Revolutionizing Marginalization in the Twenty-First Century : Algorithm and Data Collection Bias

Hannah Galbraith
Computer Science
Portland State University
Portland, Oregon, USA
hrg@pdx.edu

Anniemerlin Kanthavel
Computer Science & Electrical Engineering
Oregon Health and Science University
Portland, Oregon, USA
kanthave@ohsu.edu

William Allan Mass
Computer Science
Portland State University
Portland, Oregon, USA
wamass@pdx.edu

C.-M. Rutz
Computer Science
Portland State University
Portland, Oregon, USA
catrutz@pdx.edu

Chad M Tolleson
Computer Science
Portland State University
Portland, Oregon, USA
tolleson@pdx.edu

ABSTRACT
Through the examination of four key fields where machine learning algorithms are being aggressively applied to data sets which underrepresent and marginalize vulnerable socioeconomic, gender, or racial groups, this paper will present an examination of how those fields are affected by four key means of introduction of bias into those data sets and their algorithmic analysis. These means of introduction are: a lack of oversight in creating and implementing the algorithm; a lack of input or limited participation from stakeholders during the training and testing phases of the algorithm's creation; subconscious societal bias built into the data sets against which these algorithms are being trained, and the inherent difficulty in removing all bias from the training data set for any algorithm. As this paper will seek to demonstrate, these factors can be identified and potentially mitigated; however, it will also show through citation and example that allowing these factors to remain unmitigated (or even unidentified) throughout the training of the algorithm can have disasterous consequences when the algorithm is applied to its intended purpose in real world applications; the marginalization of these underserved groups can be exacerbated dramatically through the use of these algorithms.

CCS CONCEPTS
●Human-centered computing → Empirical studies in HCI
●Computing Methodologies → Machine Learning
●Computing Methodologies → Computer Vision
●Security and Privacy → Social Aspects of Security and Privacy
●Social and Professional topics → Socio-technical systems;Gender; Surveillance
●Social and Professional topics → Socio-technical systems; Criminal Justice; Surveillance
●CCS → Social and Professional topics → Computing/ Technology Policy → Medical Information Policy
●CCS → Software and its Engineering → Software Creation and Management → Designing Software → Requirements Analysis
●CCS → Software and its Engineering → Software Creation and Management → Designing Software → Software Implementation Planning

KEYWORDS
Data Bias; Algorithmic Bias; Gender; Transgender; Automatic Gender Recognition; Race; Facial Recognition and Computer Vision; Machine Learning; Natural Language Processing; Criminal Justice

ACM Reference format:
Hannah Galbraith, Anniemerlin Kanthavel, William Allan Mass, C.-M. Rutz, and Chad M Tolleson. 2019. Algorithm and Data Collection Bias. In *Proceedings of CS 410/510: Explorations of Data Science. Portland, OR, USA, 10 pages.*

1 Introduction

Over the past two decades the rise of big data and accessible machine learning algorithms have transformed our society into one in the midst of a data revolution. Because these algorithms can be used to evaluate larger data sets for more complex purposes, in less time than previously possible, it has become possible to pair a data set with an algorithm to produce results which provide definitive improvements in how people commute and travel travel, how they are entertained, or how they advertised to (this latter, of course

represents an improvement for the advertiser, and not necessarily for their target audience). Yet, the application of this combination of technologies does not stop there—corporate entities, law enforcement agencies, and medical professionals all seek to apply the inherent power and efficiency of this technological pairing to their own oeuvres. Frequently, these applications take into account only the narrowest range of ideal results, and fail to consider whether the result provides a result which is inclusive across genders, races, or socioeconomic divides. Inherent to all of these inconsistencies are four discrete means through which bias is introduced into these data sets: a lack of oversight in creating and implementing the algorithm itself; limited participation or input from stakeholders during the training and testing phases of the algorithm's creation; subconscious societal bias built into the data sets against which these algorithms are being trained (or into the idealized goal of the algorithm, or both); and finally, the inherent difficulty in removing *all* bias from the training data set for any algorithm.

Through the examination of four key fields where machine learning algorithms are being aggressively applied to data sets which underrepresent and marginalize vulnerable socioeconomic, gender, or racially divided groups, this paper will seek

2 Gender Bias in Natural Language Processing

Natural language processing (NLP) is becoming an increasingly prominent presence in our day-to-day lives. It's used to power everything from machine translation, to search engines, to AI-driven "smart"' assistants like Siri, Alexa, and Cortana. NLP systems often unintentionally reinforce societal prejudices and stereotypes. These biases can originate from any area single of a system, or from multiple areas. The datasets used to train the model (e.g. large corpora of texts from sources like Twitter, Wikipedia, or Google News), the language resources (e.g. pre-trained word embeddings or lexicons) used, or the method of training (e.g. the algorithm, features, and/or parameters used) can all contribute to models which reflect prejudices around certain groups of people [1].

Talking about how gender bias can be quantified in natural language processing algorithms may seem esoteric or abstract. However, because NLP systems are used in such a wide variety of contexts, learned biases can have real-world consequences which reinforce the alienation of historically marginalized groups, including women. Examples can include search engines that rank women engineers below their male counterparts on employment-oriented services [5] or voice AI which struggles to recognize women's voices [8]. Thus, the prejudices present in society are reflected in NLP systems, which in turn amplify those biases.

To showcase the variety of NLP systems that exhibit machine bias, we will explore three subdomains of NLP: machine translation, sentiment analysis and word embeddings.

2.1 Gender Bias in Machine Translation

Many people use machine translation (MT) tools to communicate in foreign languages. Google Translate alone gets over 200 million users daily [2]. Recently, however, concerns have arisen regarding Google Translate's reinforcement of gender asymmetries, with some decrying it as 'sexist' [2]. Marcelo Prates, Pedro Avelar and Luis Lamb conducted a study to demonstrate that Google Translate routinely translates sentences using gender stereotypes which do not reflect the actual demographics of certain positions [2]. Using a comprehensive list of job positions from the U.S. Bureau of Labor Statistics (BLS), they constructed sentences like "He/She is a <job position of interest>" in 12 different gender neutral languages such as Hungarian, Chinese, and Yoruba. They then translated these sentences into English using the Google Translate API, and collected statistics about the frequency of female, male, and gender-neutral pronouns featured in the translated output. The authors were able to show that Google Translate exhibits a strong tendency towards male pronoun defaults, particularly when subjects were in positions related to STEM fields, as well as other fields that have historically been male-dominated. Once obtained, the authors compared these statistics against BLS' data for the frequency of female participation in each job position, showing that Google Translate fails to reflect the real-world distribution of female workers.

Prates, Avelar and Lamb reference Kay and Kempton's 1984 analysis of the Sapir-Whorf hypothesis [3], arguing that one's language has a direct effect on one's understanding of the world and, at least to some extent, influences how a person perceives their reality. This suggests that languages which distinguish between female and male genders grammatically may enforce a gender bias in a person's perception of the world [2]. Prates, Avelar and Lamb also argue that moving towards gender neutrality in language and communication could help to promote gender equality, and translation tools should aim to keep texts gender neutral when possible instead of defaulting to male or female variants.

Issues of gender stereotyping spread across a range of translation tools. Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer showed that four popular industrial machine translation systems (Google Translate, Microsoft Translator, Amazon Translate, and SYSTRAN), as well as two recent state-of-the-art academic machine translation models are significantly prone to gender-biased translation

errors [4]. They took the opposite approach to Prates, Avelar and Lamb, testing all systems using two datasets composed of English sentences in which the subject of the sentence had been cast into a non-stereotypical gender role [4]. An example sentence is, "The doctor asked the nurse to help her in the operation". They created an automatic method for evaluating gender bias using eight target languages with grammatical gender (Spanish, French, Italian, Russian, Ukrainian, Hebrew, and Arabic). What they found was that all systems had very low accuracy when it came to preserving correct gender pronouns in their translations [4]. The best performing model on each language often did only slightly better than a random guess [4]. On the other hand, when sentences were inverted to feature subjects in gender-stereotypical roles (e.g. a female nurse), all systems were significantly more accurate at preserving the correct gender across all languages [4].

2.2 Gender Bias in Sentiment Analysis

Sentiment analysis tools are just as prone to perpetuating inappropriate biases as machine translation systems. Svetlana Kiritchenko and Mohammad Saif compiled a dataset of 8,640 English sentences carefully chosen to biases towards particular races and genders. The templates they constructed included sentences with emotional words such as "<Person> feels <emotional state word>" and "The conversation with <person> was <emotional situation word>". They also included templates of sentences which displayed no emotion, like "I talked to <person> yesterday". They then inserted female and male first names associated with being African American or European American, as well as words indicating one of four basic human emotions: anger, fear, joy, and sadness. Examples of African American male names they included are Alonzo, Lamar, and Malik and examples of African American female names are Ebony, Latoya, and Tanisha. Examples of European American males names they included are Adam, Jack, and Ryan and examples of European American female names are Amanda, Heather, and Stephanie. Some examples of emotional state words they used are 'annoyed', 'anxious', 'ecstatic', and 'depressed'. They used this corpus as a supplementary test set to analyze 219 natural language processing systems that participated in an international shared task on predicting sentiment and emotion intensity [1].

Kiritchenko and Saif instructed participants in the following way: given a tweet and an affective dimension A (anger, fear, joy, sadness, or just general sentiment), determine the intensity of A that best represents the mental state of the tweeter — a real-valued score between 0 (least A) and 1 (most A). The training set given to the participants consisted of tweets with labeled emotional intensity scores. Participants were given two test sets: one was a regular tweet test set and the other was the corpus that the researchers created. The first test set was used to evaluate and rank the accuracy of the systems' predictions, while the second set was used to perform bias analysis. The systems submitted used a variety of machine learning algorithms, including Support Vector Machines, LSTMs, and Deep Neural Networks to complete the task.

The study found that 75% to 86% of the submissions consistently marked sentences of one gender higher than another. When predicting anger, joy, or general emotional sentiment, more systems consistently awarded higher scores to sentences including female names than they awarded to the same sentences when the names were replaced with male names [1]. When predicting fear, most submissions tended to assign higher scores to sentences with male names [1]. These align with common gender stereotypes; for example, women are more emotional than men and situations involving men tend to be more dangerous and therefore more fearful. They found biased scores to be even more prevalent when comparing race [1].

Since sentiment analysis is used often in areas such as customer support and marketing [1, 9], stereotyping customers based on their gender or race mean could have repercussions like systems considering reviews from one race or gender to be less positive simply because of their race or gender, or customer support systems prioritizing a call from an angry male customer over a call from an equally angry female customer [1].

2.3 Gender Bias in Word Embeddings

The final area of Natural Language Processing that we survey is word embeddings. Bias in word embeddings carries with it serious repercussions, as word embeddings are frequently used in a variety of systems and their downstream applications. Thus, the presence of gender bias—or any other type of bias—in word embeddings can end up having far-reaching consequences.

Word embeddings are simply vector representations of words in a corpus [5, 6]. Despite only being trained on word co-occurrences within the corpus, they have the ability to act as a dictionary for NLP systems because words that are similar to each other (e.g. 'king', 'queen', 'duke', and 'duchess') tend to fall closer together in the vector space [5]. Because of this tendency, word embeddings are able to map semantic meaning into geometric space. NLP systems can then use these vectors to create mathematical models of the corpus.

2.3.1 Debiasing word embeddings

In their seminal work, Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai showed that they were able to clearly measure gender bias in a pre-trained Word2Vec word embedding by measuring the projection of gender-neutral words onto the "gender direction" they located within the word embedding, which

was obtained by subtracting the vector for the word 'she' from the vector for 'he' [5, 7]. The larger the word's projection on 'he' - 'she', the more biased it was [5, 7]. Another way to look at it is that the closer the gender-neutral word was to one gendered word in a set of gendered words in the vector space, the more biased it was (e.g. 'nurse' fell closer to 'female' than 'male') [5]. They showed these statistical biases aligned closely with pre-existing stereotypes [5, 7].

[5] detail the algorithms they implemented in order to correct the gender bias: by employing a series of linear algebra transformations, they zero the gender projection of each word on a predefined gender direction. They also take dozens of pairs of inherently gendered words (e.g. 'grandmother' and 'grandfather') and make sure that all gender-neutral words are equidistant from the two words in the pair [5, 7]. After performing these transformations, there was a significant improvement in the results. On the initial Word2Vec embedding, 19% of the top 150 analogies were judged as showing gender stereotypes, but after performing the debiasing algorithm described above, only 6% of the new embedding was judged as stereotypical [5].

2.3.2 Masked biases in word embeddings

Despite the rigorous and highly influential work that Bolukbasi, Chang, Zou, Saligrama and Kalai did, Hila Gonen and Yoav Goldberg show that the gender-direction described in the previous section is really only an indicator of gender bias and not the complete manifestation of it. Thus, the debiasing methods employed by Bolukbasi, et al. serve only to mask the bias; gender bias is still reflected in the relationships between key gender-neutral words, and a lot of the bias information can be recovered [7].

Gonen and Goldberg found that many gender-neutral words that were "socially marked" as feminine fell close together within the vector space (e.g. 'nurse' was close to 'receptionist', 'caregiver', and 'teacher') even after debiasing. They also found that words with strong previous gender bias were still easy to cluster together and the implicit gender of words with significant previous bias could still be easily predicted based solely on their vectors [7]. Thus, the methods employed by [5] were largely superficial. They served to mask gender bias to some degree, but nevertheless gender bias is quite systematic and deeply entrenched within the embedding.

The work of of Gonen and Goldberg shows that eliminating gender bias in word embeddings, or natural language processing systems in general, is not as simple as just mathematically eliminating bias present in texts. No easy solutions exist for dealing with inappropriate prejudices that manifest in NLP tools, and any efforts made to reduce bias could more than likely be bypassed. However, this does not absolve us from dealing with the ethical repercussions of the systems we build, particularly because eliminating gender bias in systems which are so prevalent in our lives could have the benefit of reducing gender bias in our society. We must absolutely continue to research and work toward eliminating gender bias, as well as other forms of bias, in NLP systems.

3 Gender and Racial Bias in Facial Recognition Systems

A facial recognition system (FRS) is a machine learning program that has been trained on a photo database of known individuals to identify a person or group of people in a photo or video. Identification of a subject is performed through the analysis and comparison of various biometrics, and FRS employ a variety of algorithms to collect these biometrics and then categorize the subject based on that data [10, 11, 12, 13]. These categorizations are then used to gather a set of comparison photos to be used to determine the identity of the subject in question.

Facial recognition software is also rife with problematic practices and implicit biases, which is easily identifiable through FRS's widely-publicized sub-par ability to correctly identify women, gender minorities, and people of color. Furthermore, the research behind many of the algorithms FRS use to categorize subjects according to biometrics such as gender and race is also inaccurate, outdated, and unfounded [10, 14, 15, 23, 16]. This section will briefly explain the basic processes FRS use to identify the subjects for photos, and then will explore the biases implicit in how these systems classify gender and race.

3.1 The Facial Recognition Process

The FRS process can be examined in more detail by examining the developer guide for Amazon Rekognition, a well-known facial recognition system. An image fed into a FRS will be first scanned for the presence of a face, either alone or in a crowd. An image can be rotated, scaled, and cropped by the FRS to better highlight a face [10, 13].

The facial recognition algorithm will then identify facial features by extracting facial landmarks from the image; commonly used biometrics include the distance between the eyes, the distance from forehead to chin, and the shape of the chin. One FRS searches for 68 various biometrics. The resulting data is translated into a mathematical model frequently referred to as a "face template" [AL]. Rekognition stores the face template as a FaceDetail object. The FaceDetail object contains a Location object, which includes the actual metrics collected such as the (x,y) coordinates of the left and right eyes and nose [17, 18, 19, 20].

FaceDetail also contains the categorizations performed by Rekognition based on the subject's individual biometrics. For example, 'Beard' records if Rekognition thinks the subject

has a beard and ‘Gender’ records Rekognition's guess as to the subject's gender, as well as the confidence level Rekognition had in that determination [20]. Categorizations are performed by the FRS prior to identifying the subject's face due to the sheer number of photos contained in the databases these systems use for comparisons: a single database can consist of millions of tagged and identified photos of up to a hundred thousand individuals [21]. Specific algorithms routinely used in categorization include automatic gender recognition (AGR) algorithms and automatic race recognition (ARR) algorithms [11, 22].

The face template, complete with both biometric markers and categorizations, is then compared to a database of identified photos; during this comparison, distinguishing features will be used to recognize a face in its entirety, or else the subject's individual metrics will be statistically compared to template values [19]. The end result of a FRS algorithm can take many forms: some algorithms return a variety of potential matches in the form of photos with identification, ranked in order of likelihood, while others return a single result [18]. When Rekognition analyzes an image, information about each potential match is stored in a PersonMatch object. PersonMatch stores details of the image the subject matched with, such as the match identification and the match's face template, as well as the match's face ID, the source image ID, and the confidence in the face similarity [20].

3.2 Automatic Gender Recognition Algorithms

Automatic gender recognition (AGR) algorithms are routinely used in facial recognition algorithms, and have been a large focus of FRS development in recent years [23]. As seen above, Rekognition clearly has methods in place to analyze a subject's gender and a confidence level, but examining this Gender type immediately exposes a limitation inherent to automatic gender recognition algorithms: the only "acceptable" genders according to Rekognition are "male" and "female", and this categorization is based on examining physical traits [20]. In fact, every automatic gender recognition algorithm examined during the research process for this paper treated gender as an explicit binary and used biometrics [20, 11, 12, 23]; University of Washington PhD candidate Os Keyes found that nearly 95% of AGR papers published in three major journals in the last five years treated gender as a binary [14].

The previous discussion on facial recognition algorithms established that the AGR algorithms used are based on biometrics collected from a subject's image. This use of biological traits to determine gender (which is defined as a cultural categorization based on behavior and societal role) suggests that sex and gender are conflated within FRS research and development. By contrast, sociologists who study gender have accepted that the idea that sex determines a person's social role, behavior, and physical presentation is inaccurate [14]. It is also apparent that physiological sex is treated as binary by AGR papers, which is a biological impossibility; intersex individuals, who are not biologically male or female but instead something else entirely, are just as common among the world population as people with red hair [24].

However, these facts are rarely reflected in AGR literature, and are particularly absent from the criteria AGR algorithms use to determine a subject's gender. The idea of recognizing a subject's gender through physical cues is ubiquitous throughout AGR literature, and is often attributed to such vague sources as the "characteristic differences between men and women" [12]. Such "characteristic differences" include biological metrics such as differences in gait and facial physiology, and also purely cultural behaviors such as the presence of makeup, hairstyles, mannerisms and even apparel [11, 12].

The views reflected above-that sex dictates a person's gender and that both sex and gender are binary-inherently excludes the experiences of transgender (trans) people, whose gender identities do not match the sex assigned to them at birth. The term "transgender" includes a wide array of experiences; some trans people fall within a gender binary (trans men and trans women), while others have a gender that does not fall within the binary (non-binary), is flexible (genderfluid), or do not have a gender at all (agender). This is contrasted with cisgender (cis) individuals, whose sex matches their gender [25, 14].

Additionally, the attitudes reflected in AGR papers often reflect a rigid and stereotyped binary view of gender, as is evident in the practice of categorizing an individual as male or female based on certain apparel, makeup, hairstyles, and mannerisms. While trans individuals are very underrepresented in the data sets that train FRS [15], when they do appear and a discrepancy between a physiology of one gender and clothing of another is seen, researchers jump to identifying the individual within the binary and do not consider that the person might be trans. Keyes notes in their research that one study specifically challenged the "female" categorization of an individual by an AGR as "a man in a wig!" [14]. In addition, use cases of AGR are typically binary and include stereotypical cultural behaviors of men and women, such as showing "ads of cars when a male is detected, or dresses in the case of females" [14].

The overall lack of scientifically-based research into the development of many AGRs and lack of transgender-inclusive use cases does not bode well for AGR's overall performance on transgender people, but concrete data that shows this is hard to find. The only experiment which explicitly compared AGR's performance on trans versus cis individuals was executed by the magazine Jezebel. Jezebel analyzed a dataset of 600 photos taken from The Gender Spectrum Collection, a collection of photos of trans and gender non-conforming individuals published by Vice [26],

and compared the results to a set of similar pictures of cis people taken from Shutterstock. Jezebel found that, on average, Rekognition was 10% more confident of its gender categorization on the Gender Spectrum Collection, and also highly inaccurate: misgendering occurred in 4% of images from the Shutterstock set, but jumped to 31% when analyzing images from the Gender Spectrum Collection. Rekognition's gender categorization also tended to be inconsistent between individuals in the Gender Spectrum Collection. Like many photo datasets, the Gender Spectrum Collection contains multiple photos of the same people; Rekognition categorized several of these people as "male" in some photos, and "female" in others [15].

It is not surprising that a gender recognition algorithm steeped in such inaccurate and unfounded data struggles to correctly categorize even binary trans men and trans women, and fails 100% of the time at identifying a trans person whose gender identity exists outside the gender binary.

3.3 Automatic Race Recognition

Automatic race recognition (ARR) algorithms are used in some facial recognition systems to classify the subject by race according to the subject's face biometrics. ARR algorithms compare these biometrics against a photo database consisting of people of known races. Racial categories vary according to the specific algorithm in question, but generally contain three categories comparable to White, Black, and Asian [27].

As with AGR algorithms, a lack of proper scholarship also appears frequently in ARR literature. One ARR paper accepted into IEEE in 2011 mentions that "most anthropologists historically have agreed on the existence of three relatively distinct groups" and goes on to describe the differences between these groups in such broad generalizations as having "medium lips", "small chin", and "low forehead". More exact measurements or descriptions are not provided, and the sole citation leads to a paper describing a photo database [27, 28].

Not only are ARR algorithms based on unfounded claims, FR systems as a whole are much worse at identifying a person of color than a white person. An article published in 2010 mentions that a digital camera sold at the time, and which included facial recognition software as a means of keeping faces in focus in pictures, often told those of Asian descent to "open their eyes" [29], and Joy Buolamwini, an AI researcher at MIT, founder of the Algorithmic Justice League, and a woman of color, has often noted that facial recognition systems she has used or even helped develop recognize her lighter-skinned colleagues and friends more easily than her. One system failed to identify Buolamwini's face, but had little trouble identifying a face Buolamwini had doodled on her hand as well as a featureless, plain white mask [29].

The difference in accuracy between categorizations of white faces vs the faces of people of color has been widely noted in publications in the media as well as major journals and independent organizations [10, 30]. Buolamwini worked with an undergraduate student at the University of Toronto, Deb Raji, to quantify just how large this disparity was in certain FR systems; they found that while virtually 100% of white men were identified correctly, that number dropped to 71% for white women and down to 69% for black women. Raji notes that this discrepancy could be caused by training datasets presented to the FR system that contained a much larger proportion of white men when compared to women and people of color [31].

3.4 Causes of Bias and the Ramifications of Bias

Although the failure of FRS to accurately identify certain populations versus white men is readily apparent, the root cause of this disparity is harder to uncover. AGR and ARR algorithms used in facial recognition systems are often proprietary, and as such their internal workings are hidden from public examination. Very few systems are tested for internal biases, and limited academic research into the causes of the biases in FRS has been undertaken [16]. In the case of AGR algorithms, even getting academic recognition of the biases in AGR algorithms and corresponding negative ramifications on the populations in question has proven to be a nontrivial task [14]. Some areas of improvement have been noted, however, and are discussed below.

3.4.1 Biased Datasets

Merely the collection of datasets for use in training FRS has proved to be problematic. Georgetown Law estimates that as many as half of Americans' faces are included in FBI datasets sourced from DMV photos, immigration records, and prison population photos [16], and many datasets are comprised solely of images included without their subject's knowledge or consent [32, 33]. Such practices are common, with several of the largest datasets consisting of images taken in mass web-scrapings. As the sourcing and use of these datasets is completely unregulated, however, these practices are unlikely to change [16].

For example, a study partially funded by the FBI and the US Army used images of transgender youtube personalities taken from transition timeline videos in order to train a FRS to recognize a person through a gender transition, ostensibly to thwart possible terrorist attacks [32]. These pictures were taken without the subject's knowledge or consent, and were used in a way that was very harmful and violating to the transgender individuals involved. The result of this study is a facial recogniition system that is now better at identifying a transgender person, and the application of this system to commercial usage is likely to be used in such a way that increases the amount of harassment that transgender people

already receive, and is espeically dangerous for vulnerable transgender peopulations such as transwomen of color, who are disproportionally the victims of police violence [25, 14].

Even when data sets are ethically obtained, too often do these datasets contain a disproportionately high amount of white cisgender males, and are therefore more likely to misidentify females, trans people, and people of color. For example, Microsoft recently declined to sell its facial recognition software to a California police department, citing human rights concerns as the software was trained primarily on white men; similarly, researchers have theorized that Amazon Rekognition's particularly high error rate results in part from a heavily skewed data set [34, 35]. Studies have shown that feeding a FRS more pictures of racial minorities and women improve the accuracy of the systems, and that facial recognition systems are more likely to correctly identify the majority race of the system's country of origin [10, 34]. However, this has not stopped Amazon from selling Rekognition to the Washington County Sheriff Department, or Microsoft from providing its FRS to an unidentified prison [36, 35].

3.4.2 Organizational and Societal Bias

Attempts to eliminate bias have also been hampered by a refusal by some prominent FR professionals and researchers to recognize that the inaccuracies currently seen in these algorithms are likely due to biased algorithms and skewed training sets, and are not findings that justify personally held biases. Patrick Grother of the US Department of Commerce's National Institute on Standards and Technology proposed in 2018 that the poor performance displayed by FR systems across the board when attempting to identify and classify black men and women was due to "nature: Female faces are naturally more similar to each other [sic]" and "photos of black subjects' faces are naturally more similar to each other [sic]". Failing that, Grother suggests that the algorithms used in FRS could be misidentifying black men due to photo artefacts, and misidentifying women due to their habit of changing makeup and hairstyles day-to-day [30].

Buolamwimi believes that the facial recognition community's failure to commonly identify and agree on the presence of bias is due to a failure to identify that commonly held goals of the facial recognition community may be problematic in and of themselves. If an algorithm meets certain accuracy and ability benchmarks established by the facial recognition community, it is considered to be performing well. However, Buolamwini states that the representativeness of the benchmarks has not been questioned, and so algorithms are not being tested in ways that could expose their biases [37].

3.4.3 Ramifications of Eliminating Bias

When considering whether it would be better for society if FR systems, AGR algorithms, and ARR algorithms were completely unbiased, one must first consider their use case. Though facial recognition systems are used in a variety of relatively benign contexts, such as advertising or children's mobile apps, they can also provide powerful tools for control and surveillance. Half of all Americans-nearly 120 million people-have their photos stored in unregulated databases used by local, state, and federal government. US Customs and Border Protection uses facial recognition at the US-Mexico border, as well as customs checkpoints in domestic airports [36, 38, 33, 16].

Facial recognition systems can recognize up to 100 people in a single image, and are capable of identifying, tracking, and analyzing people in real time. The ACLU of California discovered that Amazon's marketing tools for Rekognition, released earlier this year, specifically mention deployment by law enforcement agencies as a "common use case", citing Rekognition's tracking and identifying capabilities of multiple people in a large crowd. This technology enables governments to enact unparalleled mass surveillance [18, 39].

This is concerning, especially considering that facial recognition systems perform so poorly in correctly identifying minority populations that typically receive a disproportionate amount of law enforcement attention; increased use of FR systems as they currently are could lead to increased quantities of false arrests of women, transgender people, and people of color. However, eliminating the bias seen in FR systems is not necessarily the societal fix it may seem to be at first glance, as FR technology would then become more efficient at identifying the targets of law enforcement, who are disproportionately the same minority populations being disadvantaged now [14, 15].

There exists no panacea to cure the biases inherent in facial recognition systems due to the ways in which facial recognition software is being employed, and the problem is unlikely to improve unless the societies and organizations developing and using facial recognition technology become less biased themselves. Until that time, the effects of bias cannot be completely eliminated from facial recognition algorithms and their usage.

4 Algorithmic Bias in the Criminal Justice System

As previously asserted, there are four common problems causing algorithms to perpetuate societal biases: lack of oversight in the creation and implementation of algorithms, limited input from stakeholders during the creation and maintenance phases, societal bias built into the training data, and the inherent difficulty in removing all bias from algorithms. At least one domain expert, Cathy O'Neil, has produced several works talking about these issues in detail. Her book, "Weapons of Math Destruction" goes into detail about the various ways that algorithms, implemented incorrectly, have negative consequences for the population.

[40, 41, 42, 43, 44] Now we will inspect these problems within the world of criminal justice. There are three areas in criminal justice for which we'll discuss the implications of algorithmic bias: the allocation of law enforcement resources, assisting law enforcement officers in the field, and providing sentencing guidelines in the courtroom.

4.1 Allocation of Law Enforcement Resources

4.1.1 Lack of oversight and stakeholder input

The lack of oversight in the creation and implementation of algorithms is a common theme we will see in all three of these divisions of law enforcement as they are being considered here. They manifest themselves in this area via the development of program criteria without the input of non-profit groups representing the interests of historically underserved or over-policed people and neighborhoods. These interested parties are very much aware of the shortcomings of some policing policies, and can advocate for the development and implementation of more balanced systems.

4.1.2 Effects of resource allocation styles

There are several styles of algorithms for the allocation of police resources. Some systems attempt to allocate resources-which we will consider solely as the number and frequency of police patrols-amongst neighborhoods corresponding to their historical crime rates. Two other methods include an equal allocation of patrols across all neighborhoods, and devising a system with equal response times for all neighborhoods in the community. [45, 46] The latter two methods are considered more equitable for all populations within the community, especially the equal response time approach. A study performed by Indiana University and the University of Maryland found that the method of applying police resources based on current crime rates has some unexpected consequences. A common sense approach might lead to an allocation of more police resources to neighborhoods with the highest crime rates. A study performed by Indiana University and the University of Maryland have determined that such an approach will indeed reduce the crime rates in those neighborhoods, but with the unintended side effect of causing a rise in crime rates in other neighborhoods. Notionally this is due to people prone to criminal behavior relocating to other neighborhoods to avoid police scrutiny. This study determined that a more effective approach was to allocate police resources evenly across all neighborhoods, and then increase overall police levels if crime rates were a concern. This prevented the anomaly where crime rates rise and fall inversely related the corresponding policing levels. [45] The equal response time approach is more of a reactive system ensuring all neighborhoods receive equitable treatment based on their citizens reporting criminal activity. [46]

4.1.3 Historical bias persists

As regards algorithmically enforced bias, we are concerned about the approach of allocating resources based on historical crime rates. To start, let us assume that there is an established racial bias in historical arrest records in the United States. This is a debated theory which falls largely outside of the scope of this paper, however numerous papers support this theory, and while they will not be listed here, active links to many of them may be found on Wikipedia [47]. The problem is that certain minority populations are disproportionately represented in arrest and conviction statistics in relation to their proportion of the general population. This over-representation holds true even when adjusting for economic factors and known rates of criminality across all demographics. Also, consider the fact that a larger proportion of these same minority groups are in lower economic brackets, which in turn is known to have a correlation with higher arrest and conviction rates. These two points suggest that these minority groups are overrepresented in historic crime statistics in two different ways. Considering that there is a general tendency for people of the same socioeconomic groups to live in the same neighborhoods, we can clearly see that in a given community there will be some neighborhoods with disproportionately (or fairly represented) higher crime rates. Therefore, if a police department were to allocate their resources (patrol routes and frequency) based on these historical crime rates, then the people living in these neighborhoods will be much more likely to interact with police, and this amplified quantity of contact with the police means the crime statistics are further biased against these populations.

4.1.4 Removing bias affects style

Interesting to consider, if all of the main and proxy variables correlated with racial and economic bias are removed from the training data, the algorithm may recommend allocation of police resources evenly across the entire jurisdiction. This has the potential to resemble the allocation styles in which resources are intentionally allocated evenly across the jurisdiction with the intention of hampering criminal behavior (as opposed to encouraging the criminals to modify behaviour simply to avoid police interaction). [48]

4.2 Assisting Law Enforcement in the Field

4.2.1 Lack of oversight and stakeholder input

As stated in the discussion of police resource allocation, by involving interested stakeholders representing the different subpopulations of the general population. Had developers and owners of this software solicited and incorporated input from these stakeholders, they may have anticipated and prevented some of the problems (to be discussed) that led to the current growing trend of jurisdictions banning the use of facial recognition software. [48, 49, 50]

4.2.2 Stakeholders sometimes affect implementation

Different purveyors of these facial recognition software programs have displayed varied levels of responsibility around their products. On one hand Microsoft who has

refused to sell their facial recognition software to law enforcement agencies, out of concern for false positives imposing on the rights of civilians, and general abuse of the tool by individuals within these law enforcement agencies. However, Microsoft still contracted the software out to at least one prison system and the U.S. Department of Immigration and Customs Enforcement [51] Similarly, at least one manufacturer of body cameras used by police departments, Axon, has explicitly declined to include facial recognition software into their cameras, again for fears of the impairment of civil rights and potential for abuse. [52] One the other end of the spectrum, Amazon distributes its software with many police departments and other agencies both domestic and foreign with seemingly little worry of the repercussions. The potential for abuse has not gone unnoticed by some company stakeholders. In fact, employees and shareholders of Microsoft and Amazon have clearly asserted that the facial recognition software these companies create should not be used in the criminal justice field, or in other governmental arenas at this time, or perhaps ever. Amazon, for one, seems to have made few, if any, changes in their business practices due to the stakeholder complaints. [51, 54, 55]

4.2.3 Bias towards vulnerable groups

The biases observed with facial recognition software are myriad. One known problem with the training data is how it relies on questionable sources. For example, the US Department of Commerce's National Institute of Standards and Technology's Facial Recognition Verification Testing Program (which is America's gold standard in facial recognition accuracy verification) is predominantly reliant on images of: images of child pornography victims, U.S. visa applicants (primarily from Mexico), arrest images of people who are now deceased, and booking photos. None of these groups are able to give their consent to be included in the test data, nor were they asked to provide it. The implication here is that these populations, who are inherently more at risk than the average U.S. population, are over represented and thus more likely to be surveilled by the police. As previously stated, since there is an overrepresentation of black Americans in the mugshot dataset due to policing bias, they too are more likely to be surveilled by the police. [53]

4.2.4 Rekognize racial and gender bias

The most notorious, yet prolific, of these programs is Amazon's Rekognition FRS. Over a number of tests, it has been determined that people with darker complexion and features that are more feminine are more likely to have false positives when run thru the Rekognition. As mentioned earlier, the effect is amplified even further for non-binary persons. [54] This bias may be caused by non-representative training data, or by algorithms not sufficiently trained or properly implemented by the software's developers or the groups to whom the software is licensed, or both. Regardless, the effect is the same in that members of certain groups are more likely to be misidentified. If they are then misidentified as a person with an outstanding warrant, they are more likely to be falsely arrested. Unfortunately, arrest records, even in the case of false arrests, are difficult to remove from people's records. Arrest records could be used (legally or not) in many circumstances, such as job background checks, and a false positive from a facial recognition program may have a lasting impact on the victim's life.

As an example of just how poorly the Amazon Rekognition program performs, when given a dataset comprised solely of members of the United States Congress, Rekognition falsely matched 28 members to publicly available mugshots. [55] The software could be set to display matches where a very high probability of an accurate match is found, say 99+%. In reality, the program often simply displays multiple possible matches regardless of the positive match confidence level. [56]

4.2.5 Removing bias requires work

As we've seen in other affected areas, the bias in these algorithms is remarkably difficult to remove. Even if the training data is perfectly balanced to accurately reflect the total population, the algorithm will still produce biased results, due to programming choices, or oversights within the algorithms themselves producing incorrect results. Developers of the algorithms need to perform ongoing head-to-tail reviews of the systems surrounding the algorithms and of the algorithms themselves to ensure consistent and fair treatment of marginalized populations. One group of researchers at Google are developing a system called "InclusiveFaceNet" that "detects face attributes by transferring race and gender representations learned from a held-out dataset of public race and gender identities. Leveraging learned demographic representations while withholding demographic inference from the downstream face attribute detection task preserves potential users' demographic privacy while resulting in some of the best reported numbers to date on attribute detection in the Faces of the World and CelebA datasets" [57] So while removing the bias fully is extremely difficult, with some ingenuity and effort it is possible to produce results which both reduce bias and increase the accuracy rating of the FRS.

4.3 Providing Sentencing Guidelines in Court

4.3.1 Lack of oversight and stakeholder input

Once again, when it comes to algorithms used to assist judges with sentencing guidelines in the courtroom, the historic bias against certain minorities and lower economic groups is maintained and amplified. In addition, again, much of the oversight and input from stakeholders (other than the developers and the courts) consists mostly of organizations such as the ACLU raising concerns and challenging bias results after-the-fact. Which would be largely avoidable with proper input in the development stages. Even worse than the facial recognition software programs, many of these

sentencing guidelines / recidivism rate algorithmic programs operate in black boxes where there is no visibility for outside parties. The developers of these programs claim, and perhaps rightly so, that to do so would expose their trade secrets to competitors. There could and should be, however, some ability for an entity entrusted with oversight to audit the results of a person's algorithm score when requested. [58] This sentiment is discussed in detail in Megan Garcia's book Racist in the Machine, which is definitely worth reading a summary if not in its entirety. [60]

4.3.2 Some states exercise oversight

Many states have performed some level of review of the implementation and usage of sentencing recommendation algorithms. Unfortunately, not all states have performed reviews, and not all reviews are performed to the same degree of thoroughness. [62] A standardized countrywide review would provide a foundation from which individual jurisdictions could look at the impact of different implementation methods, and then tailor their systems to be more equitable.

4.3.3 Racial and economic bias affect scoring

These programs work by analyzing many aspects of a person's life and predict their chances of recidivism. In theory, this is great; if a person is more likely to commit a crime in the future then they should receive a longer sentence. By this point, the problems should be obvious to the reader. Historic biases towards certain minorities and to those in lower economic groups are maintained and amplified in these algorithms. For example, if a person is living in or near poverty then the algorithms will rate that person as having a higher risk score. Similarly, if a person lives in a neighborhood with high crime rates (which happen to reflect racial divides, and which again happen to correspond with economic levels) then they receive a higher risk. This has led to some confusing and apparently biased results. There is a widely documented case where a black woman and white man stole property of about the same value, but the white man received a much lower risk score compared to the black woman. Despite the fact that she had a relatively sparse history with the police, opposed to his long criminal record that included armed robbery. [58, 59]

4.3.4 Risk scoring accuracy problems

ProPublica performed a review of the predicted rates of recidivism for 7,000 people arrested in Broward County, Florida. They determined that the algorithm correctly predicted those who would commit additional crimes within the next two years only 20% of the time. [58] Considering how inaccurate the algorithm proved to be, one has to ask if its relevance should be diminished if not made obsolete. In fact, these algorithms are already relied on to widely different degrees.

4.3.5 Inconsistent application

Software programs designed to assist in suggesting appropriate sentencing guidelines in courts are intended to be just one aspect judges use to determine their sentences. The usage of these programs are often times left to the individual judges' discretion. As with all aspects of life, some judges use the software appropriately, some use it exclusively without any other input, and some judges ignore the recommendations entirely. [61]

4.3.6 Removing bias doesn't have to be hard

Again, the bias in these algorithms is difficult to remove via training data cleanup alone. In this arena, the algorithms themselves could be tweaked to level out the risk scores and sentencing guidelines. For example, the algorithm can be designed to run the analysis multiple times whereby all the variables hold the same except for those that differentiate for race and economic status. Then the risk score returned to the user (judge or court official) either the: most lenient, most severe, or average version of the score produced from the different scenarios. This modified score would be applied to all people, regardless of their racial or economic status In this manner much of the bias for certain sub-populations would be minimized as it would help ensure that people from different sub-populations are treated equally.

4.4 Bias in Criminal Justice Summary

To summarize our inspection of algorithmic bias in the criminal justice system, there is a dearth of oversight and stakeholder input for algorithm usages. The historical bias towards certain minorities and economic classes are baked into the very data used to train these algorithms. As a result, these algorithms are perpetuating and even amplifying the biased treatment towards these groups. Removing the bias in the training data is somewhat, but not totally, possible. To truly remove the bias from the algorithms, developers need to both ensure the training data is balanced, and also ensure that each component of the algorithms themselves produce unbiased results at each stage.

5 ALGORITHM BIAS IN THE MEDICAL FIELD

AI and machine learning are transforming the healthcare industry, potentially changing outcomes for good, revolutionizing the way doctors provide health care. AI systems can collect data over time, access the data stored in other computers and go over the data written on the internet, medical journals and research notes in a matter of seconds and make an educated decision based on all the data it’s been over. It helps physicians with fast and accurate diagnostics of the diseases in patients. AI may be most effective at reducing human error.AI assisted robotic surgery allows doctors to perform complex procedures with greater control than conventional approaches. With the emerging technologies patient can get doctor assistance without visiting the hospitals/clinics which results in cost cutting. From interacting with patients to directing patients to the most effective care setting, virtual nursing assistants are available

24/7 to monitor patient's health and providing wellness checks through voice and AI.

Conversely, the use of AI in healthcare raises some important clinical safety and ethical questions. Certain machine learning models (e.g. deep learning) are less transparent than others and harder to interpret. It is very important in clinical decision making to understand how these algorithms arrive at their conclusions for better patient outcomes. AI systems are not capable of adapting the developments or changes in medical policies, as they are trained using the historical data. Biased data that are used to train the AI model can exacerbate the current health disparities in healthcare.

5.1 Racial Bias in Healthcare

5.1.1 Basic instances of bias in Healthcare

Machine learning algorithms are capable of synthesizing and interpreting the data in the medical record effecting an improvement of clinical decision support to guide medical diagnosis and treatment. However, these algorithms are subject to the same four biases which appear above, and which additionally include biases related to data missing from the medical record, insufficient sample sizes of the patient, and misclassification or measurement error due to implicit biases [63]. Most of the problems are caused by overreliance on machines, algorithms trained on biased data and algorithms that do not provide clinically meaningful information.

Algorithmic Risk Score is an algorithmically generated value, which reflects the level of risk in the presence of some risk factors, including patients' chronic symptoms of disease or risk of mortality etc. It takes data about the health factors of an individual patient and assigns each piece of information a different weight. Then the risk algorithm aggregates the weight for the chosen target variables and produces an output risk score. These risk algorithms are managed by multiple risk-based contracts [64], each of which might calculate risk using a different algorithm. Thus, even though two patients may have very similar medical conditions, the health care system might charge differently for their care. Ziad Obermeyer and Senthil Mullainathan's 2019 paper *Dissecting the Racial Bias in an Algorithm that Guides 70 Million People* discusses this algorithm at length, and notes that it is already deployed widely across the American healthcare system. [AK3]

5.1.2 Algorithmic Care Management for Complex Patients

One of the major problems in the healthcare system is the poor management of complex patients, defined as patients with multiple chronic conditions, who are on many different medications and who visit multiple doctors or specialists-who may provide contradictory diagnoses of the underlying issues. These patients commonly have multiple emergency visits and lengthy hospitalizations when their health deteriorates.

Care management programs offer patients with high complexity needs a dedicated phone service available 24/7, trained nurses to help with the medication and help get the appointment of specialists as needed. The outcome of care management programs is generally positive, it identifies patients before their health conditions worsen and provide the needed help. Because of the cost of allocation of resources for these patients, health care systems rely on algorithms-using the previously defined risk scores-to determine whether a patient qualifies based on their predicted future health needs, and whether it is predicted that they will benefit from the program. The algorithm also predicts next year's cost and predict those using data from this year using the insurance claims data. In the study by Obermeyer and Mullainathan, the algorithm is applied to patients admitted at a primary care hospital, which produced a dataset consisting of 40000 patients, out of whom only 12 percent are black. Obermeyer and Mullainathan focus on the policy effect of the algorithm; which patients get enrolled in the program as a measure of bias. Other health care systems purchase this evaluation software from the manufacturer, and run it on the data of the patients in their own primary care practices.

Patient data is run once or twice a year and a risk score S is generated, which is used as a screening mechanism for who will get into the program. If the patients are at or above the 97th percentile, they get auto-identified and enrolled into the program; if the score is between the 55th – 97th percentile, the patient's data are collected and presented to their primary care physician for recommendation. However, for the same risk score ratings, blacks and whites have very different realized health. For example, the highest-risk black patients (those at the threshold where patients are auto-enrolled in the program), have significantly more chronic illnesses than white enrollees with
the same risk score.

How does the bias creep in? The algorithm is given a data frame with (1) Yit (label), total medical expenditures ('costs') in year t; and (2) Xi,t−1 (features), fine-grained care utilization data in year t−1 (e.g., visits to physicians, cardiologists, number of x-rays, etc.). The predicted risk of developing complex health needs is thus in fact the predicted costs. Both hospitals and insurance companies focus on the cost factor of their patients when thinking about their patients 'health care needs'. These calculations invariably find that blacks cost more than whites on average; First, blacks bear a different and larger burden of disease, making them costlier. But this difference in illness is offset by a second factor: blacks cost less when holding at a constant their exact chronic conditions, a force that dramatically reduces the overall cost gap. So, the fact that blacks cost less than whites conditional on health means an algorithm that predicts costs accurately across racial groups will necessarily also generate

biased predictions on health.

The root cause of this bias is not in the procedure for prediction,
or the underlying data, but the algorithm's objective function itself. It arises from the choice of variables selected to enroll patients into the program and not from the measurement of these variables. In this case the financial motives of the healthcare industry workers acts as a detrimental health factor for the black patients.

5.2 Sex and Gender Differences in Healthcare

Sex and gender play a role in how health and disease affect individuals. Sex differences in medical field calculations are based on biological factors, including reproductive function, concentartions of sexual harmones, the expression of genes on X and Y chromosomes. Gender is associated with behaviour, lifestyle and life experience of the individual [66]. Because sex and gender affect a wide range of physiological functions, they have a impact on a wide range of diseases including those of the cardiovascular, pulmanory and autoimmune systems.

For example, Osteoporosis is more prevalent among women, because they have less bone tissue than men and experience a rapid phase of bone loss due to the hormonal changes around menopause; it is therefore considered a 'woman's disease.' Because of this designation, it can go undetected in men who are older than 50[67]. Women are twice as likely as men to experience depression, with some women experiencing mood symptoms related to hormone changes during puberty, pregnancy and perimenopause. However, women are more likely to admit to negative mood states and to seek treatment for mental health issues than men are, so again the gap in frequency may be partially explained by underreporting. Women and men have different symptoms during a heart attack: for both men and women, the most common heart attack symptom is chest pain or discomfort. However, women are more likely than men to have shortness of breath, nausea and vomiting, fatigue and pain in the back, shoulders and jaw.

Despite the wealth of data on the differences, medical practice does not take gender into account in diagnosis, treatment or disease management. This lack of consideration of sex and gender differences in common diseases causes a qualitative gap when dealing with particular issues for one gender or the other, and renewed effort must be maed to address this gap in order to improve health and healthcare for everyone. Researchers should include sex and gender differences in research questions and research design, obtain a gender sensitive dataset. Doctors should opt for sex and gender sensitive optimized treatment strategies in life threatening cardiovascular diseases. Finally, Health insurance companies should offer long term prevention programmes, tailored to the biological differences inherent to sex and the lifestyle choices inherent to gender differences.

5.3 Causes of Bias in Healthcare using EHR Data

Machine learning algorithms are capable of synthesizing and interpreting the data available in the medical record. Integration of machine learning with clinical decision support tools such as computerized alerts and diagnostic support offer physicians and people who work in the healthcare targeted and timely information that can improve the clinical decisions [63]. However these algorithms may be subject to bias which could exacerbate the disparities in healthcare.

5.3.1 Missing Data

Machine learning algorithms use the data available in the medical record or data derived from communicating sources like sensor and patient reported data. If data are missing, are not accessible or represent metadata (such as the identity of a notewriter) , then the machine learning algorithms may misinterpret the available data. Future comparison using the algorithm would not offer any benefit to people missing from or not acknowledged in the training dataset.

Studies have found that individuals from vulnerable populations including those with low socioeconomic status, psychosocial issues and immigrants[68] are more likely to visit multiple institutions to receive healthcare. Clinical decision support tools that identify patients based on having a certain number of encounters with a particular disease code or medication will be less likely to find patients who have had the same number of visits across different healthcare than those who receive their care in one system.Such patients may have insufficient information in the health record to qualify for disease definitions in the tool. If an algorithm cannot observe and identify certain individuals, the machine learning model cannot assign an output to them, which widens the existing disparity.

5.3.2 Sample Size

Even if data are available for certain group of patients, insufficient sample sizes may make it difficult for the medical data to be interpreted through the machine learning model. Low sample sizes and underestimation of minority groups are a common issue in randomized controlled trials and genetic studies.

Genetic studies have been criticized for not fully accounting for genetic diversity in non European patients. Studies [69] found that looking at 2511 studies , 81 percent of participants in genome mapping studies were of European descent. As the number of populations included in a study increases, the number of variables which must be controlled for also increases. In trying to keep things as simple as possible geneticists favor existing cohorts, such as that of the Framingham heart study or other large datasets generated by

well established medical centers. Such organizations collect samples and information from people in the same geographic location. Repeated sampling of the same population, perpetuates and exacerbates the disparity issue.

5.3.3 Misclassification or Measurement Error

A common source of error in observational studies are caused due to errors by practitioners. Healthcare providers are as susceptible to unconscious bias as everyone else. Many non-medical factors influence medical decisions, including a patient's style of dress, their race, ethnicity, gender, insurance status, and the clinical setting. For example, uninsured patients receive substandard medical care more frequently than those with insurance. Individuals with low socioeconomic status may be more likely to be seen in teaching clinics, where documentation or clinical reasoning may be less accurate or systematically different than the care provided to patients of higher socioeconomic status. If patients receive differential care or are differentially incorrectly diagnosed based on socio-demographic factors, algorithms will reflect the practitioner bias and misclassify patients based on those factors.

5.4 Elimination of Bias in Healthcare

Many of the problems occur due to the overreliance on machines, algorithms based on biased data, or algorithms that do not provide clinically meaningful results. Computer scientists and bioinformaticians together with medical practitioners and biostatisticians should outline the "intent behind the design of the tools,"[70] including choosing appropriate questions and settings for machine learning use, interpretation of the findings and how best to conduct meaningful follow-up studies. Clinical decision support algorithms should be tested for the potential introduction of discriminatory aspects throughout each stage of data processing. Race/ethnicity should be captured in the medical record to avoid bias later in the model. Feedback loops should be designed to monitor and verify the machine learning output and validity.

Efforts to measure the utility of machine learning should focus on the demonstration of clinically important improvements in the relevant outcomes rather than strict outcomes like accuracy or area under the curve. While both accuracy and efficiency are important, ensuring all races/ethnicities and socioeconomic levels are adequately represented in the model is necessary to ensure a homogenous level of accuracy in general use for any algorithm.

Machine learning algorithms have the potential to revolutionize the medical field by providing clinical decision support based on the predictions of the model. However, attention should be paid to the data that are being used to produce these algorithms. Existing health care disparities should not be amplified by thoughtless or excessive reliance on machines.

CONCLUSION

We have seen that across several disciplines, using several different machine learning implementations, that the biases inherent in societal treatments of underrepresented or marginalized groups are easily transferred from data set to algorithm to practice. While it is difficult (or perhaps impossible) to remove all the bias from a data set, there are means of mitigating the bias which will allow these algorithms to perform more fairly toward all the stakeholders they are evaluating. By considering the communities affected by the algorithm, data scientists can ensure that data sets are utilized which provide more accurate representation to those communities. The engagement of community advocacy groups, and the continued scrutiny of these issues by groups such as the ACLU is integral in the remediation of the bias, and of bringing these issues to the wider community. Finally, we can work to acknowledge and attempt to change inherent societal biases. By addressing theese biases at their root and working to change them in society, we can slowly but effectively work toward their removal from data sets. It must be noted that while this is the most difficult means of addressing this issue, it is likely one which will have the greatest net effect on eliminating the issue of data bias.

The use of machine learning in data science has become such a ubiquitous method of data analysis that, without any governing body providing oversight into it, is beginning to create data privacy and surveillance issues which can not be ignored. Three municipalities—San Francisco, California; Somerville, Massachussets; and most recently Oakland, California—have banned the use of facial recognition software by government entities and law enforcement agencies[71, 72, 73]. Other municipalities—most recently Portland, Oregon—are passing data privacy resolutions which limit not merely how facial recognition software can be used by agencies in the city, but outlines the responsibilities of the city toward those individuals from whom it collects data as well. Although this is a step in the right direction, and it's laudable that city governments are being forward thinking in outlining their responsibilities, allowing this level of decision-making to rest in the hands of individual municipalities will create a patchwork of local-level data privacy laws, leaving individuals unaware or unsure of what to expect. A better overall solution might be implementing a federal-level data fiduciary policy, which mandates that corporations and agencies exercise a responsibility to use your data in your interests. It's important to note that the concept of data fiduciaries should provide a *minimum guaranteed* amount of data privacy for everyone, and that individual municipalities could still opt to legislate stricter data privacy guarantees than a data fiduciary might provide.

ACKNOWLEDGMENTS

We would like to thank Dr. Kristin Tufte, the instructor for this course, for the invaluable advice, ideas, and resources she provided for this project..

REFERENCES

[1] Svetlana Kiritchenko and Mohammad Saif. 2018. Examining gender and race bias in two hundred sentiment analysis systems. arXiv:1805.04508.
Retrieved from: https://arxiv.org/pdf/1805.04508.pdf

[2] Marcelo Prates, Pedro Avelar, and Luis C. Lamb. 2019. Assessing gender bias in machine translation — A case study with Google Translate. arXiv:1809.02208v4.
Retrieved from: https://arxiv.org/pdf/1809.02208.pdf

[3] Paul Kay and Willett Kempton. 1984. What is the Sapir-Whorf hypothesis?. *American Anthropologist*, 86, 1 (Mar. 1984), 65-79. DOI:
Retrieved from: https://doi.org/10.1525/aa.1984.86.1.02a00050

[4] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. (2019). Evaluating gender bias in machine translation. arXiv:1906.00591. Retrieved from: https://arxiv.org/pdf/1906.00591.pdf

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. arXiv:1607.06520. Retrieved from: https://arxiv.org/pdf/1607.06520.pdf

[6] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115, 16 (Apr. 2018), E36335-E3644.
DOI: https://doi.org/10.1073/pnas.1720347115

[7] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender bias in word embeddings but do not remove them. ArXiv:1903.03862.
Retrieved from: https://arxiv.org/pdf/1903.03862.pdf

[8] Joan Palmiter Bajorek. 2019. Voice recognition still has significant race and gender biases. *Harvard Business Review* (May 2019).
Retrieved from: https://hbr.org/2019/05/Voice-recognition-still-has-significant-race-and-gender-biases

[9] Jia Wertz. 2018. Why sentiment analysis could be your best kept marketing secret. *Forbes* (Nov. 2018).
Retrieved from: https://www.forbes.com/sites/jiawertz/2018/11/30/why-sentiment-analysis-could-be-your-best-kept-marketing-secret/#4ed056c32bbe

[10] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. 2012. Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security 7, 6 (December 2012).

[11] C. R. Vimal Chand. 2010. Face and Gender Recognition Using Genetic Algorithm and Hopfield Neural Network. Global Journal of Computer Science and Technology 10, 1 (April 2010), 2-3.

[12] Feng Lin, Yingxiao Wu, Yan Zhuang, Xi Long, and Wenyao Xu. 2016. Human Gender Classification: A Review. In International Journal of Biometrics (IJBM), 8, 3/4 (2016).

[13] Luan Tran, Xi Yin, and Xiaoming Liu. 2018. Representation Learning by Rotating Your Faces. In IEEE Transactions on Pattern Analysis and Machine Intelligence, September 2018. Retrieved August 7, 2019 from https://arxiv.org/pdf/1705.11136.pdf.

[14] Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. In Proceedings of the ACM on Human-Computer Interaction, 2, CSCW. Jersey City, NJ. Retrieved August 4, 2019 from https://ironholds.org/resources/papers/agr_paper.pdf.

[15] Anna Merlan and Dhruv Mehrotra. 2019. Amazon's Facial Analysis Program Is Building A Dystopic Future For Trans And Nonbinary People. Jezebel. Retrieved August 4, 2019 from https://jezebel.com/amazons-facial-analysis-program-is-building-a-dystopic-1835075450.

[16] Clare Garvie, Alvaro M. Bedoya, and Jonathan Frankle. 2016. The Perpetual Line-Up: Unregulated Police Face Recognition in America. Center on Privacy and Technology report. Georgetown Law, Washington, DC. Retrieved August 5, 2019 from https://www.perpetuallineup.org/.

[17] Symantec Corporation. 2019. How Does Facial Recognition Work?. Security Center. Retrieved August 4, 2019 from https://us.norton.com/internetsecurity-iot-how-facial-recognition-software-works.html.

[18] Electronic Frontier Foundation. 2019. Face Recognition. Street-Level Surveillance. Retrieved August 4, 2019 from https://www.eff.org/pages/face-recognition.

[19] Sreekar Krishna, John Black, and Sethuraman Panchanathan. 2006. Using Genetic Algorithms to Find Person-Specific Gabor Feature Detectors for Face Indexing and Recognition. In Advances in Biometrics: International Conference, ICB 2006, Hong Kong, China, January 5-7, 2006, Proceedings. Springer Science+Business, Berlin, Germany, 182-191.

[20] Amazon.com, Inc. 2019. AWS Documentation: Amazon Rekognition: Developer Guide: API Reference: Data Types. Retrieved August 4, 2019 from https://docs.aws.amazon.com/rekognition/latest/dg/API_Types.html.

[21] IAPP.org. 2019. Largest facial recognition database pulled from the web. International Association of Privacy Professionals, Inc. Retrieved from https://iapp.org/news/a/largest-facial-recognition-database-pulled-from-web/.

[22] Y. Sun, M. Zhang, Z. Sun and T. Tan. Demographic Analysis from Biometric Data: Achievements, Challenges, and New Frontiers. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 40, 2 (February 2018), 332-351. DOI:10.1109/TPAMI.2017.2669035

[23] Md. Nurul Ahad Tawhid, Emon Kumar Dey. 2018. A Gender Recognition System from Facial Image. In International Journal of Computer Applications (0975 - 8887), 180, 23 (February 2018). Retrieved August 8, 2019 from https://pdfs.semanticscholar.org/2e58/ec57d71b2b2a3e71086234dd7037559cc17e.pdf.

[24] Hida. 2015. How Common is Intersex? An Explanation of the Stats. Intersex Campaign for Equality. Retrieved f[AO] IAPP.org. 2019. Largest facial recognition database pulled from the web. International Association of Privacy Professionals, Inc. Retrieved from https://www.intersexequality.com/how-common-is-intersex-in-humans/.

[25] James, S. E., Herman, J. L., Rankin, S., Keisling, M., Mottet, L., and Anafi, M. 2016. The Report of the 2015 U.S. Transgender Survey. National Center for Transgender Equality, Washington, DC.

[26] Broadly. 2019. The Gender Spectrum Collection. Vice Media, LLC. Retrieved from https://broadlygenderphotos.vice.com/.

[27] S. Md. Mansoor Roomi, S.L. Virasundarii, S. Selvamegala, S. Jeevanandham, and D. Hariharasudhan. 2011. Race Classification Based on Facial Features. In 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics. IEEE, Hubli, Karnataka, India. DOI:10.1109/NCVPRIPG.2011.19

[28] Intelligent Multimedia Lab. Asian Face Image Database PF01. Department of Computer Science and Engineering technical report. Pohang University of Science and Technology, Korea.

[29] Adam Rose. 2010. Are Face-Detection Cameras Racist?. Time Magazine. Retrieved August 5, 2019 from http://content.time.com/time/business/article/0,8599,1954643,00.html.

[30] Patrick Grother. 2018. Demographic Effects in Face Recognition. National Institute on Standards and Technology, US Department of Commerce. International Face Performance Conference 2018 address. Retrieved August 7, 2019 from https://nigos.nist.gov/ifpc2018/presentations/17_grother_ demographics.pdf.

[31] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In Association for the Advancement of Artificial Intelligence 2019, January 2019.

[32] James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software. The Verge. Retrieved August 4, 2019 from https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai- facial-recognition-dataset.

[33] U.S. Customs and Border Protection. 2019. Biometrics. Travel. Retrieved August 4, 2019 from https://www.cbp.gov/travel/biometrics.

[34] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Conference on Artificial Intelligence, Ethics, and Society. Association for the Advancement of Artificial Intelligence, Palo Alto, CA.

[35] Joseph Menn. 2019. Microsoft turned down facial-recognition sales on human rights concerns. Reuters. Retrieved August 4, 2019 from https://www.reuters.com/article/us-microsoft-ai/microsoft-turned-down-facial- recognition-sales-on-human-rights-concerns-idUSKCN1RS2FV.

[36] Amazon.com, Inc. 2019. Customers. Rekognition. Retrieved August 4, 2019 from https://aws.amazon.com/rekognition/customers/.

[37] Ian Tucker. 2017. 'A white mask worked better': why algorithms are not color blind. The Guardian. Retrieved August 4, 2019 from https://www.theguardian.com/technology/2017/may/28/joy-buolamwini-when- algorithms-are-racist-facial-recognition-bias.

[38] Keyes, O. 2019. Counting the Countless. Real Life. Retrieved August 4, 2019 from https://reallifemag.com/counting-the-countless/.

[39] Matt Cagle and Nicole A. Ozer. (2018). Amazon Teams Up With Law Enforcement to Deploy Dangerous New Face Recognition Technology. ACLU Northern California. Retrieved August 4, 2019 from https://www.aclunc.org/blog/amazon-teams-law-enforcement-deploy- Dangerous-new-face-recognition-technology.

Chad's References

[40] Cathy O'Neil. 2016. Weapons of Math Destruction (1st. ed.). Crown, New York, NY. https://weaponsofmathdestructionbook.com/

[41] Evelyn Lamb. 2016. Review: Weapons of Math Destruction. (August 2016). Retrieved July 8, 2019 from https://blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/

[42] Wikipedia. 2019. Wikipedia: the Free Encyclopedia - Weapons of Mass Destruction. Retrieved July 8, 2019 from https://en.wikipedia.org/wiki/Weapons_of_Math_Destruction

[43] Cathy O'Neil. 2017. The era of blind faith in big data must end. Video (September 7, 2017). Retrieved July 8, 2019 from https://www.youtube.com/watch?v=_2u_eHHzRto

[44] Cathy O.Neil. 2016. Weapons of Math Destruction - Talks at Google. Video (November 2, 2017). Retrieved July 8, 2019 from https://www.youtube.com/watch?v=TQHs8SA1qpk

[45] Indiana University. 2016. Study shows allocation of police resources affects economic welfare, inequality. (July 13, 2016). Retrieved July 8, 2019 from https://phys.org/news/2016-07-allocation-police-resources-affects-economic.html

[46] Ayan Mukhopadhyay, Chao Zhang, Yevgeniy Vorobeychik, Milind Tambe, Kenneth Pence, and Paul Speer. 2016. Optimal Allocation of Police Patrol Resources Using a Continuous-Time Crime Model. Vanderbilt University, Nashville, TN & University of Southern California, Los Angeles, CA. Teamcore 2016. Retrieved July 8, 2019 from http://teamcore.usc.edu/papers/2016/Optimal-Allocation.pdf

[47] Wikipedia. 2019. Wikipedia: the Free Encyclopedia - Race and Crime in the United States - Discrimination by law enforcement and the judicial system. Retrieved July 8, 2019 from https://en.wikipedia.org/wiki/Race_and_crime_in_the_United_States#Discrimination_by_law_enforcement_and_the_judicial_system

[48] Kate Conger, Richard Fausset, and Serge F. Kovaleski. 2019. San Francisco Bans Facial Recognition Technology. (May 14, 2019). Retrieved July 8, 2019 from https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

[49] Sarah Wu. 2019. Somerville City Council passes facial recognition ban. (June 27, 2019). Retrieved July 8, 2019 from https://www.bostonglobe.com/metro/2019/06/27/somerville-city-council-passes-facial-recognition-ban/SfaqQ7mG3DGulXonBHSCYK/story.html

[50] Sarah Ravani. 2019. Oakland bans use of facial recognition technology, citing bias concerns. (July 19, 2019). Retrieved July 8, 2019 from https://www.sfchronicle.com/bayarea/article/Oakland-bans-use-of-facial-recognition-14101253.php

[51] Drew Harwell. 2018. Microsoft calls for regulation of facial recognition, saying it's took risky to leave to tech industry alone. (July 13, 2018). Retrieved July 8, 2019 from https://www.washingtonpost.com/technology/2018/07/13/microsoft-calls-regulation-facial-recognition-saying-its-too-risky-leave-tech-industry-alone/

[52] Axon. 2019. First Report of the Axon AI & Policing Technology Ethics Board. (June 2019). Retrieved August 14, 2019 from https://static1.squarespace.com/static/58a33e881b631bc60d4f8b31/t/5d13d7e1990c4f00014c0aeb/1561581540954/Axon_Ethics_Board_First_Report.pdf

[53] Os Keyes, Nikki Stevens, and Jacqueline Wernimont. 2019. The Government is Using the Most Vulnerable People to Test Facial Recognition Software. (March 17, 2019). Retrieved on July 8, 2019 from https://slate.com/technology/2019/03/facial-recognition-nist-verification-testing-data-sets-children-immigrants-consent.html

[54] Anna Merlan and Dhruv Mehrotra. 2019. Amazon's Facial Analysis Program is Building a Dystopic Future for Trans and Nonbinary People. (June 27, 2019). Retrieved on July 8, 2019 from https://jezebel.com/amazons-facial-analysis-program-is-building-a-dystopic-1835075450

[55] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots. (July 26, 2018). Retrieved on July 8, 2019 from https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

[56] Drew Harwell. 2019. Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong?. (April 30, 2019). Retrieved on July 8, 2019 from https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/?noredirect=on&utm_term=.5822334cf717

[57] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. 2018. InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity. arXiv:1712.00193v3. Retrieved on July 8, 2019 from https://arxiv.org/pdf/1712.00193.pdf

[58] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. (May 23, 2016). Retrieved July 8, 2019 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[59] Sam Corbett-Davis, Sharad Goel, and Sandra Gonzalez-Bailon. 2017. Even Imperfect Algorithms Can Improve the Criminal Justice System: A way to combat capricious and biased nature of human decisions. (December 20, 2017). Retrieved on July 8, 2019 from

https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html

[60] Megan Garia. 2016. Racist in the Machine: The Disturbing Implications of Algorithm Bias. World Policy Journal, Volume 33, Number 4, Winter 2016/2017, pp.111-117. DOI: https://doi.org/10.1215/07402775-3813015 Retrieved July 8, 2019 from https://muse.jhu.edu/article/645268/pdf

[61] Neal B. Kauder and Brian J. Ostrum. 2008. State Sentencing Guidelines: Profiles and Continuum. National Center for State Courts. Williamsburg, VA. Retrieved on July 8, 2019 from https://www.ncsc.org/~/media/Microsites/Files/CSI/State_Sentencing_Guidelines.ashx

[62] Electronic Privacy Information Center. 2019. Algorithms in the Criminal Justice System. Electronic Privacy Information Center. Washington, DC. Retrieved on July 8, 2019 from https://epic.org/algorithmic-transparency/crim-justice/

[63] Milena A, Suzanne Tamang, Jinoos Yazdany.(2018) Potential biases in Machine learning algorithms using Electronic HealthCare Data Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/#R9

[64] David Champagne,Sastry Chilukuri, Martha Imprialou and Jordan Vanlare (Dec 2018) Machine learning : Avoiding Hype , realizing potential Retrieved from https://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/machine-learning-and-therapeutics-2-0-avoiding-hype-realizing-potential

[65] Ziad Obermeyer, Senthil Mullainathan (Jan 2019) Dissecting Racial bias in an algorithm that guides health decisions for 70 million people Retrieved from http://delivery.acm.org/10.1145/3290000/3287593/p89-Obermeyer.pdf?ip=50.53.50.141&id=3287593&acc=NO%20RULES&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E307F3F2B9B0216CC%2E4D4702B0C3E38B35&__acm__=1565906153_e5db013b15e7344d41cf4bd1ea018bd1

[66] Vera Regitz - Zagrosek Sex and gender differences in health Retrieved from https://www.embopress.org/cgi/doi/10.1038/embor.2012.87

[67] NIH National Institutes of Health - How SEX and GENDER influence health and disease Retrievd from https://orwh.od.nih.gov/sites/orwh/files/docs/SexGenderInfographic11x17_508_Final_2.pdf

[68] Arpey NC, Gaglioti AH, Rosenbaum ME. July 2017 -How socioeconomic status affects patient perceptions of health care. J Prim Care Community Health Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/28606031

[69] Alice B. Popejoy , Stephanie Oct 2016 Genomics is failing on Diversity Retrieved https://www.nature.com/news/genomics-is-failing-on-diversity-1.20759

[70]Char DS, Shah NH, Magnus D.May 2018 Implementing machine learning in health care Retrieved on July 16th from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5962261

[71] Sarah Emerson, May 2019. San Francisco Bans Facial Recognition Use by Police and the Government. Retrieved on July 15th from https://www.vice.com/en_us/article/wjvxxb/san-francisco-bans-facial-recognition-use-by-police-and-the-government

[72] Caroline Haskins, A Second U.S. City has Banned Facial Recognition Retrieved on July 16th from https://www.vice.com/en_us/article/paj4ek/somerville-becomes-the-second-us-city-to-ban-facial-recognition

[73] Kate Kaye, Portland lawmakers pass privacy resolution to guide policies for facial recognition, other data use Retrieved on July 16th from https://www.geekwire.com/2019/facial-recognition-video-surveillance-highlight-new-privacy-resolution-passed-portland-lawmakers/