The Impact of Biases in Facial Recognition Artificial Neural Networks

Ezra M.Wingard
Candidate for B.A. Degrees in Cognitive Science and Psychology

State University of New York, College at Oswego
College Honors Program

May, 2023

Abstract

This study probes how biases are formed, and then mitigated within artificial neural networks for facial recognition. In current research on facial recognition neural networks, it has been shown that there are many ways that biases/prejudices can negatively affect the accuracy of the network on characteristics such as gender status and identity. In order to test this, two pre-trained neural networks were fed novel datasets - one on cisgender faces and one on transgender faces. The two pre-trained models were then analyzed with regards to gender identity and status variables on accuracy rates calculated from the direct prediction outputs provided by the neural networks. Notable biases were found within both datasets and models on gender characteristics.

Table Of Contents

Introduction

Neural Networks

Artificial neural networks (ANNs) are a subset of artificial intelligence and machine learning that can provide us with predictions of information we care about. They are based on human neural networks located within our brain, that consists of neurons, synapses, and electrical impulses (Gupta, 2013). To put it extremely simply, they consist of inputs and outputs. The inputs, which can be any type of information, are fed into the Neural Network (NN) in order to make it "learn" and can be used to "test" how well the algorithm performs after it has 'learned' sufficient information. This "learning" is achieved through the shifting of weights in the artificial neurons within the neural network which simulates what humans think of as learning and facilitates the "conversations"[1] that happens between neurons. These conversations form connections between neurons, sometimes in more than one layer (called a hidden layer[2]), that will then converse with each other to form predictions and share outputs which serve as the output. The outputs can be thought of as transformations of your inputs (DeepAI, 2019). For example, for the scope of this paper, the inputs will be images of cisgender and transgender people of different races and ages, and the output is a prediction of their gender. The neural network used in this study is a classification algorithm that takes an input and classifies the facial images into a select few demographics that are of interest to be measured (which are shown through the outputs).

---

[1] Conversations in the context of neurons within ANNs means that one neuron or node within the algorithm will pass along its information to similar nodes to form connections.

[2] A hidden layer in NN terminology denotes that there is one or more layers in between the input and output layers that can create more connections and provide more detailed and potentially accurate outputs.

Before you can create said outputs, you must train your neural network. It must learn all the information about what you want to make predictions about, so that you can make sure your predictions will be accurate. The training process can be relatively short to a very lengthy amount of time, depending on the volume of the inputs that you want to train it on. Generally speaking, a greater number of inputs for training leads to better success with accurate outputs. There is also such a thing as too many inputs during the training phase which can lead to your neural network being unbalanced and not forming good outputs, such as within Deep Learning NN architectures. This is known as ‘overfitting’ and denotes that your algorithm has gotten used to the items in your training dataset and will not accurately generalize to other novel datasets, such as those used to test your NN (Guo et al. 2015).

After you feed your neural network the inputs, and it has sufficiently taken them into account and processed them, you should be able to see that your network has new weights. With each new input, the neural network will change its neurons’ weights, just like what is hypothesized happens within the human brain. These weights will be used to later create those predictions that we need to make with the network. The weights are extremely important to how a network can function – both in humans and in machines. In humans and machines, the weights determine if a singular neuron activates or not when thought of in terms of a threshold. The higher the number, the stronger the weight is and the more potential to fire, and vice versa for the lower the number. The actual number representing the weight is mostly arbitrary, but the magnitude of the number is the

thing that matters. The neuron (whether artificial or biological) will fire if and only if the total number of the weights combined is greater than the set threshold. The threshold is the thing that makes sure that the neuron will not fire when it's not supposed to– it should receive only enough excitatory signals that it exceeds the threshold, and if it doesn't have enough power to exceed the threshold, it doesn't activate (DeepAI, 2019).

One example of the most basic neural networks that can help to cement these ideas is a single-layer perceptron. Although Perceptrons were originally introduced by Rosenblatt (1958), they were popularized by two infamous computer scientists, Minsky and Pappert, within the late 1960s through their uniquely named book *Perceptrons*. Perceptrons, at the time of publication in 1969, can be thought of as the ancestors to modern Artificial Intelligence (AI) algorithms. Within the example of a single-layer perceptron, the *AND* perceptron can explain the most basic ideas that happen within machine learning algorithms. Within this perceptron, it takes 2 inputs, and has 1 output node. For the *AND* perceptron, both of the inputs must be more than 0 to fire. Let's say for example that the weights for the inputs are both 1, and the threshold is a little bit less than 2. This means that the neuron is likely to fire, because the weights combine to 2, and 2 is more than the threshold. If it did not fire, and it looks everything was done correctly with the weights and threshold being lower than the combined weight total, then lower the threshold and try again. If it still doesn't fire, keep lowering the threshold and try again each time to make sure that the threshold is at the perfect amount to just make it fire. However, if one input weight was 0, and

the other input was 1, with a threshold still at a little below 2, it would not fire, because the sum must exceed 2. The same is true for if the inputs were both 0, or even if they were -1. The signal that the inputs are trying to put into the neuron are inhibitory, and will not allow the neuron to fire and pass along its message. Although this perceptron may seem basic, it can be used to show the inner workings of a neural network, and perceptrons in general are used as the building blocks of more complex NN algorithms used today (Bhardwaj, 2020). We can further abstract away from the *AND* perceptron by adding more layers, inputs, and other infrastructure to create what is used more frequently in modern Machine Learning (ML) algorithms, such as the neural network created for this thesis.

The specific type of neural network that was used for this project is a Convolutional Neural Network (CNN). This network is different from a traditional Artificial Neural Network (ANN), because it relies on layers of convolution to compute the output predictions rather than weights and nodes. These differences can be summed up to the CNN being a two-part NN: it does feature extraction and feature mapping, while ANNs typically only extract features. In CNNs, the convolutional layer uses the mathematical process of convolution, the process of taking two functions and creating a third to see the difference between them. To put it more simply, the convolutional layers extract important information from the inputs and then transform those inputs into a map onto the output image. The output from one convolutional layer will be provided to the next convolutional layer on and on until the end of the hidden layer. Additionally, pooling layers typically follow the convolutional layers within the

CNN architecture, allowing for the convolution done within the previous layer to be fitted into the next convolutional layer - it does this by reducing the dimensions of the convolutional layer while preserving important information and details to pass on to the following layers. This process repeats going from a convolutional layer to a pooling layer until the end of the "hidden layer" within the CNN architecture. At the end of the hidden layer and after the last pooling layer, a fully-connected layer is typically used in order to convert the two dimensional features that have been fed forward into a one dimensional vector that will supply the output that is readable to humans (Guo et al., 2015). One example of a previously commonly used Convolutional Neural Network is AlexNet (Krizhevsky, 2014). Although there are other neural networks with greater accuracy findings (depending on the training dataset used), this neural network is one "household" name of image classification neural network algorithms. It has been used for computer vision[3] and as with any artificial neural network, when its function is put simply, it takes an input and transforms the data into an output. An example of an output from a computer vision NN is detecting a face within an image, such as what was done within the algorithm used for this study. Additional things that can be accomplished with computer vision include, but are not limited to: image classification[4], and facial recognition, detection, and classification[5].

---

[3] Computer Vision denotes the type of inputs and outputs that a NN can predict. There is a subset of NN algorithms which are trained on 'Vision' tasks, such as by taking image, video, or live web camera feedback. A Neural Network trained within the Computer Vision task excels and is specifically trained to provide feedback based on these 'visual' inputs.

[4] An example of image classification is AlexNet. It's able to be fed an image of a bird and classify it as 'bird'

[5] Detecting a face within an image, recognizing a specific individual's face, and classifying certain characteristics of the face like gender, race, age, etc. are all things that Computer Vision algorithms can do with respect to facial analysis.

The actual process of using convolution within a neural network contains a lot of complex details and jargon such as kernels that will not be discussed heavily within this thesis. As such, convolution will not be delved into deeply within this paper, because the neural network model's outputs and what we can learn from said outputs are of more importance. Because of the multiple different types of layers that CNNs employ, this allows such models to excel at computer vision tasks in comparison to regular ANNs. The fully-connected layer comes after the last pooling layer in the network, and contains a lot of the parameters from the whole CNN model's architecture – about 90% of them (Guo et al., 2015). This layer is extremely important as it allows the network to compile the information processed within previous layers, and then form the final output that we will get. For this paper's purposes, that would be output predictions of the gender of the individuals in our input images. With the architecture of CNNs in mind (specifically the inclusion of pooling and fully-connected layers), we can easily understand why this model of neural network excels at processing images over other models – computer vision and image processing requires a lot of attention to detail, especially in cases which include classification. When taking into account the tasks that one wants to perform with ML algorithms, these things must be considered to get the best results. Thus, since this thesis is on facial recognition software (a subset of computer vision), it was imperative to use CNN models.

Why Is This Important?

Historically, transgender people and non-white individuals are left out of important discussions, even when it predominantly affects them. ML / AI are fields that have historically neglected to have such conversations. Throughout the years, research on artificial neural networks has been progressing rapidly and providing a lot of development with such technology. With that being said, however, the ethics and ramifications of using such technology has lagged far behind the pace of research and development.

There have been conversations involving research and development on AI for a long time, however according to the The AI Index 2022 Annual Report (Zhang et al., 2022), the topic of ethics within machine learning started to be discussed more prevalently circa 2014. Additionally, Zhang et al. reported that the number of publications on the subject of AI ethics have increased by a factor of 5 since 2014, with 71% of the publications pertaining to organizations and corporations in the sector of Industry. Lastly, the AI Index Annual Report noted that the number of papers is expected to continually grow more and more with each coming year (2022). There were also several notable projects that launched in 2017 regarding ML ethics include Canada's Pan-Canadian Artificial Intelligence Strategy and Japan's AI Technology Strategy (EPRS, 2020), MIT's Ethics and Governance of Artificial Intelligence as part of MIT's Media Lab (MIT Media Lab, 2017), and many more. This spike in discussions about how NNs can be used for harm has led to notable publications on the subject of facial recognition, such as Dr. Buolamwini's dissertation that came from the MIT Media Lab - one of the

projects that was formed during the first ‘boom’ of research on these ethics (Buolamwini & Gebru, 2018). Such notable literature that are discussed below show how pertinent the issue of ethics is whenever new technology such as NNs are being formed.

Many researchers have found that transgender individuals and Black people - specifically Black women - are most affected by this technology in real-world contexts, which is why the issue of ethics is so important. Buolamwini and Gebru’s *Gender Shades* (2018) findings documented that darker skinned Black women are more likely to be misgendered by computer vision software. Scheuerman et al.’s work on *How Computers See Gender* (2019) has found that transgender people are also often misgendered by computer vision algorithms. These two studies are not the only research that have been done on the subject, however it’s plain to see that there are potential harms of these technologies and an intense need to help rectify this.

Despite the documented dangers and problems associated with facial recognition software, machine learning and neural networks are rapidly gaining popularity– not only within personal usage but also within the commercial domain. Specifically, facial recognition software has been the subject of many papers that discuss ethical dilemmas when it comes to certain situations that neural networks may be used. Scholars within the field have raised concerns regarding use of Automatic Gender Recognition (AGR) NNs that may adversely affect transgender-specific populations and people who have not consented to having their data run through an algorithm. Additional ethical dilemmas

associated with AGR include algorithms that could misgender transgender people and does not allow for their algorithmically-assigned gender to be changed through user input or other means (Keyes, 2018; Scheuerman et al., 2019, 2020).

Others have noted that there are some governmental programs that have been found to use AI in ethically questionable ways, such as the police in Detroit, Michigan (Johnson, 2022) as well as England and Wales (Radiya-Dixit, 2022). Additionally, concerns have been raised by AI scholars on the ramifications that NNs can have on transgender people who are misgendered. According to several authors, there have been ethical ramifications noted with AI that could misgender transgender people and does not allow for their algorithmically-assigned gender to be changed through user input or other means (Keyes, 2018; Scheuerman et al., 2019, 2020). These staggering anecdotes on potential harms that may come out of unethical uses of such technology is part of the reason that similar projects are needed within these rapidly developing areas of technology.

In addition to the research findings mentioned above, it has been documented that these same groups of people (transgender and non-white people) are typically underrepresented in the datasets used to train and test the algorithms (Wu et al., 2020 ; Karkkainen & Joo, 2021). This is important because during the training phase, if a neural network is not exposed to diverse demographics, it will not perform well on even basic recognition tasks of such individuals. The training phase, as was mentioned in the previous section, is critical for the neural network's outputs to be accurate and representative of the population it's being tested on. Although the importance of the training phase in ML development is

fairly well-known, many individuals within the field have opted to train their facial recognition algorithms on datasets that contain mostly older cisgender white men. This may be because they have disregarded the need for more inclusion of transgender people and racially diverse individuals, opted to not focus on adding diversity into their datasets because of less readily available large-scale training sets, or some other rationale. Some datasets which are very unbalanced on race/gender characteristics that are cited widely within literature include the LFW (Labeled Faces in the Wild), CASIA-WebFace, MS-Celeb-1M datasets, or the Adience and VGG Face datasets (Papers with Code, 2023). All of the aforementioned datasets seem to lack the thing that FairFace prides itself on - having a larger number of faces with backgrounds in diverse racial identities, genders, and ages than the typically high numbers of older cisgender white men's faces. The advantages to such large-scale datasets, however, is that they can be split between training, validation, and testing datasets which can allow for ease of preparing the NN for projects. One may not have to go searching for, or create, their own datasets when there are larger-scale options available. Even though there is an immense need for such diversity, it seems that a lot of the larger ML projects that use facial recognition software opt to use the larger, unbalanced datasets that have been cited more often in literature within the field of AI. Although, as was mentioned, these pre-created datasets may be convenient, they may be detrimental to the equity and accuracy of the most impacted groups that this software will be used on.

There have been other attempts in the past to specifically create datasets to help NNs train on transgender faces, such as the HRT Trans Database (AIAAIC, 2023). The dataset in question was bashed by several authors such as Keyes and Scheuerman for questionable ethics for multiple reasons, including but not limited to: lack of information on consent from people in the photos, lack of publication of the dataset despite being publicly funded, and feuds over the dataset being created for "national security purposes" (AIAAIC, 2023). Other papers that have included transgender people in their testing and training datasets have opted to not publish their datasets online because of potential ethical concerns caused by doing so, such as the dataset mentioned above. Because of the consistent targeting of marginalized communities such as transgender people, there may be concerns about datasets that specifically include these populations. Thus, when considering how to train a neural network to debias on such populations, we must find other options than readily available datasets like those commonly used within larger-scale models.

It is important to note that not all of the biases that come from ML algorithms are formed during the training process (especially when unbalanced datasets are used). There are mounds of research that have been and are being conducted on the very topic of how these biases form, as well as ways in which we can mitigate them before there is detrimental harm to humans (Wu et al., 2020; Scheuerman et al., 2020; Buolamwini & Gebru, 2018; Google, 2022). Balancing datasets is one of the many ways that have been proposed in order to lessen the biases that can be formed within AI applications. Although that is

important to note and consider with regards to such technology, many researchers have noted specific ways that we can reduce biases within these algorithms as a way of harm reduction through dataset balancing (Joo & Kärkkäinen, 2021; Wu et al., 2020). Others have argued that balancing datasets will not be enough for bias mitigation within NN models (Wang et al., 2019A ; Wang et al., 2019B; Alberio et al., 2020; Gong et al., 2020; Zhang et al., 2018).

Thus, with regards to this honors thesis, it was important to try to see what methods can potentially reduce biases/prejudice against transgender people. As there are many potential real-world implications that this technology can have, it is important to consider ways that harms can be reduced. The thought was that by creating novel datasets based on scraped images of transgender people and cisgender people on two identity categories, there could be testing for (and possibly mitigation of) potential biases that may arise, even from a model that prides itself on being balanced on race, gender, and age. It was questioned if FairFace's dataset balancing on cisgender people would show different accuracy rates within non-cisgender populations. Additionally, it was questioned whether or not using a different model that was trained on a non-balanced dataset would make a difference in regards to gender classification outputs.

Methods

Models

In order to find and mitigate biases, I used a pre-trained neural network model from FairFace (Karkkainen & Joo, 2021) that was said to be trained on a balanced [6]dataset. This is in direct contrast to more well-known and used datasets (which do not have the same claims of being 'balanced') - these 'unbalanced' datasets typically contain high amounts of cisgender white men, and do not seem to account for diversity of race, age, and gender during the training phase. As discussed above, it was hypothesized that because of the FairFace model being trained on a more diverse dataset, that it may be less prejudiced against transgender people, especially those of different age and race groups.

The specifications for race, gender, and age outputs used were those that were already specified with the pre-trained model - 4 races (White, Black, Asian, Indian), 7 races (White, Black, Latino_Hispanic, East Asian, Southeast Asian, Indian, Middle Eastern), gender, and age (0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+). Although the code which contained the pre-trained FairFace model used language such as "Male" and "Female", this was changed to "Man" and "Woman" for purposes of focusing on gender-based language. The switch to gender-based rather than biologically sex-based language to include transgender people within this study, as transgender people may not identify with their assigned sex at birth.

---

[6] "Balanced" within NN datasets means that the number of images within the dataset that were used to train the neural network were sorted in a way to ensure that the amount of photos pertaining to race, age, and gender were equal. This 'balancing' is thought to potentially decrease the negative effects of a NN model experiencing overfitting on one specific population, such as older cisgender white men.

To test the biases of the pre-trained, balanced FairFace neural network model, code was taken from their GitHub repository (Joo & Kärkkäinen, 2021). Some of the code that was taken directly from the repository needed to be changed so that it could be used to generate the necessary outputs, such as file names. Once the code was sufficiently modified, testing was able to be completed and outputs were able to be received.

In order to empirically test on potential differences from the model based on fairness of gender categories, a second pre-trained model was used with the same novel training datasets. The model in question is InceptionResNet v1, which has made no claims to debias or balance the datasets that the neural network was trained on. Similar parameters were from the FairFace model for outputs (7-race, man and woman) for consistency purposes. However, certain characteristics were not able to be parsed into outputs within this model, such as a 4-race contrast, and age outputs. The only outputs that were measured within data analysis, however, were gender groups across both models.

The required code for implementing the InceptionResNet v1 model was taken from the author's GitHub repository (Sandberg, 2023). Unlike the FairFace model, the IRNv1 model was provided with no additional runnable code (even with modifications) for the purposes of this study. Thus, original code was written for this model’s classification outputs and predictions. The IRNv1 model was pre-trained on the VGGFace2 dataset, which consists of over 3 million images (Cao et al., 2017). The number of men reported within the dataset is

approximately 59.7% , with no reports on racial demographics within the original paper (Cao et al., 2017).

Datasets

Because there are no known (ethical) datasets that are available on the internet consisting of transgender people for facial recognition purposes, I used the Instaloader Python API (Instaloader, 2023) to scrape images of transgender and cisgender people from Instagram. An original Python script was created and used the API in order to download necessary images from target demographics. Through this script, a hashtag was inputted as a string for the scraper, so that all images within the inputted hashtag would be downloaded. Examples of hashtags used for transgender populations include #GirlsLikeUs, #FTMtransgender, etc. The scraped images were then cleaned accordingly (images consisting of subject matter other than binary transgender individuals faces were omitted), and then the testing datasets were thus created. The original images scraped using Instaloader were separated into different folders based on demographics. The images from the datasets were found and used based off of public information available on Instagram, in accordance with Instagram and Instaloaders privacy regulations and guidelines.

The images were then renamed post-sorting, without the usage of any personal identifiers - example of a transgender man's image name was "TM1_1", indicating that it is the first Trans Man (TM) in the set, and the first image of said man. Each person within the datasets were given a different identifier to track the number of images from each person - TM1 and TM2 would signify two different

individuals who are both transgender men[7]. If there were more than one image of the same person (TM1), then it would correspond to a difference in number after the underscore ("TM1_2"). This was used in order to identify misgendering in the between groups and within groups outputs for analysis purposes.

The individuals within the cisgender dataset were gathered based on self-identification of race through hashtags, and all identified as cisgender at the time of the image collection. Hashtags used to find cisgender populations on Instagram using the scraper script include #BlackBoyJoy, #Latinoman, etc. Categorization of images into folders were done by racial self-identification (7 race categories) rather than gender (as done with *transgender* men/women) to denote that they were in the cisgender dataset. At the time when the images were scraped for the datasets, all individuals self-identified as cisgender within this dataset.

In order to run the neural network models, each dataset had to have a comma separated value (csv) file created with its image paths. These csv files were used to provide the 'inputs' to the models to be processed. In order to create the accurate files to feed to the neural networks, a novel python script was created to read through the image dataset folders and automatically create a csv file for the corresponding dataset. There was one csv file for each dataset, which consisted of all of the image paths to be fed into the neural network. The transgender and cisgender csv files were separated in order to maintain congruence with inputs and outputs for posterity sake.

[7] Each different demographic was given a different shorthand signifier. Transgender women would be named as TW#_# in the same naming convention. Similarly with the cisgender dataset, Latino men would be LM#_#, and Black women would be BM#_#.

While running the pre-trained models, images from the constructed datasets were cleaned and cropped around faces found within the files and automatically sorted into a folder of detected faces. This process happened automatically, without supervision or action required on the part of the user. After the files were saved to a folder of detected faces, predictions took place and were saved to an output csv file. Because the neural network models automatically created the output csv file from the detected faces folder, there were two folders created to differentiate between the two datasets (cisgender vs transgender). Thus, when predictions took place, outputs for the entirety of one dataset were placed in a csv file. These files, which served as containers to access the NN models outputs, were thus used to compute accuracy and other analyses to determine if biases were present in the models used.

Within the two specific datasets scraped for this study, there were varying Ns between demographics. Table 1 shown below shows the exact number (N) of images within each dataset created. The exact number of images that were used within each dataset is reported, as well as the rounded percentage that each gender identity (men and women) account for within the whole dataset.

Table 1

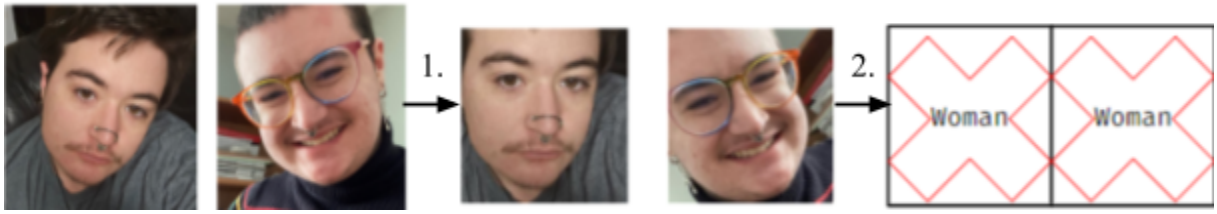
Total Number Of Images Per Dataset

Datasets	N of Women	N of Men
Transgender (n = 414)	222 (53.62%)	192 (46.38%)
Cisgender (n = 550)	280 (50.91%)	270 (49.09%)

Note. Although the Ns per dataset differ, within the analyses conducted such discrepancies are accounted for.

Figure 1

Examples Of Dataset Inputs To The Fairface And Irnv1 Models



Note. The two sets of images in the figure were taken of the author as examples of dataset inputs to the NN Models. The leftmost set of images represents the initial inputs to the NN models. Following 1., in the middle, the initial inputs were cropped around the detected faces in each photo. Once all detected faces were found, following 2., the NN models provided outputs based on the detected faces and provided gender predictions. The author was misgendered within both images across the two NN models.

Analyses

Preliminary analyses of the gender accuracy rates were conducted in Google Sheets[8] through a formula traditionally used in accuracy calculations for machine learning (Google Developers, 2023):

$$\frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

Within the context of this study's neural network model classifications, the output was binary and thus only required the calculation of a True Positive and False Positive ratings to calculate the accuracy rates. For individuals whose images were included in the test set of this study, their self-identified gender identity was

[8] Calculations that were conducted in Google Sheets were also confirmed using Confusion Matrix calculations in RStudio

used to mark correct/incorrectness. This means that a transgender woman who had been classified as a woman would be correct and noted as a True Positive, and vice versa for transgender men).

Because of the nature of True and False Negatives, such variables were not compatible with the binary classification ratings. The neural network either misclassified them or classified them correctly based on their self-given identities and thus were omitted from the accuracy calculation. Thus, the formula that was adapted for the use of this study is as follows:

$$\frac{True\ Positives + False\ Positives}{N}$$

In the above adapted formula, the True Positives represent the correctly classified individuals, the False Positives represent the misclassified individuals, and N represents the total number of images within each dataset.

Further analysis on the test datasets (both the transgender and cisgender test sets) on both models were conducted using R Studio and the R programming language. A logistic regression was performed to determine analysis of each predictor variable on the outcome variable. Predictor variables within the analyses were model (FairFace vs InceptionResNetv1 or IRNv1), gender identity (man vs woman), and gender status (transgender vs cisgender self-identification). The sole outcome variable measured is the accuracy rates calculated.

Results

Both Models

With all results mentioned below, there are varying statistical tests completed for each model. It is important to note that unless otherwise specified, gender classification rates only include gender identity. Outside of specific instances, analyses to be described below take into account general gender identity without respect to gender status. Table 2 shown below takes into account all accuracy rates per demographic (gender identity and gender status) within both models.

Table 2

Accuracy Rates per Demographic

	Datasets		Demographics	
	Gender Status	Total Accuracy	Accuracy Men	Accuracy Women
IRNv1	Cis	68.7%	69.5%	67.5%
	Trans	60.6%	47.4%	72.2%
FF	Cis	95.3%	93.5%	97.5%
	Trans	66.7%	53.6%	77.9%

Note. FF and IRNv1 on the vertical axis of the table represent the models (FairFace and InceptionResNetv1) used within the study. The highest accuracy percentage is for cisgender women within the FairFace model (93.5%), and the lowest accuracy is for transgender men within the IRNv1 model (47.4%).

A logistic regression was calculated to determine the odds ratios and effects between both models with respect to the gender identity and gender status on accuracy rates. There was a significant main effect of the model employed,

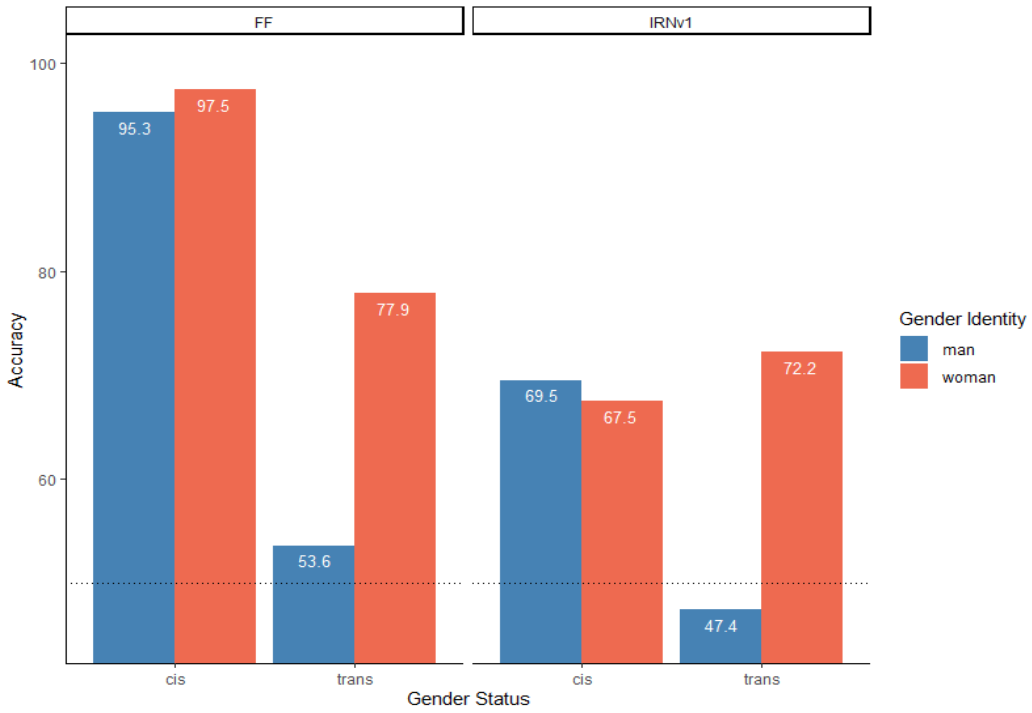
gender status, and gender identity. In regards to the models, it was found that FairFace was 6.27 times more likely to be accurate on general gender classifications than the IRNv1 model, $p < .001$, 95% CI [-.30, -.18]. Based on the gender status, the odds of a cisgender person being gendered correctly was 12.34 times more likely than a transgender person, $p < .001$, 95% CI [-.47, -.33]. Lastly, with gender identity, it was found that the odds of a woman being gendered correctly was 2.73 times more likely than a man, $p < .001$, 95% CI [-.02, .11].

In the same logistic regression, two out of four interactions were found to be significant. The two significant interactions were between model and gender identity, as well as model and gender status. FairFace was found to have a higher accuracy on men, with the odds of a man being gendered correctly as 2.99 times higher than within the IRNv1 model, $p = .05$, CI [-.15, .03]. Additionally, FairFace was more likely to gender a cisgender person correctly around 4.89 times more than the IRNv1 model, $p < .001$, CI [.07, .28]. The interactions between gender status and gender identity (1.12, $p = .82$, 95% CI [.10, .30]), as well as model, gender identity, and gender status were not found to be significant (2.83, $p = .1$, 95% CI [-.08, .21]).

Within a logistic regression the Odds Ratios provided can be used as the effect size. We can see that the largest effect on the accuracy outputs that was measured was from gender status (an Odds Ratio of 12.34 for the gender status alone on accuracy rates). The smallest effect size noted was from a statistically insignificant interaction between gender status and gender identity (an Odds Ratio of 1.12 for the interaction between gender status and gender identity).

Figure 2

Examples Of Dataset Inputs To The Fairface And Irnv1 Models



Note. The bar graph is separated between Model - FairFace (FF) and InceptionResNetv1 (IRNv1), Gender Status - cisgender (cis), transgender (trans) testing dataset, and Gender Identity - man (blue / left bars), woman (pink / right bars). The dotted line denotes the chance-level accuracy for classification (50%).

FairFace Model

Preliminary analyses were conducted on accuracy rates for the FairFace model. The binary transgender testing dataset showed an initial accuracy rate of 66.7% overall on binary transgender status. For trans women specifically, the accuracy rate was 77.93%, in contrast to trans men accuracy rates of 53.65%. This is in stark contrast to the binary cisgender testing dataset, where classification rates for cisgender women had a 97.5% accuracy rate, with cisgender men lagging

behind with an accuracy rate of 93.5%. The total accuracy rate when tested on cisgender status was 95.3%.

A logistic regression was conducted to determine the FairFace model's correct rate of gender classification on the identity of woman given the correct rate of gender classification on the identity of man. It was found that, when holding gender status constant, the odds of a woman being gendered correctly within the FairFace model was 2.73 times higher than a man being gendered correctly, $p = .05$, 95% CI [0.18, 19.95]. With respect to gender status, while holding gender identity constant, the odds of a cisgender person being gendered correctly was found to be 1.08 times higher than transgender people within this model, $p < .001$, 95% CI [-3.06, -2.01]. The interaction between gender status and gender identity within the FairFace model was not found to be significant (1.12), $p = .82$, 95% CI [-.92, 1.04].

InceptionResNetv1 Model

Preliminary calculations were done on the IRNv1 Model's accuracy rates. The overall accuracy rates on the binary transgender testing dataset was 60.7%. For trans women specifically, the accuracy rates within the IRNv1 model predictions was 72.2%, while on trans men it was 47.4%[9] The overall accuracy rates on the cisgender test dataset was 68.6%. Cisgender men had an accuracy of 69.5% while cisgender women had an accuracy of 67.5%.

An additional logistic regression was conducted to determine the IRNv1 model's correct rate of gender classification on the identity of woman given the

[9] This accuracy rate of 47.7% was the lowest of ALL average accuracy rates across demographics measured in both models. For contrast, the highest average accuracy rate for a demographic was 98.2% accuracy on cisgender women within the FairFace model.

correct rate of gender classification on the identity of man. There was a significant main effect of gender status and interaction between gender status and gender identity. When holding gender identity constant, the odds of a cisgender person being gendered correctly within the IRNv1 model was 1.47 times higher than a trans person, $p < .001$, 95% CI [-1.32, -.58]. Additionally, it was found that the odds of a transgender woman being gendered accurately within this model was 3.22 times higher than a cisgender man, $p < .001$, 95% CI [.64, 1.71]. However, the effect of gender identity within the IRNv1 model on accuracy rates was not found to be significant (2.47), $p = 0.56$, 95% CI[-.45, .24]. With respect to the interaction between gender identity and gender status, it was found that a transgender gender status created an effect in gender identity that was not present outside of the interaction, because the main effect of gender identity was found to be statistically insignificant.

Discussion

NN Model Outcomes

The results of this study provide important insights into the accuracy of gender identity and status classification between two different Deep Learning models, FairFace and IRNv1. The two models within this study were chosen because of the datasets used within pre-training. The creators of the FairFace model have made claims that their dataset used for training was balanced on race, age, and gender; IRNv1 has made no such claims[10].

Overall, both models were typically better at gendering women (regardless of gender status), [11] however with regard to gender status alone, cisgender people were more likely to be gendered correctly than transgender people. The study found that FairFace was significantly more accurate in general gender classification (both gender status and gender identity) than the InceptionResNetv1 model.

Further analysis revealed significant interactions between all predictor variables - model used, gender identity, and gender status. FairFace was more accurate in gendering men and cisgender individuals than InceptionResNetv1. The study also found that the FairFace model was more accurate in correctly gendering women compared to men. The InceptionResNetv1 model, on the other hand, had no significant difference in accuracy between gender identities. This

[10] The creators of the VGGFace2 dataset, which IRNv1 was pre-trained on, claimed that their dataset was more balanced on gender (men and women) than previous datasets with comparable size.

[11] There is one exception to this generalization across models. Within the cisgender dataset on the IRNv1 model, cisgender men were 2%

suggests that the FairFace model may be biased towards gendering women correctly.

The findings of this study have important implications for the development and use of gender classification algorithms. It is crucial to consider the potential biases that may exist in these models and to ensure that they are trained on diverse and representative datasets. Additionally, the study highlights the importance of considering the datasets used to pre-train NN models. Although FairFace was technically better on all gender classification tasks than the InceptionResNetv1 model, there were still discrepancies pertaining to gender status classification. Despite claims of FairFace being 'fair' and 'balanced' with respect to gender, because those operations were done solely on cisgender individuals[12] that may have contributed to a major lapse in accuracy on the transgender dataset. Within both models, the transgender accuracy rates were dismal in comparison to the cisgender accuracy rates, especially in regards to transgender men[13], who were misgendered the most frequently across both models. It remains to be seen if adding transgender individuals into the datasets similar to those used to train FairFace and/or IRNv1 would result in better accuracy rates for such populations.

Interestingly, the findings from this study echoed similar research previously done by authors on the subject of facial recognition software classifying transgender individuals - with transgender women often being

[12] Within the FairFace paper, there was no notable mention of the inclusion of transgender people within their attempts to create a balanced dataset. Thus, it is assumed that because there was no identification of transgender individuals there are only cisgender people within the FairFace dataset used for pre-training.

[13] Although transgender men had the worst accuracy rates, both models generally performed worse on men within both gender status groups.

misgendered less than transgender men (Scheuerman et al., 2019). Through the results between and within models, there are noticeable differences in transgender vs cisgender accuracy rates. Specifically, significant main effects from the logistic regressions with respect to gender status were found in both the IRNv1 and FairFace models, as well as the between model comparison. When both models were taken into account, in general, cis people were 12.34 times more likely to be gendered correctly on both models than transgender people. When interpreting the Odds Ratio results from the logistic regressions, we can take them into account for effect size as well. Interpretation of the 'size' of the Odds Ratio in a similar fashion to Cohen's *d* was done following calculations from a 2010 study (Chen et al.) So, when discussing the effect that gender status had on each model and across models, we can see that for the IRNv1 and FairFace models, there was a small effect of gender status on accuracy rates. When taking into account both models, there was an extremely large effect of gender status on accuracy rates. The interaction between model and gender status also had a large effect on accuracy rate outcomes.

Within the FairFace model alone, the findings suggest that the FairFace model may have some limitations in accurately classifying gender for transgender individuals, particularly for trans men. Generally, the FairFace model's accuracy rates for gender classification were influenced more by the gender status of the individual (cisgender vs. transgender) rather than their gender identity (male vs. female). Because there was an insignificant interaction between gender status and gender identity within this model, we can say that the model's accuracy in

classifying gender status was not massively affected depending on whether the person was a man or woman.

Within the IRNv1 model alone, the results suggest that there is a large discrepancy between gender identity and gender status accuracy rates. It was found that there was a significant interaction between gender identity and gender status but not a significant main effect on gender identity. This suggests that gender identity may be a more critical factor in accurately classifying gender identity for transgender individuals within this model. However it was found that that the model performs significantly better on cisgender people as a whole than transgender people, without respect to gender identity.

As stated above, across both models there seemed to be critical differences within the accuracy rates of transgender individuals and cisgender individuals. Depending on the within-models analyses and between-models analysis, there were differences in gender classification accuracy rates. However, the large takeaway is that there are tangible biases that are seen in the models used within this study, regardless of the debiasing precautions used (i.e., balancing a dataset). This may be because of a lack of inclusion of transgender individuals, or potentially other confounding variables such as race and age. If we solely look at Figure 2, the graph shows that there are extreme differences between models and within models on transgender accuracy rates.[14] We can see that FairFace performed extremely well on cisgender people, with the best average accuracy rates of the datasets and models included. Additionally, the visual contrast

[14] Although the percentages displayed in that figure demonstrate major differences, it is important to note that the data points shown were averages between each demographic. Within the logistic regressions calculated, all data outputs from the model were included.

between transgender men's average accuracy rates and cisgender men's accuracy rates is extremely large. Lastly, there is also a large difference between average gender classification rates on women as a whole across all datasets and models in comparison to men. It is unknown as to why exactly both the FairFace and IRNv1 models on average performed worse on men's gender classification than women's. Further research is needed to understand the reasoning behind the differences discussed here.

Future Work

For future studies in a similar vein, a few recommendations are in place. Although the man/woman dichotomy was still used within this project's scope, switching a focus towards a spectrum of masculinity and femininity would be preferable. As suggested by Hamidi et al. (2018), the incorporation of gender identity and presentation as a spectrum would be beneficial in the event of inclusion of gender-diverse individuals such as non-binary individuals. Moving away from a dichotomous view of gender would not only allow for the incorporation of non-binary individuals, but would also create more ethical algorithms towards cisgender and other individuals who do not have a gender presentation that strictly resides within societal gender norms. This could potentially help with the misgendering of men within NN models, since men were among the most misgendered within almost all sectors of this study (in regards to cisgender and transgender men). Lastly, by classifying gender on a spectrum, we could better identify biases within humans as well as AI with respect to masculinity and femininity, and potentially gender as a whole.

The inclusion of transgender individuals in datasets to train ML algorithms is a must. The misgendering of transgender people in real life is already an epidemic, and it would be irresponsible to allow NN algorithms to misgender transgender people as well. By including transgender people's faces in facial recognition training datasets, we can potentially mitigate some of these biases that led to extreme misgendering within this study, especially on transgender men. More studies should be done on the opinions of marginalized groups (such as transgender and gender-diverse individuals) on the ethics and potential uses of such technology, such as the one conducted by Scheuerman et al. in 2019. When considering the ethical implications of the uses and misuses of any technology, we need to consider the opinions of the people who would be most affected (i.e., marginalized groups such as transgender people, non-white individuals, and a diverse range of ages).

It is also recommended to include more checks for biases within NN algorithms during the training and validation phases in order to potentially check such biases earlier on. If there are biases found similar to those shown within this study, such as against transgender individuals (with varying degrees based on the dataset used, however balanced it may be), we need to implement rules and procedures to thoroughly mitigate these biases. Creating tangible rules and guidelines to detect and mitigate biases within NNs, especially those used commercially or through the government, is critical.

Study Limitations

There were several limitations within this study, including but not limited to unbalanced datasets used to test the NN algorithms, racial and age groups exclusion, and tangible bias mitigation measures. Specifically, the datasets that were created using a Python API scraper contained uneven amounts of images based on gender identity and gender status. Although the cisgender dataset was more "balanced" with respect to age and gender categories than the transgender dataset[15], the number of images between datasets was also uneven.[16] Within the accuracy and logistic regression analyses, race and age were also not considered as factors. This could lead to age and/or race potentially acting as confounds within this study that could have affected analysis outcomes and thus interpretations of the results. Within the IRNv1 model, age and race were not able to be programmed correctly at the time of gender analyses, and thus were omitted from all calculations and predictions used within this study. Additionally, this study did not include specific bias mitigation measures that were recommended by several researchers within the field of ML. Specific bias mitigation measures that were noted were the inclusion of the gender variables as continuous rather than discrete and the use of bias detection programs such as InsideBias (Hamidi et al., 2018; Serna et al., 2021). There are major implications of asserting the binary gender distinctions of male/female or man/woman, especially when including transgender individuals. Within the context of this study, it was unknown how to

[15] The cisgender dataset had around a 1% difference in men vs women, and images were gathered based on racial groups. The transgender dataset had around 7.24%, and images were not balanced on race.

[16] The number of images within the cisgender dataset was 550, and the number of images within the transgender dataset was 414 - a difference of 136 images or 28.22%.

implement gender variables as a spectrum, and thus the scope of the study had to change to only include binary transgender individuals. With respect to bias detection materials such as InsideBias (Serna et al., 2021), there was not enough time to implement such programs within this study. Overall, the most intrusive study limitation was time; there was not enough time to pursue the study through the original intended scope.

Conclusion

Although dataset balancing can provide some benefits in regards to gender classification amongst gender identities, NN classification based on gender status is still lagging behind. The findings from this study highlight the urgent need for further research and development of AI models that are sensitive to the nuances of gender, especially in regards to gender classification. Additionally, we must critically examine the underlying biases and prejudices that may be ingrained in these models, and work to address and mitigate them. This research also underscores the importance of diverse and inclusive datasets for training AI models, as biased data can lead to biased outcomes.

As AI continues to play an increasingly prominent role in our lives, it is crucial that we strive to ensure that these systems are fair, equitable, and just. If biases within NN algorithms are left unchecked, marginalized groups may be severely affected within real-world applications using AI. By recognizing and addressing biases in NN models, we can move towards a more inclusive and equitable future for all individuals, regardless of characteristics such as gender identity, gender status, or gender expression.

References

1. AIAAIC - HRT Transgender Dataset. (2023). Retrieved 12 April 2023, from https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies/hrt-transgender-dataset

2. Albiero, V., S., K., Vangara, K., Zhang, K., King, M., & Bowyer, K. (2020). Analysis of Gender Inequality In Face Recognition Accuracy. Retrieved from https://arxiv.org/abs/2002.00065

3. Bhardwaj, A. (2020, October 11). What is a Perceptron? – Basics of Neural Networks. Towards Data Science. https://towardsdatascience.com/what-is-a-perceptron-basics-of-neural-networks-c4cfea20c590

4. Cao, Q., Shen, L., Xie, W., Parkhi, O., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age. https://arxiv.org/abs/1710.08092#

5. Cao, Q., Shen, L., Xie, W., Parkhi, O., & Zisserman, A. (2017). VGGFace2: A dataset for recognising faces across pose and age. https://arxiv.org/abs/1710.08092#

6. Chen, H., Cohen, P., & Chen, S. (2010). How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies, Communications in Statistics - Simulation and Computation, 39:4, 860-864.

7. davidsandberg. (2023). facenet/inception_resnet_v1.py at master · davidsandberg/facenet. https://github.com/davidsandberg/facenet/blob/master/src/models/inception_resnet_v1.py

8. DeepAI. (2019). Weight (artificial neural network). DeepAI. Retrieved from https://deepai.org/machine-learning-glossary-and-terms/weight-artificial-neural-network

9. DeepAI. (2020). Hidden Layer. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/hidden-layer-machine-learning

10. European Parliament. Directorate General for Parliamentary Research Services. (2020). The ethics of artificial intelligence: issues and initiatives. Publications Office.

11. Gong, S., Liu, X., & Jain, A. K. (2019). DebFace: De-biasing face recognition. arXiv preprint arXiv:1911.08080.

12. Google. (2023). Google Developers. Classification: Accuracy. http://developers.google.com/machine-learning/crash-course/classification/accuracy

13. Guo, Y., et al. (2015). Deep Learning for Visual Understanding: A Review. Neurocomputing.

14. Gupta, N. (2013). Artificial neural network. Network and Complex Systems, 3(1), 24-28.

15. Hamidi, F., Scheuerman, M.K., & Branham, S.M. (2018). Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

16. Instaloader. (2023). GitHub - instaloader/instaloader: Download pictures (or videos) along with their captions and other metadata from Instagram.

17. Johnson, K. (2022, March 7). How Wrongful Arrests Based on AI Derailed 3 Men's Lives. WIRED. https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/

18. Johnson, K. (2022, March 7). How Wrongful Arrests Based on AI Derailed 3 Men’s Lives. WIRED. https://www.wired.com/story/wrongful-arrests-ai-derailed-3-mens-lives/

19. Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).

20. Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. Proceedings of the ACM on human-computer interaction, 2(CSCW), 1-22.

21. Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997.

22. Meta AI Research. (2023). Face Recognition. Papers With Code. https://paperswithcode.com/task/face-recognition#datasets

23. Minsky, M., & Papert, S. (1969). Perceptrons. M.I.T. Press.

24. MIT Media Lab. (2017). Ethics and Governance of Artificial Intelligence. MIT Media Lab. https://www.media.mit.edu/groups/ethics-and-governance/overview/

25. Radiya-Dixit, Evani, A Sociotechnical Audit: Assessing Police Use of Facial Recognition (Cambridge: Minderoo Centre for Technology and Democracy, 2022).

26. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65 6, 386-408 .

27. Salas‐Eljatib, C., Fuentes-Ramirez, A., Gregoire, T.G., Altamirano, A., & Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators, 85, 502-508.

28. Scheuerman, M.K., Paul, M. J., and Brubaker, J. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. Proc. ACM Hum.-Comput. Interact. 3, CSCW: Article 144.

29. Scheuerman, M.K., Wade, K., Lustig, C., and Brubaker, J. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. Proc. ACM Hum.-Comput. Interact. 4, CSCW1: Article 58.

30. Serna, I., Pena, A., Morales, A., & Fierrez, J. (2021). InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics. 2020 25th International Conference on Pattern Recognition (ICPR), 3720-3727.

31. Wang, T., Zhao, J., Yatskar, M., Chang, K. W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5310-5319).

32. Wang, Z., Qinami, K., Karakozis, I., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2019). Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. Retrieved from https://arxiv.org/abs/1911.11834

33. Wu, W., Protopapas, P., Yang, Z., & Michalatos, P. (2020). Gender Classification and Bias Mitigation in Facial Images. 12th ACM Conference on Web Science (pp.106-114).

34. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

35. Zhang, D., Maslej, N., Brynjolfsson, E., Etchemendy, J., Lyons, T., & Manyika, J. et al. (2022). The AI Index 2022 Annual Report. https://arxiv.org/abs/2205.03468?context=cs.AI