slight increase in racial bias. Equal opportunity measures how often each race occurs in the results compared to all total occurrences. The increase in measured bias reflects that not all races are equally represented in the results, disadvantaging individuals of some races. Because these metrics are aggregated, knowing which race is more affected is difficult. There are also seven race classes, so any slight differences in the distribution of each race reflect highly, more so than in the gender measurement.

Notably, racial bias in CLIP ViT-B/16 is much higher in every measure than gender bias. The elevated racial bias in CLIP ViT-B/16 can prove more difficult to reduce than gender bias. Although CLIP was debiased on race, there was a more significant reduction in gender bias. A similar instance happened in Berg et al., where they reported that Debias-CLIP-gender showed a reduction in racial bias (Berg, et al., 2022). CLIP-Race, however showed a less significant reduction in gender bias, although the occurrence supports the hypothesis of debiasing on one axis, such as race, will additionally reduce gender bias. These results reflect the hypothesis that debiasing on one aspect, race, can also reduce bias in another protected class, such as gender. While the measured reduction in race is less than ideal for a model such as this, the implication of reduced bias in other areas is promising.

4.3 Intersectional-CLIP

We debias CLIP on the intersection of ethnicity and gender using labeled images from Balanced Faces in the Wild with the same debiasing protocol as CLIP-Race. The results below in Table 10 show bias present in CLIP ViT-B/16 and Intersectional CLIP when measured on a balanced subset of the BFW validation set.

Table 10: Results from debiasing CLIP by training on the Balanced Faces in the Wild dataset. These measures show a decrease in the Ethnicity/Gender classes, which are eight classes encompassing ethnicity and gender. The decrease happens explicitly in the MaxSkew@1000 measure, while the NDKL slightly increases. The increase could be from several reasons, but having one way of encompassing the complex concept of intersectionality is almost impossible.

Intersectional CLIP Bias Testing with Balanced BFW				
Model	Bias Measure	MaxSkew@1000		
		Equal Opportunity	Demographic Parity	Differential Fairness
CLIP ViT-B/16 (Radford, et al., 2021)	Ethnicity	0.4893	0.4893	0.1204
	Gender	0.2334	0.2334	0.0077
	Ethnicity/Gender	0.7195	0.7195	0.3214
Intersectional-CLIP	Ethnicity	0.4939 +1%	0.4939 +1%	0.1033 -14%
	Gender	0.0915 -61%	0.0915 -61%	0.0
	Ethnicity/Gender	**0.5989 -17%**	**0.5989 -17%**	**0.1993 -38%**

Measuring the presence of intersectional ethnicity/gender bias showed a 17% reduction, with a greater reduction of 61% for gender bias and a minimal increase in the presence of ethnicity bias. Upon investigation of these results, we found that the small reduction lies in the mathematical nature of the metrics. The FairFace testing set is unbalanced between the classes, so the metrics reflect the ideal results as such causing the equal opportunity and demographic parity measurements to be identical. The subset of data taken from BFW was balanced, so it was necessary to create an unbalanced testing set and run the measurements again for improved results.

4.3.1 Testing Intersectional CLIP on an Unbalanced Dataset

The first test of Intersectional CLIP uses a testing set with 1,386 occurrences of each ethnicity/gender class. Ideally, the makeup of a dataset should not matter, but NDKL and MaxSkew@1000 rely on the dataset to generate the ideal distribution. Due to the nature of these metrics, the results can be skewed and show more bias than in real-world

applications by using an unrealistic dataset for testing. By using a testing set that is not perfectly balanced between ethnicity and gender, the results improve and show a reduction in ethnicity and gender bias when measured using MaxSkew@1000.

Table 11: Results of testing Intersectional-CLIP on an unbalanced test set. With these results, the symmetric measurements were replaced by better, if not more consistent, results, removing the strong influence the validation set has on the presented results.

Intersectional CLIP Bias Testing with Unbalanced BFW				
Model	Bias Measure	MaxSkew@1000		
		Equal Opportunity	Demographic Parity	Differential Fairness
CLIP ViT-B/16 (Radford, et al., 2021)	Ethnicity	0.5286	0.5287	0.1493
	Gender	0.2082	0.2134	0.0062
	Ethnicity/Gender	0.7322	0.7296	0.3317
Intersectional-CLIP	Ethnicity	0.5099 -4%	0.4854 -8%	0.0940 -37%
	Gender	0.0599 -71%	0.0608 -72%	0.0
	Ethnicity/Gender	0.5920 -19%	0.5626 -23%	0.1633 -51%

After testing with an unbalanced dataset, the MaxSkew@1000 measure showed a decrease in bias of up to 51%. Gender bias decreased in both MaxSkew@1000 and NDKL measures of up to 80%. The change in testing data to include different-sized samples of each ethnicity/gender caused some scores to improve and show a decrease in bias, especially MaxSkew@1000. NDKL, on the other hand, did not decrease and showed an increase in measurement for both CLIP-Race and Intersectional CLIP. The measured increase reflects that NDKL may not be ideal for measuring bias reduction for cases with more classes, such as race and ethnicity/gender. See section 4.1.1 for more insight on NDKL as a metric for this experiment.

4.3.2 Accuracy of Debiased Models: Linear Probe Evaluation

To test the functionality of the debiased models, we perform a linear probe evaluation to show an accuracy comparison. Linear probe evaluations are helpful when assessing the quality of learned features within a neural network and the overall transferability of features to different tasks without requiring further in-depth model modifications. Linear probes are widely used in testing visual representation tasks (Asano, Rupprecht, & Vedaldi, 2020) (Radford, et al., 2021) on pretraining tasks. Linear probes work by introducing a linear classifier, in this case, logistic regression, on top of the extracted features and training to perform a classification task. It is more effective than other strategies at comparing the retained feature extraction capabilities of CLIP with the debiased models (Radford, et al., 2021). The linear probe results are shown below in Table 12.

Table 12: Shows the results of a linear probe evaluation for CLIP ViT-B/16, CLIP-Race, and Intersectional CLIP. The linear probe uses logistic regression on features extracted from the specific CLIP model to classify testing images using the CIFAR-100 dataset. The accuracy scores show that the feature extraction capabilities from CLIP ViT-B/16 are not lost in the debiasing process.

Model	**CIFAR-100**	**ImageNet-1K**
CLIP ViT-B/16	79.99%	67.88%
CLIP-Race	82.51%	73.53%
Intersectional CLIP	82.56%	73.55%

This linear probe uses logistic regression and a CLIP model to classify the CIFAR-100 dataset (Krizhevsky, 2009) and a portion of the ImageNet1k dataset (Russakovsky, et al., 2015). Using extracted features in a linear probe evaluation, we test CLIP ViT-B/16, CLIP-Race, and Intersectional CLIP. The logistic regression model trains for a maximum of 1000 iterations with a regularization strength of 0.316. The results of this probe show that the feature extraction and classification capabilities of CLIP are not lost in the debiasing process. The accuracy of baseline CLIP ViT-B/16 is comparable on both datasets

as tested in the paper by Radford et al. There is also a noted increase in accuracy, which can be due to the further training CLIP receives in the debiasing protocol.

4.4 Discussion

The results presented in this thesis showed a reduction of bias for most measurements. In CLIP-Race, MaxSkew showed a reduction in racial and gender bias, except for equal opportunity. This increase is due to equal opportunity requiring that the presence of each class is equal in the positive results. For the ideal distribution, each class should make up about 14%. This small percentage is more impacted by any change in distribution when compared with the measurement of gender. Any outliers or overrepresentation could account for this increase in measure. For both models, NDKL showed an increase in bias for almost all results except gender bias measurements.

4.4.1 Similarity Scores

CLIP works by taking in an image and a set of text queries and outputting a similarity score for each query. Similarity scores are presented as percentages; each score added together equals one. Considering these scores as one entity is essential to interpretation.

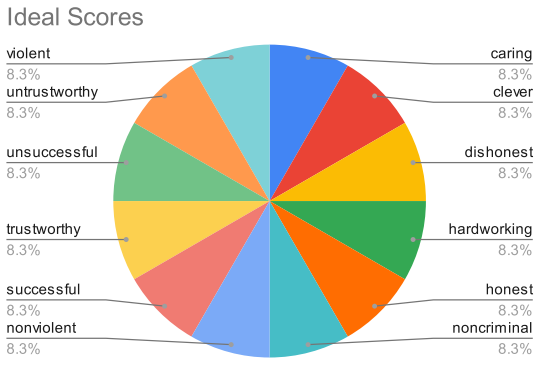


Figure 5: The ideal distribution of Intersectional CLIP similarity scores. When debiasing, the ideal scores for each class are to be equal, if and only if the class describes a non-physical characteristic.

We train the debiased models to adjust the scores to be equal for words that do not correlate with physical attributes, such as "lazy", "criminal", or "violent", as none are more related to an image of a face than another. These are referred to as sensitive text queries. Figure 5 shows the ideal distribution of the 12 similarity scores and reflect what the debiasing protocol is attempting to achieve. Figure 6 below shows the similarity scores output by CLIP ViT-B/16 and Intersectional CLIP when both receive 12 sensitive text queries and batches of 50 images containing one BFW ethnicity/race class.

For test cases, Intersectional CLIP receives 12 text queries and 50 images, each of eight classes: Black Men (BM), Black Women (BW), White Men (WM), White Women (WW), Indian Men (IM), Indian Women (IW), Asian Men (AM), and Asian Women (AW). The similarity scores output by CLIP ViT-B/16 and Intersectional CLIP are above in Figure 6 (charts for all classes are in Appendix B). The y-axis shows the sensitive text tokens used in the prompt template "a photo of a {} person" given to both models, and the similarity score percentage is in the x-axis.

CLIP ViT-B/16 presents a wide range of similarity scores when classifying all ethnicity/gender classes. All cases presented above show the tokens "untrustworthy" and "noncriminal" with high scores when scored by CLIP ViT-B/16. "Untrustworthy" has an inherently negative meaning, and the very high score, specifically for women, by CLIP ViT-B/16 reflects learned biases associated with the model's training. In a flawed, biased model such as CLIP ViT-B/16, reinforcing data to algorithm and user-to-data bias can be harmful.

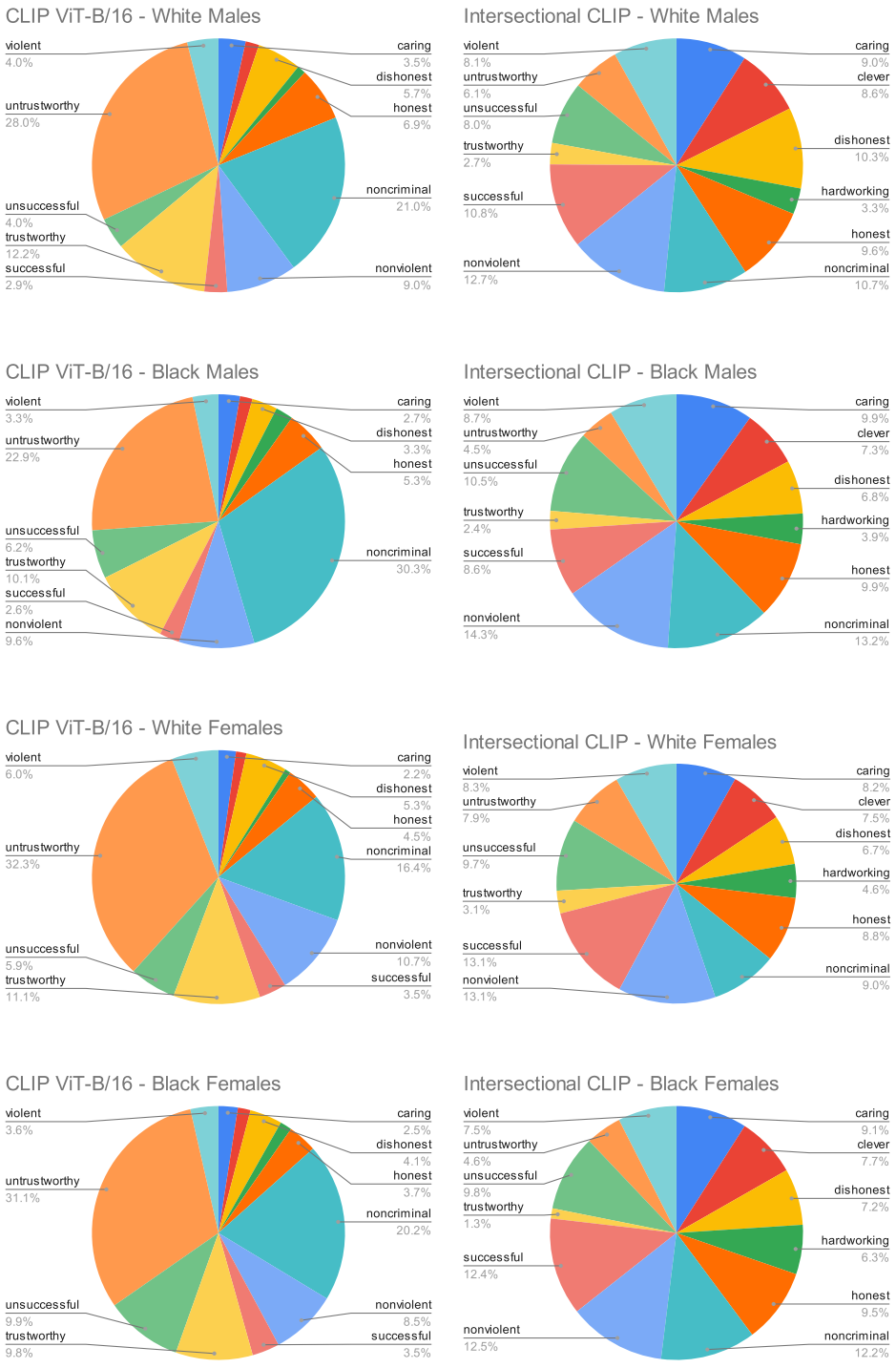


Figure 6: Similarity scores for twelve sensitive text queries when CLIP ViT-B/16 and Intersectional CLIP are given batches of 50 images for each gender/ethnicity class. This figure includes scores for White and Black males and females; the other classes are in Appendix B. Intersectional CLIP reduces high scores given by CLIP ViT-B/16 and adjusts the scores to be more even, removing significant associations to text that does not represent physical attributes.

Intersectional CLIP presents more equal similarity scores with strong associations to any sensitive tokens reduced closer to the "ideal" score of 8%. Ideally, if the debiased model produces all similarity scores at the same value for each of the 12 tokens, each token is assigned a score of 1/12. Intersectional CLIP reduces high similarity scores, with the highest score of 13% for White females and 12% for black females. Although not all similarity scores adjusted to the same value, they are closer than CLIP ViT-B/16's output, with high similarity scores of upwards of 30% for both White and Black females. "Untrustworthy" was the highest similarity score from CLIP ViT-B/16 for females in both race groups, and it was reduced to scores of 7% and 4% by Intersectional CLIP. Scores above the ideal score include tokens such as "noncriminal", "nonviolent", and "successful", with scores around 12% for both female classes. Although these scores are above the "ideal" measure, the higher association with non-harmful tokens is positive.

White and Black males show different trends in similarity scores. CLIP ViT-B/16 returns the highest similarity score for Black males as "noncriminal" at 30% similarity but only at 21% for White males. This token, although positive, paired with a higher score for Black males and other male classes of color (see Appendix B), possibly reflects the debiasing done by OpenAI (see section 1.3). They did not release their debiasing protocol in their report but did implement safeguards in CLIP and the DALL·E system. Similarity for "untrustworthy" is reduced from ~30% to 6% for White males and 4% for Black males. "Noncriminal" was also reduced to 10% for Black males and 13% for White males.

For all examples, some scores from CLIP ViT-B/16 that have a small similarity were augmented in Intersectional CLIP such as “caring”, “successful”, and “violent”. Regardless of the semantics of these tokens, the increase is due to the model learning to assign all

tokens that relate to a non-physical characteristic the same value. For example, CLIP is a pretraining model for DALL·E. When asked to generate images, DALL·E uses CLIP to classify the segments of the photo used for generation. Assume DALL·E is using Intersectional CLIP. If the user asked for “a photo of a violent person”, assuming there are no safeguards in place, DALL·E would output images representing diverse genders and ethnicities because being “violent” should not be associated with any ethnicity, race, or gender. Any word that cannot be physically seen in an image should represent a diverse population of results, as removing bias would remove any associations with a protected class. This example is extreme but necessary for perspective.

4.4.2 Identifying Similarity Scores for Physical Characteristics

The evidence presented above shows the effects of debiasing CLIP ViT-B/16 and how that reduced bias in the model's output. To ensure CLIP ViT-B/16 retains its ability to recognize physical features, we experiment with a dataset of 50 images containing various ethnicities and genders from the FairFace dataset. Using the twelve sensitive text tokens from Figure 6 and one physical token, "smiling", we show how CLIP ViT-B/16, Intersectional CLIP, and CLIP-Race assign similarity scores when given text input correlating to a physical descriptor.

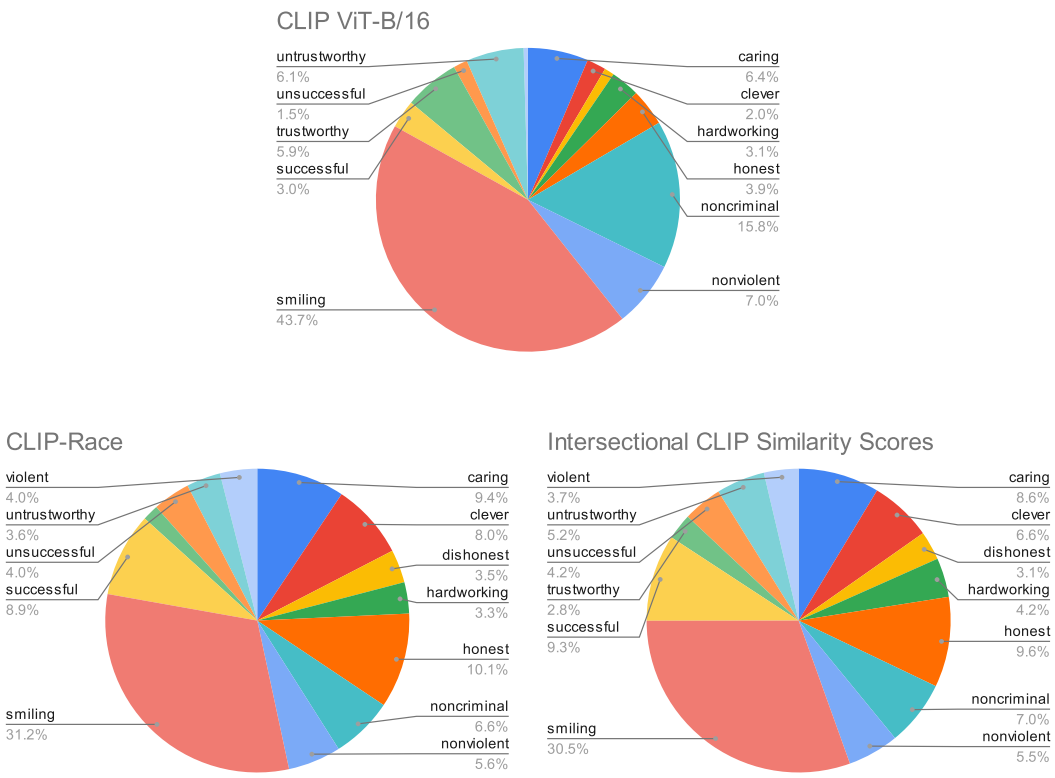


Figure 7: The similarity scores when CLIP ViT-B/16, Intersectional CLIP, and CLIP-Race receive sensitive text tokens and a physical token "smiling". These scores are aggregated for 50 facial images from FairFace. While the score for "smiling" does decrease in the debiased models, the scores for the other sensitive tokens become more even, fulfilling the purpose of the model.

Figure 7 shows thirteen text tokens' similarity scores when CLIP ViT-B/16, Intersectional CLIP, and CLIP-Race process them. The scores are aggregated among the 50 facial images used. CLIP ViT-B/16 shows a similarity score of 44% for "smiling", while Intersectional CLIP and CLIP-Race show a decrease with scores of ~31%. "Smiling" decreased due to other scores, such as "successful" and "honest", increasing to meet the "ideal" similarity of ~7%. Because "smiling" has a physical representation recognized by the debiased models, it receives a similarity score over four times higher than the "ideal" score. The reduction in the "smiling" score from CLIP ViT-B/16 reflects the other scores

increasing to the "ideal" score so that the non-physical tokens become equal, subtracting from the large scores "smiling" and "noncriminal".

Additionally, we test for similar physical tokens. “Grinning” produces results similar to smiling with the similarity score for “grinning” slightly increasing from original CLIP. This finding is extended to the term “smirking”, with an increase in similarity from original CLIP. Although the facial expressions relating to smiling are easily recognized by all three models, this notion is not found when measuring for other physical characteristics. For example, two datasets of 20 images were crafted: one containing people wearing hats, the other made of people wearing glasses.

For the subgroups of people wearing hats, CLIP ViT-B/16 assigns a 63% similarity to the set. Our debiased models, however, show a drastic decrease in similarity, with both debiased models assigning a 31% similarity score. This decrease in “wearing hats” similarity can be attributed to an increase in other positive terms such as “successful”, “caring”, and “honest”. Although the debiased models assigned a smaller similarity score, the physical value is still the largest. In a classification task, the image would still be correctly classified. We had the same results in the classification of “wearing hats”. These results portray that while the debiased models can still be used for classification, as shown in our linear probe evaluation, the debiasing process can reduce the magnitude of similarity scores when it comes to classifying images such as these that have many classification labels describing non-physical attributes.

CHAPTER 5

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

The proposed CLIP-Race and Intersectional CLIP presented showed a bias reduction in most cases. When measuring racial bias in CLIP-Race, there was between 33-98% reduction when using MaxSkew@1000, a more robust measure of bias. Although race was the attribute being debiased, there was also a significant reduction in measured gender bias, up to 78%. These findings support Berg et al. by showing that debiasing one attribute, such as race, can also reduce bias in other protected classes. Intersectional CLIP showed up to a 51% reduction in combined race and gender bias when using MaxSkew@1000 as a measure. Similar to CLIP-Race, there was also reduced bias in the separate categories of race and gender, reinforcing the hypothesis that by debiasing on one protected class, bias should also be reduced on other protected classes.

NDKL was an included measure of bias in these experiments, but it did not present results that correlated with the measure MaxSkew presented. By reflecting an increase in bias in most cases, NDKL presented as a weaker metric for bias measurement. NDKL is inefficient at measuring bias for cases where there are many classes, like race, due to it not differentiating between bias of equal but opposite extent. Because MaxSkew@1000 measures the maximum skew of the top 1,000 results, it reflects these changes in bias in a better fashion.

In addition to the measures of bias, experiments showing CLIP's effectiveness after debiasing support the notion that CLIP-Race and Intersectional CLIP retain the initial capabilities shown in CLIP ViT-B/16. We perform a linear probe evaluation to test the feature extraction capabilities of all models, and the results support the retention of feature extraction and classification. Additionally, by investigating the similarity scores of baseline CLIP and the two debiased models, we show that CLIP can still identify image features while reducing similarity to harmful or non-physical input tokens within a SoftMax function.

5.1 Limitations

Although there is a notable amount of literature on bias in artificial intelligence and machine learning models, debiasing protocols are more niche, needing more benchmark metrics and models used in similar experiments. For example, MaxSkew was proposed by Geyik et al. in 2019 and used as a measure by Berg et al. in 2022. NDKL did not effectively show bias due to the many classes it measured compared to its original use for gender. A metric that is more suited to handle many small changes in bias would be better suited for debiasing on these many factors. More published debiasing work would enhance this experiment and provide a better and more well-known way to measure bias in these large systems.

Additionally, more inclusive labeled data would further debias CLIP. The datasets used in this thesis are the most diverse datasets publicly available, yet they could still be more inclusive. A prominent face dataset including a more diverse population on other protected classes, including sexuality, disabled people, and more than just two genders, would be beneficial and a practical next step in training and testing inclusive AI.

More time would also benefit this experiment, as further training could allow the models to train for more epochs and make more adjustments. Better computing power would also speed up training these models and allow for more experimentation.

5.2 Future Work

A couple of next steps could be taken to further the research done in this thesis. This debiasing protocol could be applied to other transformer-based image classification models that contain both image and text encoders. It could also be used on stable diffusion (Rombach, Blattmann, Lorenz, Esser, & Björn, 2022). Additionally, this proposed protocol could be adapted to language models such as GPT4 to debias race, gender, age, and ability on text output.

REFERENCES

Asano, Y. M., Rupprecht, C., & Vedaldi, A. (2020). A Critical Analysis of self-supervision, or What We Can Learn from a Single Image. *International Conference on Learning Representations.*

Baeza-Yates, R. (2018). Bias on the Web. *Communications of the ACM, 61*(6), 54-61.

Berg, H., Hall, S., Bhalgat, Y., Kirk, H., Shtedritski, A., & Bain, M. (2022). A Prompt Array Keeps the Bias Away: Debiasing Vison-Language Models with Adversarial Learning. *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).*

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Conference on Neural Information Processing Systems* . Barcelona, Spain: NIPS.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Machine Learning Research Conference on Fairness, Accountability, and Transparency.* PMLR.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Ziserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. *IEEE International Conference on Automatic Face and Gesture Recognition.* IEEE.

Cho, S., Crenshaw, K. W., & McCall, L. (2013). Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs*, 785-810.

Collins, P. H. (2015). Intersectionality's definitional dilemmas. *Annual Review of Sociology*, 1-20.

Collins, P. H., & Bilge, S. (2020). *Intersectionality.* Wiley.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 1241-1299.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through Awareness. *3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). ACM.

Friedman, B., & Nissenbaum, H. (1996, July). Bias in Computer Systems. *ACM Transactions on Information Systems, 13*(2), 330-347.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. *KDD '19 Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2221-2231). New York, NY, USA: Association for Computing Machinery.

Hardt, M., Price, E., & Srebro, N. (2016, October 11). *Equality of Opportunity in Supervised Learning.* Retrieved from arXiv: arXiv

Islam, R., Keya, K. N., Pan, S., Sarwate, A. D., & Foulds, J. R. (2023). Differential Fairness: An Intersectional Framework for Fair AI. *Entropy*(25).

Kärkkäinen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. *IEEE Winter Conference on Applications of Computer Vision* (pp. 1547-1557). Waikoloa, HI, USA: IEEE.

Kirkpatrick, K. (2017, February). It's not the algorithm, it's the data. *Communications of the ACM, 60*(2), 21-23.

Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. *ACM Conference on Fairness, Accountability, and Transparency.* Seoul, Republic of Korea: ACM.

Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images.*

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021, July 13). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54*(6), 1-35.

Mishkin, P., & Ahmad, L. (2022, April). *DALL-E 2 Preview - Risks and Limitations.* Retrieved from https://github.com/openai/dalle-2-preview/blob/main/system-card.md#probes-and-evaluations

Offert, F., & Phan, T. (2022). *A Sign That Spells: DALL-E 2, Invisual Images and The Racial Politics of Feature Space.*

OpenAI. (2022, July 18). *Reducing Bias and Improving Safety in DALL-E 2*. Retrieved from OpenAI: https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . Sutskever, I. (2021). *Learning Transferable Visual Models from Natural Language Supervision.* San Francisco, CA: OpenAI.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents.* San Francisco, California, USA: OpenAI.

Robinson, P. J., Livitz, G., Henon, Y., Qin, C., Fu, Y., & Timoner, S. (2020). Face Recognition: Too Bias, or Not Too Bias? *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).* Seattle, WA, USA: IEEE.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Björn, O. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10674-10685). New Orleans, LA, USA: IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 211-252.

Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine LEarning. *Equity and Access in Algorithms, Mechanisms, and Optimizations*.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., . . . Li, L.-J. (2016, February). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, pp. 64-73.

Zhao, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 2337-2348.

Appendix A: Results from CLIP-race and Intersectional CLIP debiasing including NDKL

Table 13: Results from the debiasing protocol of race on CLIP. These results used MaxSkew@1000 and NDKL to evaluate with three fairness criteria: demographic parity (DP), equal opportunity (EO), and differential fairness (DF). Results show a decrease in racial bias overall, with the most being MaxSkew@1000 measure using differential fairness and the criteria of demographic parity of the NDKL measure. Gender bias was also measured, and there was a decrease in gender bias across the board. The outlier measure of differential fairness with MaxSkew@1000 of 0.0 could be because either the original measure was already tiny or the balance of the two genders was almost perfect, so it did not get registered by the measurement criteria.

Model	Bias Measure	Measured Bias					
		MaxSkew@1000			NDKL		
		EO	DP	DF	EO	DP	DF
CLIP ViT-B/16 (Radford, et al., 2021)	Race	0.5428	0.4328	0.0728	0.0561	0.0394	0.0394
	Gender	0.1916	0.1916	0.0039	0.0176	0.0176	0.0176
CLIP-Race	**Race**	**0.5721 +5%**	**0.2915 -33%**	**0.0013 -98%**	**0.0667 +19%**	**0.0374 -5%**	**0.0374 -5%**
	gender	0.1416 -26%	0.1416 -26%	0.0	0.0081 -54%	0.0081 -54%	0.0081 -54%

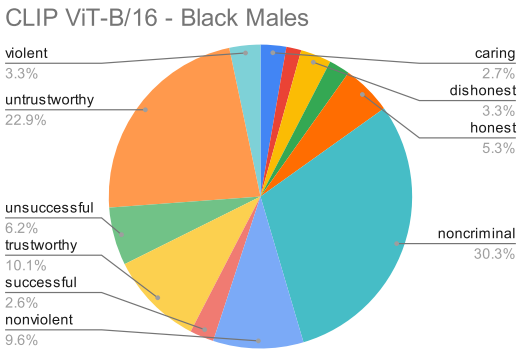
Table 14: Results from debiasing CLIP by training on the Balanced Faces in the Wild dataset. These measures show a decrease in the Ethnicity/Gender classes, which are eight classes encompassing ethnicity and gender. The decrease happens explicitly in the MaxSkew@1000 measure, while the NDKL slightly increases. The increase could be from several reasons, but having one way of encompassing the complex concept of intersectionality is almost impossible.

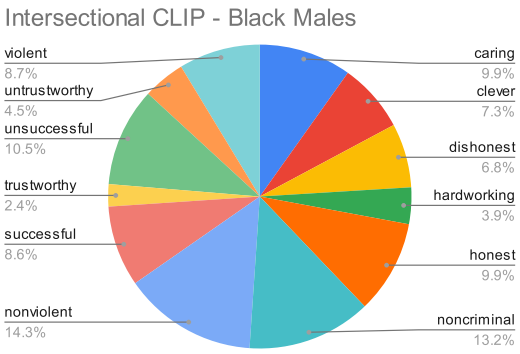
Model	Bias Measure	Measured Bias					
		MaxSkew@1000			NDKL		
		EO	DP	DF	EO	DP	DF
CLIP ViT-B/16 (Radford, et al., 2021)	Ethnicity	0.4893	0.4893	0.1204	0.0652	0.0652	0.0652
	Gender	0.2334	0.2334	0.0077	0.0290	0.0290	0.0290
	Ethnicity /Gender	0.7195	0.7195	0.3214	0.0989	0.0989	0.0989
Intersectional -CLIP	Ethnicity	0.4939 +1%	0.4939 +1%	0.1033 -14%	0.0939 +44%	0.0939 +44%	0.0939 +44%
	Gender	0.0915 -61%	0.0915 -61%	0.0	0.0055 -81%	0.0055 -81%	0.0055 -81%
	Ethnicity /Gender	0.5989 -17%	0.5989 -17%	0.1993 +33%	0.1043 +5%	0.1043 +5%	0.1043 +5%

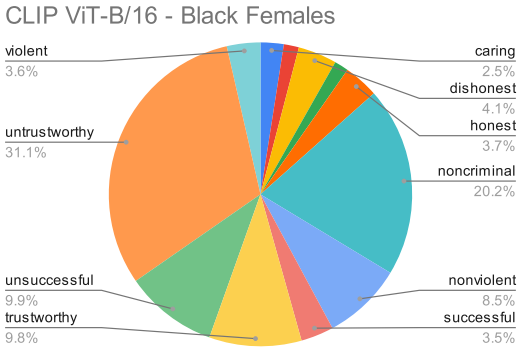
Table 15: Results of testing Intersectional-CLIP on an unbalanced test set. With these results, the symmetric measurements were replaced by better, if not more consistent, results, removing the strong influence the validation set has on the presented results.

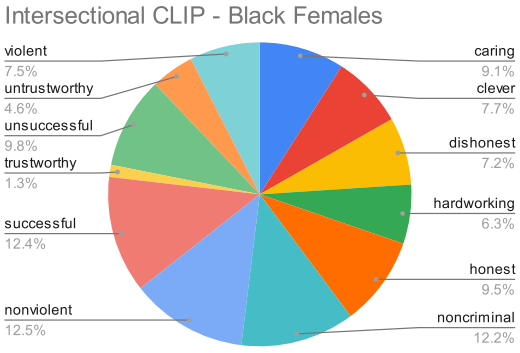
Model	Bias Measure	Measured Bias					
		MaxSkew@1000			NDKL		
		EO	DP	DF	EO	DP	DF
CLIP ViT-B/16 (Radford, et al., 2021)	Race	0.5286	0.5287	0.1493	0.0631	0.0617	0.0617
	Gender	0.2082	0.2134	0.0062	0.0224	0.0232	0.0232
	Ethn. Gender	0.7322	0.7296	0.3317	0.0905	0.0896	0.0896
Intersectional -CLIP	Race	0.5099 -4%	0.4854 -8%	0.0940 -37%	0.0930 +47%	0.0887 +43%	0.0887 +43%
	Gender	0.0599 -71%	0.0608 -72%	0.0	0.0039 -83%	0.0041 -82%	0.0041 -82%
	Ethn. Gender	0.5920 -19%	0.5626 -23%	0.1633 -51%	0.1017 +12%	0.0972 +8%	0.0972 +8%

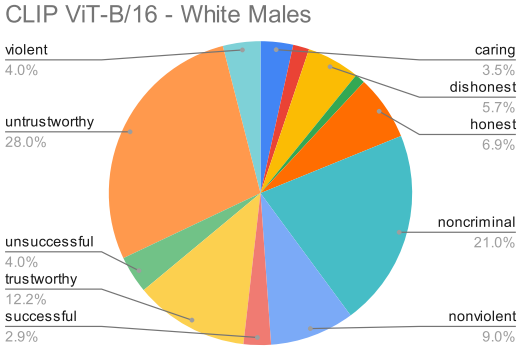
Appendix B: Similarity Scores for All Ethnicity/Gender Pairs

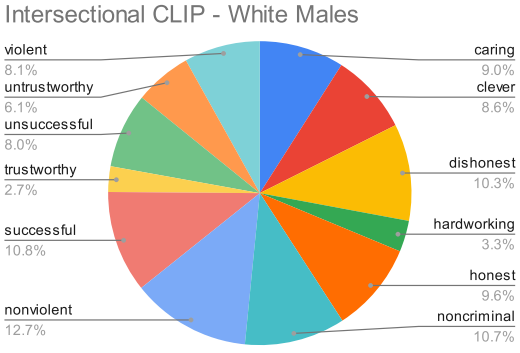




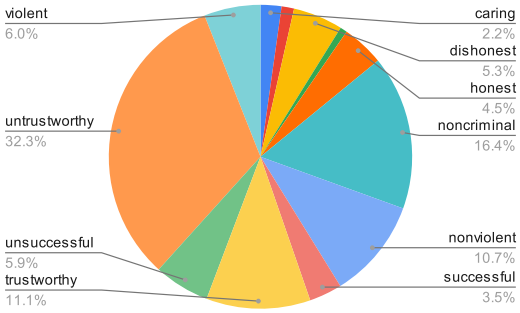




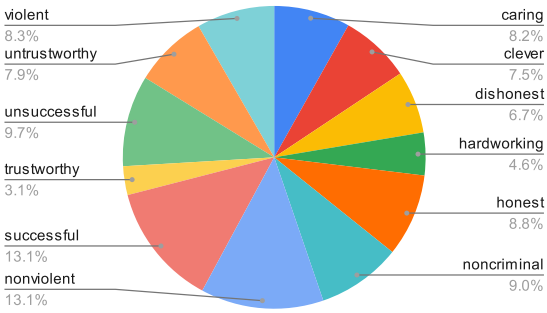




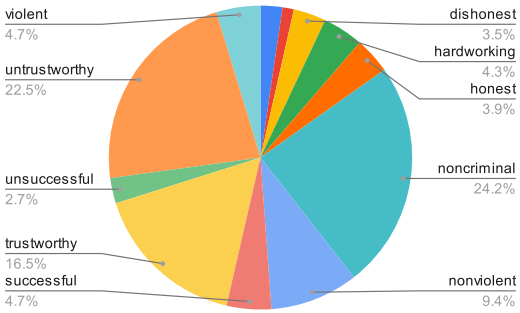
CLIP ViT-B/16 - White Females



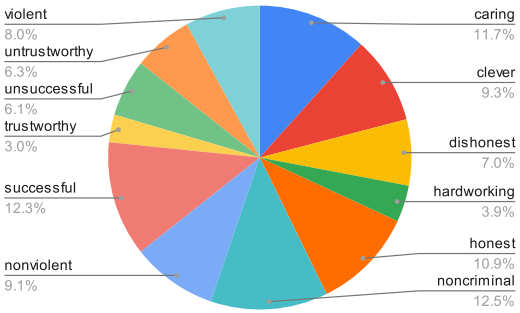
Intersectional CLIP - White Females



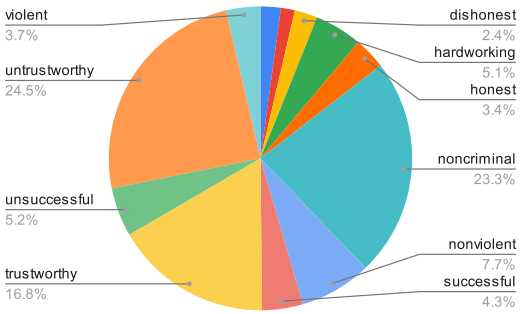
CLIP ViT-B/16 - Asian Males



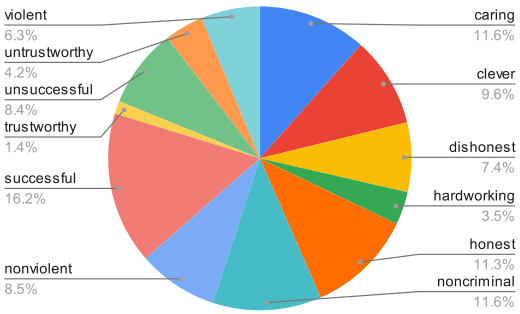
Intersectional CLIP - Asian Males



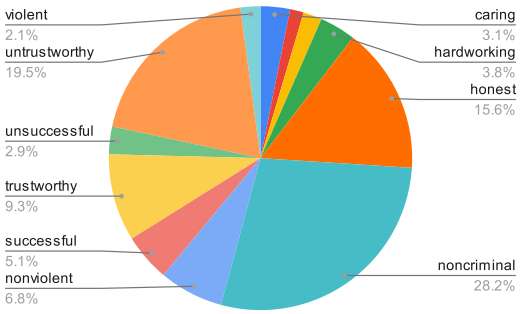
CLIP ViT-B/16 - Asian Females



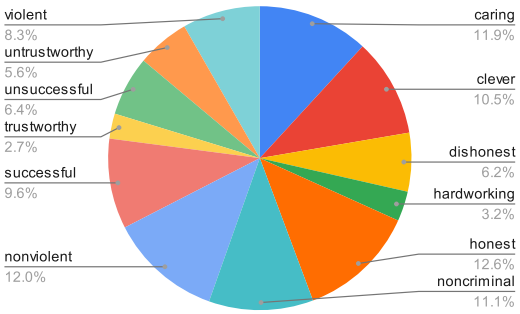
Intersectional CLIP - Asian Females



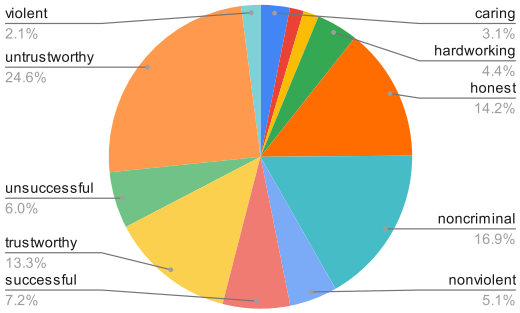
CLIP ViT-B/16 - Indian Males



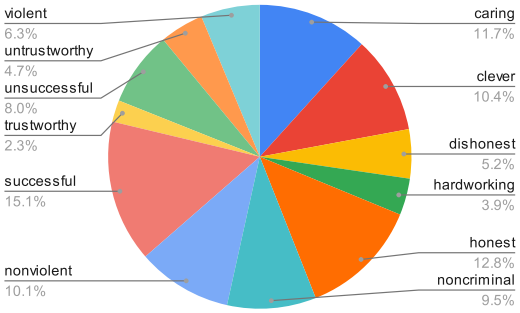
Intersectional CLIP - Indian Males



CLIP ViT-B/16 - Indian Females



Intersectional CLIP - Indian Females



Appendix C: Code

The code for this thesis can be found in GitHub, here:

https://github.com/ehoepfin/CLIP-Debiasing/tree/main.