Exploring Fairness and Bias in Algorithms and Word Embedding

Michael A. Sosnick
Thesis Advisor: Aaron Roth, PhD
Engineering Advisor: Sampath Kannan, PhD

Senior Capstone Thesis (EAS 499)
University of Pennsylvania
School of Engineering and Applied Science
Department of Computer and Information Science
December 04, 2017

Table of Contents

Abstract

Artificial intelligence and machine learning are powerful tools that give people the ability to solve important problems. However, their application can also present new challenges that people never imagined. Creating autonomous systems with the use of artificial intelligence engenders ethical dilemmas while machine learning algorithms can amplify bias that already exists in our data and in the world. This is especially true of *word embedding*, a popular framework for machine learning in natural language processing. There have been many papers that discuss algorithmic bias and propose different ways to minimize it. However, each paper uses different definitions of bias and fairness, so the questions "What is a fair algorithm?" and "What is bias?" remain; this paper will attempt to collate and critically discuss the different definitions of bias and algorithmic fairness. Then, the paper will do a deep dive defining and quantifying bias in word embedding. While certain papers have shown that significant gender bias exists in word embedding, more work needs to be done to explore other types of biases. This paper will attempt to quantify and qualitatively describe biases that have yet be discussed. Additionally, the paper will attempt to add a new way to compare biases across different target directions.

Introduction

Artificial intelligence was the stuff of science fiction and lived in the mind of eager technologists for a long period of time. However, many recent advances in theory and computational power have made this dream more and more of a reality. For example, artificial intelligence is used all the way from self-driving cars to voice-enabled smart assistants like SIRI. Machine learning is a one of the methods to achieve artificial intelligence (Copeland, 2017). Recent advances in computational power and underlying hardware have made the implementation of machine learning algorithms possible.

Algorithms and machine learning are making decisions that affect a range of activities from advertising and driving to treating or diagnosing patients to hiring, lending, policing, and criminal sentencing (Clifton, 2016; Joseph et al., 2016; Miller, 2015; Byrnes, 2016; Rudin, 2013; Barry Jester et al., 2015). This rapid integration of powerful computers and artificial intelligence is forcing us to consider new ethical questions that we have never even imagined before. For example, self-driving cars need to make decisions that value the lives of the driver, passengers, pedestrians, and other drivers on the road. Additionally, we see that many of the algorithms responsible for these decisions may actually be biased. For example, in some cases where algorithms decide which defendants awaiting trial are too dangerous to be released, black defendants are substantially more likely than their white counterparts to be incorrectly rated as high risk (Corbett-Davies et al., 2017). Moreover, researchers at CMU ran an experiment manipulating the gender on online profiles; men, with otherwise identical profiles, were five times as likely to see ads for high paying positions (Simonite, 2015). Lastly, the paper will show that word embedding, a popular framework for machine learning in natural language processing, is biased along racial, gender, and identity lines. For example, the occupation most closely associated with Hispanic names is a *mobster*. Because algorithms are making larger and more important decisions every day, it is crucial that we are critical about the fairness of the algorithms that we trust.

Therefore, this paper will introduce artificial intelligence and machine learning and give an overview of the different types of biases that can emerge from them. Then the paper will attempt to collate and critically discuss a list of definitions for algorithmic fairness. Next, the paper will introduce word embedding, discuss how bias has been defined for it, and examine what research has revealed about current bias in the framework. Finally, the paper will use these methods to find and quantify new biases that are present in the framework.

Artificial Intelligence

Artificial intelligence has been defined as "such a program which in an arbitrary world will cope not worse than a human" (Dobrev, 2004). However, approaches to artificial intelligence can be split up into four main categories: thinking humanly, acting humanly, thinking rationally, and acting rationally. Thinking humanly requires both an understanding of how humans think and a way to implement it. On the other hand, acting humanly doesn't care about how the machine works, just that it can act like a human. The most famous formulation of this is the Turing Test, designed by Alan Turing, where a person must interact with a machine and then intuit whether it was interacting with a machine or another human being. Thinking rationally would require all

known laws of logic to be codified in a way that a machine could follow and act upon. The logicist tradition within AI tried to do this, but came across significant obstacles and current models of AI have shied away from this approach. Acting rationally, or the rational agent approach, is an approach that tries to create machines as rational agents that can perceive their environment, set and pursue goals, and act in a way to achieve the best expected outcome given any uncertainty in a system (Russel and Norvig, 2010).

Machine Learning

Machine learning is formally defined as giving computers "the ability to learn without being explicitly programmed," usually through large data sets and statistical modeling (Arthur Samuel). Machine learning algorithms can be split up into three different types of learning. In supervised learning, machines are given a series of inputs and a series of correct output labels to learn a mapping from input to output. In reinforcement learning, machines learn from a series of rewards or punishments based off their actions. In supervised learning, machines can learn patterns without being given any feedback or correct answers. Some people also point to a fourth type of learning: semi-supervised learning, where machines are given a training set of inputs and outputs and then need to use that to make guesses about new, unfamiliar inputs (Russel and Norvig, 2010).

In order to better understand where ethical issues may arise, we will break down general machine learning algorithms into four distinct phases noted by Diakopoulos and Koliska: data, model, inference, and interface. The data phase is made up of the inputs to the algorithm. The data that we use to train our algorithms can, and probably does, already have bias within it. The model phase is a "simplified or optimized reality of the world, often using data and a process that predicts, ranks, associates, or classifies." We note that algorithms that predict, rank, associate, or classify may have different definitions of bias and fairness. One of the models that we will explore in depth in this paper is word embedding for natural language processing. The inference phase consists of the actual results or recommendations of the algorithm. The last phase, interface, is what actually interacts with the outside world and the people who use the algorithm. In this phase transparency is crucial in alleviating many ethical problems.

Algorithmic Bias

Algorithmic bias is one of many ethical issues that arise with artificial intelligence. There are many different examples of algorithmic bias. We will outline the different examples, some of which are explicitly programmed while others are very hard to combat. Some of these examples are based off of scenarios outlined in *Data preprocessing techniques for classification without discrimination* and the "Catalog of Evils" in *Fairness Through Awareness*. It is important to note that many of these examples and their underlying causes overlap. Examples one through four display obvious biases within the model whereas examples five through nine display biases within the data or problems with the data itself.

1. **Explicit discrimination**.
   This is when decisions are based off of a specific, sensitive attribute. This means that people that belong to a specific group or have a sensitive attribute are explicitly denied a

positive outcome. For example, an algorithm for a bank could deny a loan based on the applicant's race or disability status.

2. **Redlining**.
   Redlining is the discrimination of people within certain neighborhoods because the majority of residents of that neighborhood have a specific attribute like being poor or being people of color (Hunt, 2005). Here, location is used as an indicator for other attributes. For example, an algorithm could discriminate against those from Harlem, New York because it has historically been a neighborhood made up of a clear majority of people of color.

3. **Sensitive Attribute as Proxy**.
   This is where one marker, like address or school, indicates a disproportionately higher likelihood of some other protected attribute, like race or sexual identity. For example, an algorithm can discriminate against someone from San Francisco, California because there is a higher percentage of LGBTQ+ folk there than in other parts of the country. We note that this doesn't mean that the majority of residents there identify as LGBTQ+, just that there is a disproportionate percentage. This more generalized version of redlining is incredibly important to note because it is harder to prevent and identify than regular redlining or explicit discrimination.

4. **Redundant Encoding**.
   In this scenario, an algorithm is blind to specific, protected attributes, but can deduce those same attributes to near accuracy from a combination of other data points. For example, Jernigan and Mistree developed an algorithm with a type of "gay-dar" which could detect a person's sexual orientation based on that person's network of Facebook friends alone (Jernigan and Mistree, 2009). Algorithm architects must be careful to ensure that their algorithms are not doing this without their knowledge.

5. **Historical Discrimination**.
   Algorithms may look at historical data when evaluating new inputs, but historical data tends to be more biased because of old policies and past cultural norms. Therefore, algorithms may learn historical discrimination if they just use historical data. For example, if an algorithm uses historical data to predict successful top employees, then the algorithm will skew overwhelmingly white, straight, and male at many companies.

6. **Encoded Existing Bias**.
   Similar to historical discrimination, the current data that an algorithm trains with can also be biased. Therefore, even if an algorithm only uses current data, it can still encode many existing biases. For example, word embedding software has been shown to encode gender bias because many people still talk and write with these biases. A large portion of this paper will be devoted to looking at this bias and other types of biases in word embedding.

7. **Multiple Data Sources**.

Training data sets are oftentimes made up of different sources or based on different populations. These different sources and populations have very different properties which may signify a positive outcome. For example, if a bank only used SAT scores to grant loans, it would disproportionately negatively affect lower income groups and people of color because SAT scores are tied very closely to a family's ability to pay for a tutor. This is important to recognize in order to both prevent bias against certain groups and incorporate crucial domain knowledge.

8. **Data Collection Feedback Loops**.
   Algorithms can make decisions for a group of people that increases (or decreases) their chances of receiving a negative (or positive) outcome in the next cycle. For example, if an algorithm gives low income people higher insurance prices, then that can cause their debt to rise and credit scores to plummet, which can lead to fewer job prospects, so that the next time they are evaluated by the algorithm they will receive even higher insurance prices (Couch, 2017). Another example is a bank that uses an algorithm that gives out loans only to white people, so that when it evaluates who has paid back their loan on time, there is historical discrimination and it will only choose white people again.

9. **Scarcity of Minority Population Data**.
   There is a scarcity of proper data about gender and minority populations (Dahal, Me and Bisogno, 2007). This effects how well a machine can learn about those groups. Because of this, many machines may be able to evaluate majority groups with higher accuracy than minority groups. Additionally, algorithms may not be able to detect important domain knowledge about minority groups that would help them.

Algorithmic Fairness

There are many papers that focus on algorithmic bias and how to alleviate it. In many of these papers, the writers offer their own definition for algorithmic fairness and use that to evaluate their work. Various papers have been written about discrimination across multiple fields (Romei and Ruggieri, 2014; Barocas and Selbst, 2016). Additionally, very recent work has been done to try to measure the discrimination in decision making and produce a unified view of performance criteria for discrimination in new algorithms (Žliobaite, 2017).

Discrimination measures can be categorized into four categories: statistical tests, absolute measures, conditional measures, and situation measures (Žliobaite, 2017). First, statistical tests are formal techniques in which statistical hypotheses are either accepted or rejected. Second, absolute measures can show the magnitude of difference between a metric for two groups. Next, conditional measures try to capture the amount of discrimination between two groups that isn't (or is) due to other characteristics. Lastly, situation measures try and quantify direct discrimination where individuals in the dataset can identify if they were discriminated against.

The discrimination measures used across these four categories range from proportion ratios to balanced residuals. These metrics can be used to expertly judge the efficacy of various

proposed solutions that are meant to combat discrimination. Additionally, they can be used to judge the bias in many different systems. For example, the impact ratio is the ratio of positive outcomes for the protected class over the general group; $r = p(y^+|s^1)/p(y^+|s^0)$, where a ratio of $r = 1$ is fair. In US courts this is used to quantify discrimination, where a ratio of less than 80% is considered discriminatory. Additionally, we see that some of these metrics are used in the field of computer science to measure discrimination or are used in specific definitions of fairness. For example, Hajian et al. use extended lift (*elift*) ratios to define redlining rules. However, some of the other measures like mutual information (MI), which measures mutual dependence between variables, are used more to quantify discrimination rather than claiming what is or what isn't fair. Because these measures focus on measuring discrimination rather than proposing definitions of fairness, it is still important to create a standard list of fairness definitions.

The definitions of fairness that have been proposed in academic works combating algorithmic bias set a standard to meet across various metrics. A more unified view of the definitions of fairness will encourage authors to think and argue critically for why a particular definition of fairness is appropriate. This is especially important because one definition of fairness should not be taken as a given. Additionally, authors will be better able to concede which definitions of fairness or discrimination metrics their proposals do and do not address. This will help readers understand the strengths and weaknesses of these academic papers and algorithms and lend credibility to the field. Therefore, we start to formulate a list of major definitions of fairness based on the definitions set forth in important work that is being done in the field and critically evaluate them.

1. **Statistical Parity**.

   Statistical parity is defined as the "property that the demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole" (Dwork et al., 2011). For example, if a percentage *X* of applicants are supposed to be approved for a loan, then an *X* percentage of each protected class should also be approved for a loan.

   This tries to achieve group fairness or fairness for a group as a whole, rather than individual fairness or for individual people. Various statistical tests can be used to ensure that this is the case. This definition is even legally mandated in certain cases (Kleinberg et al., 2016). However, Dwork et al. shows that statistical parity can actually lead to some unfair situations. For example, if we achieve statistical parity, we may be trying to equalize between groups that are unequal in talent or positive characteristics. Take a recruiter who is looking for a software engineer and tries to achieve statistical parity between majors; obviously, there could be more qualified candidates in the computer science major than in a non-STEM major like English. Additionally, statistical parity can also hurt disadvantaged groups rather than help them. If statistical parity is pushed so that unqualified "token" candidates are chosen just for show, then their possible poor performance can be used as justification for bias afterwards.

   Therefore, we see another version of this definition called *conditional statistical parity*. In this altered definition, we control for a limited set of "legitimate" risk factors. For example, in the case where an algorithm decides if a defendant awaiting trial is high risk, *conditional statistical parity* would mean that within a group of people with similar

number and types of prior convictions, black and white defendants are labeled high risk at the same rate (Corbett-Davies et al., 2017).

2. **Calibration**.

   An algorithm is considered well-calibrated if when it "identifies a set of people as having a probability *X* of constituting positive instances, then approximately an *X* fraction of this set should indeed be positive instances" (Kleinberg et al., 2016). We define positive instances to mean actually having the characteristic that the algorithm is trying to identify. For example, if an algorithm identifies members of group *Z* to have an *X* percentage chance of paying back their loans, then an *X* percentage of people in group *Z* should pay back their loans. Calibration could exist within groups or across groups; calibration within groups would mean that each group *Z* would have its own percentage X_Z that it would need to meet, whereas calibration across groups would require every group to have the same percentage *X* that they would need to meet.

   This definition of fairness, especially calibration within groups, is usually paired with other definitions as well. Both calibration within groups and calibration across groups tries to achieve group fairness, but calibration within groups does so with a limited scope as it does not prevent biases like historical discrimination, the encoding of existing biases, or data collection feedback loops. For example, an algorithm may correctly predict that less women will stay with a company for an extended period of time and hire based on that information; however, if the company continues to hire less women, then it may also be seen as a less female-friendly environment to work, which will encourage them to leave earlier than their male counterparts.

3. **Balance for the Negative Class**.

   This requires that the people who are classified negatively in different groups should have the same average score (Kleinberg et al., 2016). For example, let's say an algorithm must consider applicants for a loan, then all white people who default on their loan should be given the same average score as all black people who default on their loan.

4. **Balance for the Positive Class**.

   This requires that the people who are classified positively in different groups should have the same average score (Kleinberg et al., 2016). This is the flip side of the previous definition of Balance for the Negative Class. For example, let's say an algorithm must consider applicants for a loan, then all white people who pay back their loan on time should be given the same average score as all black people who pay back their loan on time.

5. **Predictive Equality**

   Predictive Equality requires similar accuracy across groups as defined by the rate of false positives (Corbett-Davies et al., 2017). However, this could be expanded to use the false negative rate as well. A false negative occurs if the algorithm classified someone negatively, but the person actually had the positive characteristic. Additionally, a false positive occurs if the algorithm classified someone positively, but the person actually had the negative characteristic. This measure clearly speaks to group fairness; however, it also ensures that individuals would have had the same chance of being falsely identified

in any group. Similarly, in statistics, a Type II error is when the algorithm fails to reject a null hypothesis that is actually false for a given person and a Type I error is when the algorithm rejects a null hypothesis that is actually true for a given person. It is important to note, however, that false negatives and false positives are not called Type I and Type II errors when they are caused by bias (Banerjee et al., 2017).

At the same time, these types of errors are hard to detect and cannot usually be quantified (Banerjee et al., 2017). For example, after a bank rejects an applicant for a loan, it has no way of knowing if that applicant *would* have been able to pay back the loan unless it ended up getting a loan from a different bank and there is access to that information. Therefore, this definition of fairness is hard to implement.

6. **Individual Fairness**.

This requires that "any two individuals who are similar with respect to a particular task should be classified similarly" (Dwork et al., 2011). This is based on the notion of *Fair Equality of Opportunity* set forth by John Rawls, which states that social positions, or the like, be meritocratically allocated such that any two individuals who are similar with respect to some determining characteristic like talent or ambition will have the same prospects of success (Rawls, 1999; *Stanford Encyclopedia of Philosophy*, 2017). There are different ways to try and achieve this definition of fairness.

In order to quantify this, Dwork et al. assumes some distance metric that captures similarity with respect to a particular task between different people and proposes a Lipschitz condition on that metric. A Lipschitz condition requires that any two individuals *x, y* which are at a "distance $d(x, y) \in [0, 1]$ map to distributions M(x) and M(y), respectively, such that the statistical distance between M(x) and M(y) is at most d(x, y)" (Dwork et al., 2011). This means that any two individuals *x, y* who are within some distance $d(x, y) \in [0, 1]$ from each other should have indistinguishable outcomes.

In our context, *Fair Equality of Opportunity* could also be defined as an algorithm that never preferentially chooses individual *X* over individual *Y* if individual *X* is not as qualified as individual *Y*. More formally, we say that for every choice an algorithm makes between individuals *x, y*, the probability that it chooses *x* should be greater than the probability it chooses *y* only if the quality or expected outcome of individual *x* is greater than that of individual *y* (Joseph et al., 2016).

These are only some of the major descriptions of fairness in algorithms. At first glance, it would seem appropriate to try and achieve all of these in every algorithm. Unfortunately, we see that oftentimes this is not possible. Kleinberg et al. show that in most cases it is impossible to achieve calibration within groups, balance for the positive class, and balance for the negative class at the same time. The only time that you can achieve all three simultaneously is when the algorithm has perfect prediction or when the different groups have equal base rates (the same fraction of members in the positive class). Additionally, Dwork et al. show that their definition of individual fairness using a Lipschitz condition can imply statistical parity if and only if the Earthmover distance (based on the similarity metric) between the two groups is meaningfully small. Additionally, we also know that some of these definitions of fairness by themselves can be considered very unfair under other definitions.

Most of the definitions of fairness described above apply to classification algorithms where inputs or individuals are classified into groups that will receive different benefits. However, machine learning can also be used for other types of algorithms, which can also exhibit bias. Notably, machine learning used in natural language processing can also display significant gender bias (Bolukbasi et al., 2016; Zhao et al., 2017; Schmidt, 2015). The definitions of fairness described above are not easily applicable to the case of word embedding as there are no individuals or a priori classification problem (Bolukbasi et al., 2016). Therefore, defining fairness and bias in machine learning algorithms for language is incredibly important.

Word Embedding

Word embedding is a popular framework for language modeling where words or phrases in text data are mapped to vectors of real numbers, which are then used in many machine learning algorithms. Formally, each word or phrase is represented as a *d*-dimensional word vector. Word embeddings can act as a sort of dictionary for computer programs that need to "understand" what each word means (Bolukbasi et al., 2016). We also see that many linguistic patterns can actually be represented as linear translations, i.e. vector differences show relationships between words and vector arithmetic can show analogies between sets of words (Mikolov et al., 2013). For example, we can show clear relationships through subtraction such as:

$$\overrightarrow{Spain} - \overrightarrow{Madrid} \approx \overrightarrow{France} - \overrightarrow{Paris}$$

We can also show other types of relationships as well. We can easily manipulate the arithmetic to show word analogies as well. For example:

$$\overrightarrow{man} - \overrightarrow{woman} + \overrightarrow{queen} \approx \overrightarrow{king}$$

We see that word embedding can capture relational similarity for a range of word groups like superlatives, past participles, sports, politics and occupations (Heuer, 2105).

Additionally, the similarity of two words is found using the inner product of their vectors. We define the similarity as the cosine of the angle between two word vectors such that given arbitrary word vectors *u*, *v:*

$$cos(u, v) = \frac{u \cdot v}{\| u \| \| v \|}$$

Because word vectors are already normalized, $\cos(\vec{u}, \vec{v}) = \vec{u} \cdot \vec{v}$ (Bolukbasi et al., 2016). We also note that word vectors can be used to compare large text sources as well (Figure 1). From all of these examples we see that we can learn many interesting things from word vector comparison.

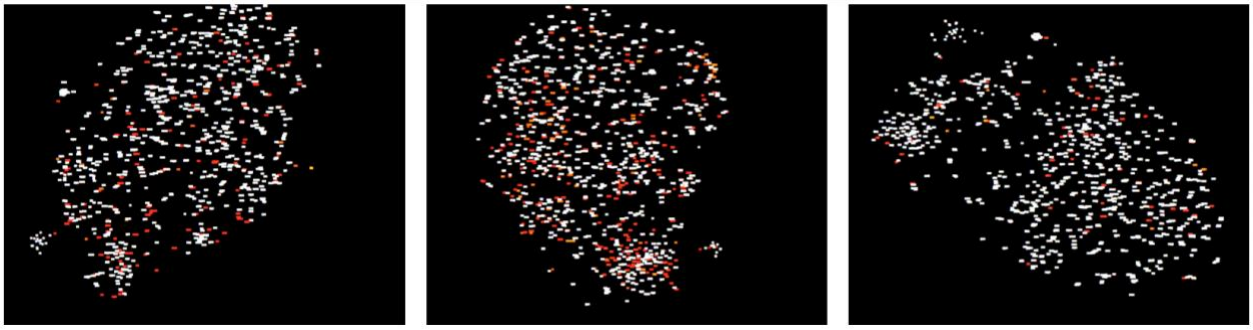

Figure 1 Global cluster of the Wikipedia articles on the United States (left), Game of Thrones (middle), and Word War II (right) The word embeddings are projected down to 2D using t-SNE. (Heuer, 2015)

Metrics for Bias in Word Embedding

Clearly, word embedding is powerful; however, its ability to capture relationships between words also allows it to capture bias as well. We know that significant research has been done in the field and has shown that word embedding software like word2vec and GloVe retain significant gender bias (Bolukbasi et al., 2016; Zhao et al., 2017; Schmidt, 2015). However, similar to bias in classification algorithms, we must also define metrics for bias in the field of natural language processing (NLP) in order to help further work to debias NLP algorithms.

1. **Direct Bias**.

   The main description of direct bias that we use is taken from Bolukbasi et al. in their work *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* and referenced in Chakraborty et al. in their work *Reducing gender bias in word embeddings*. This metric is used to identify the direct bias in a word embedding application. It is defined as follows.

$$DirectBias = \frac{1}{|N|}\sum_{w \in N} |\cos(w, d)|^c$$

   We define a direction *d* as the relationship between two sets of words. For example, in order to describe a *gender direction*, we combine the directions of multiple gender word pairs like $\overrightarrow{man} - \overrightarrow{woman}$ and $\overrightarrow{he} - \overrightarrow{she}$. This allows us to better capture a bias around gender that doesn't include the noise associated with particular words. For example, the word *man* can be used to describe a male human being, but it can also be used as a verb in "*man the station*" or an exclamation in "*oh man!*". Additionally, we define a set of words that should be neutral along the direction that we want to test for bias as *N*. Lastly, we define a measure *c* to denote how strict we would like the test to be. For example, if we set *c = 0*, then the metric will return *1* if any bias is found and *0* if there isn't any bias found. We can also set *c = 1*, which would result in a more gradual bias, where we allow for words to be weighted by frequency.

2. **Indirect Bias**.

We also introduce the idea of *indirect bias*, which creates relationships between words that are not specific attributes, but clearly arise because of them. For example, Bolukbasi et al. note that *receptionist* is much more similar to *softball* than it is to *football*. This could be because *receptionist* and *softball* are both associated with *female*. We note that there are many relationships between words which may not exhibit bias just because both words are also close to the same demographic or attribute marker. Bolukbasi et al. show that this is the case with *mathematician* and *geometry*, which both show a strong *male* association, but are closely related for other reasons.

This description of indirect bias is taken from Bolukbasi et al. in their work *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* and referenced in Chakraborty et al. in their work *Reducing gender bias in word embeddings*. We define the metric as the attribute component to the similarity between two vectors *w*, *v* and as follows:

$$\beta(w, v) = \frac{\left(w \cdot v - \frac{w_\perp \cdot v_\perp}{||w_\perp||_2 \cdot ||v_\perp||_2} \right)}{w \cdot v}$$

We say that given a word vector *w*, then $w = w_d + w_\perp$ where w_d is the "contribution" from an attribute direction *d* and $w_d = (w \cdot d)d$. Therefore, this metric is actually the inner product of the renormalized vectors after removing any attribute direction component. We note that $\beta(w, w) = 0$ because a word's similarity to itself is not dependent on its attribute direction component. We also note that $\beta(w, v) = 0$ if $w_d = v_d = 0$ because neither word has an attribute direction component and so the two words cannot have any similarity due to the attribute. Lastly, we note that $\beta(w, v) = 1$ if $w_\perp = v_\perp = 0$ because the two words have no meaning outside of the attribute direction *d* (Bolukbasi et al., 2016).

3. **Word Embedding Association Test (WEAT)**.

This test is from Caliskan et al. in their work *Semantics derived automatically from language corpora contain human-like biases*. It is built off of the Implicit Associations Test (IAT) and tests for a form of indirect or implicit bias. It is in the form of a test statistic where the null hypothesis is that two sets of non-attribute specific words are equally similar to corresponding attribute word groups. For example, two sets of non-gender specific occupation words (e.g. the set {computer programmer, construction worker, …} and the set {nurse, librarian, …}) should be equally similar to each gender set (e.g. the set {man, he, …} and the set {woman, she, …}). The test statistic is defined as follows:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

and

$$s(w, A, B) = mean_{a \in A} \cos(\vec{w}, \vec{a}) - mean_{b \in B} \cos(\vec{w}, \vec{b})$$

We define *X* and *Y* as the two sets of target non-attribute specific words and *A* and *B* as the sets of corresponding attribute words. We also define the size of the of the bias effect as:

$$\frac{mean_{x \in X} s(x, A, B) - \ mean_{y \in Y} s(y, A, B)}{std - dev_{w \in X \cup Y} s(w, A, B)}$$

4. **Secondary Bias**.

Another description of bias is taken from Zhao et al. in their work *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. They use the metric below to rate the bias in multi-label object classification and visual semantic role labeling. Although this isn't originally used directly to capture the bias in word embedding, we argue that it can still be useful. Because word embedding software may be used with or in labeling applications, this metric can still relay valuable information about bias. We name this metric *secondary bias* in regards to its relationship to word embedding. We define the metric as follows:

$$b(o, g) = \frac{c(o, g)}{\sum_{g' \in G} c(o, g')}$$

We define *o* as some subset of the output of a labeling system and g as some subset output variable which corresponds to a sensitive demographic attribute. For example, *o* may be the activity or verb that the labeling system is trying to identify and g is the gender of the agent that the labeling system is trying to identify such that $g \in G = \{man, woman\}$. We define *c(o,g)* as the number of *o* and g in the corpus. If we apply this to the case of verbs and the gender direction, then the metric would be as follows:

$$b(verb, man) = \frac{c(verb, man)}{c(verb, man) + \ c(verb, woman)}$$

where $b(verb, man)$ would be the gender bias towards *man* for each verb. We then say that if $b(o, g) > \frac{1}{||G||}$, then *o* and g are positively correlated and there may be bias. (Zhao et al., 2017).

Gender Bias

Multiple recent papers have been published that help reveal and quantify the bias in word embedding and most of them reveal significant gender bias (Bolukbasi et al., 2016; Zhao et al., 2017; Schmidt, 2015). Gender bias in word embedding reflects many of the biases in our current society; however, word embedding can also have the ability to amplify these biases. For example, the word *doctor* is closer to *he* than it is to *she* along the gender direction. So, when searching for a doctor in a particular field, given everything else being equal, male doctors will appear higher than their female counterparts. This would make it harder for women in medicine and perpetuate the very bias that it reflects. Even more, it is easy to hypothesize situations where

a female doctor who is better for the job and can make the difference between life and death is not properly found because her male counterpoints were placed higher than her in search results. This is only one hypothetical situation, but we could imagine many more where the gender bias negatively affects both men and women and perpetuates gender stereotypes. We also see that this gender bias is even magnified to a greater degree in languages that have grammatical gender associations like Spanish and German (McCurdy et al., 2017).

Bolukbasi et al. use the word2vec software that is trained on a corpus of Google News texts consisting of 3 million English words, which they call w2vNEWS. They show gender bias is exhibited strongly in the set of occupation words and amongst analogies. Qualitatively, we can see how these biases appear in the figures above. We see from Figure 2 that the occupations as projected on the *she-he* gender direction contain significant bias and fall within common stereotypes of gendered occupations. Additionally, we see that many of the analogies generated with word embedding reveal significant bias. The analogies demonstrate noteworthy implicit bias that reflect how gender bias in word embedding is far reaching and affects occupation words, action words, item words, activity words, and adjectives. In figure 4, we see how indirect bias connects words like softball and receptionist mostly because of gender. Quantitatively, Bolukbasi et al. found that 19% of the top 150 analogies created using w2vNEWS were judged to show gender bias by a majority of crowd workers. Additionally, they found that the set of occupation words have a DirectBias score of 0.08, which shows significant bias. Quantifying implicit bias across a set is difficult, but some scores can be found in Figure 4.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 2 The most extreme occupations as projected on the she-he gender direction. (Bolukbasi et al,. 2016)

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 3 Analogies generated along the she-he direction using w2vNEWS. For example, the first analogy is she:sewing to he:carpentry. The analogies in the top list were rated as stereotypes by at least 5 out of 10 crowd workers that were polled. The analogies in the bottom list were rated as appropriate. (Bolukbasi et al., 2016)

softball **extreme**	**gender portion**	**after debiasing**
1. pitcher	-1%	1. pitcher
2. bookkeeper	20%	2. infielder
3. receptionist	67%	3. major leaguer
4. registered nurse	29%	4. bookkeeper
5. waitress	35%	5. investigator
football **extreme**	**gender portion**	**after debiasing**
1. footballer	2%	1. footballer
2. businessman	31%	2. cleric
3. pundit	10%	3. vice chancellor
4. maestro	42%	4. lecturer
5. cleric	2%	5. midfielder

Figure 4 An example of indirect bias. These are the 5 most extreme occupations along the softball - football direction. The gender component or indirect bias β of each word due to the gender direction is shown. (Bolukbasi et al., 2016).

Chakraborty et al. show similar results, but focus on the GloVe word embedding software. They find a DirectBias score of .114 over their corpus of words. Additionally, they provide an interesting visual of words projected onto the *she-he* gender direction. We see in Figure 5 that words like *lion*, *dictator*, and *great* are very biased towards *he* and the words *fetus*, *sperm*, *nude*, and *lovely* are very biased towards *she*. Additionally, they find similar results to Bolukbasi et al. in indirect bias as well. They find that the words *receptionist*, *podiatrist*, *caregiver*, and *publisher* have gender portions of 64%, 42%, 26%, and 24% respectively. It is important to note that this implies that the biases present in word embedding are not unique to the word2vec algorithm, but are present in word embedding in general as word2vec and GloVe are two of the most popular, trusted word embedding algorithms.

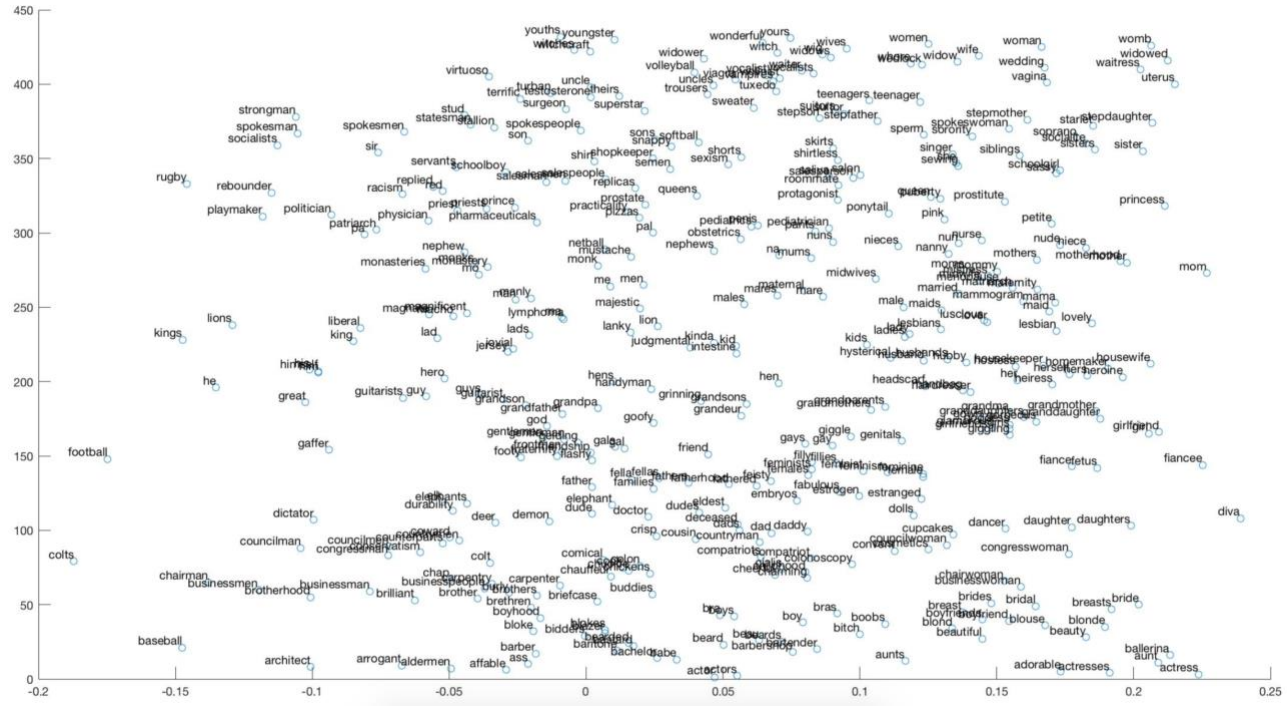

Figure 5 Select words projected on to the she-he direction. Words to the left are extreme he words and words to the right are extreme she words. (Chakraborty et al., 2016).

Other Biases

In a very recent study, Caliskan et al. show that applying machine learning to human language results in many historic biases. Based on the IAT, they created the WEAT discussed above to compare different targets along a direction. For example, they run tests on a *flower-insect* direction using a set of pleasant words that include *caress*, *freedom* and *health* and a set of unpleasant words like *abuse*, *crash*, and *filth* (see Appendix 1 for a full list of pleasant and unpleasant words used). These tests used 25 names of different flowers and 25 names of different insects to generate the direction. Additionally, Caliskan et al. did WEAT studies comparing target directions where common IAT tests have shown significant bias in the past. We see from Figure 6 that word embedding contains many of the same biases that we also find in IAT literature, with the study being able to replicate every stereotype that it tested. In fact, the effect of many of the biases in the IAT literature are not only matched, but actually amplified in word embedding (see rows 1, 3, 6, 7, and 9 in Figure 6).

Target words	**Attrib. words**	**Original Finding**				**Our Finding**			
		Ref	**N**	**d**	**p**	$\mathbf{N_T}$	$\mathbf{N_A}$	**d**	**p**
Flowers vs insects	Pleasant vs unpleasant	(*5*)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(*5*)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(*5*)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*5*)	(*7*)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*9*)	(*7*)	Not applicable			16×2	8 × 2	1.28	10^{-3}
Male vs female names	Career vs family	(*9*)	39*k*	0.72	$< 10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs arts	Male vs female terms	(*9*)	28*k*	0.82	$< 10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs arts	Male vs female terms	(*10*)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(*23*)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(*9*)	43*k*	1.42	$< 10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Figure 6 Compare results from 8 well-known IAT findings (rows 1-3 and 6-10) to WEAT findings using word2vec. Rows 4 and 5 show bias in hiring. The WEAT compare two sets of words from target concepts like flowers and insects with two sets of attribute words like pleasant and unpleasant. Further, N *is the number of subjects in the IAT,* N_T *is the number of target words in WEAT studies,* N_A *is the number of attribute words in WEAT studies,* d *is the effect size in each study, and* p *stands for the p-values associated with each* d *score. (Caliskan et al,. 2017).*

Importantly, this study starts to show, quantify, and prove other types of biases that are present in word embedding besides for just gender bias. It gives a high level, essential breath on bias in various areas and implies that many – if not all – historic biases are also present in word embedding. The study also importantly starts the work of investigating racial bias against African Americans. We see that significant bias occurs across the European American to African American direction. However, more work needs to be done to show how these biases play out, both quantitatively and qualitatively.

New Findings using Occupations

Bolukbasi et al. and others first showed that significant gender bias exists and should be corrected. Then, Caliskan et al. show that these biases occur across a host of different directions. Therefore, we argue that continued research needs to be done to quantify and qualitatively describe these other biases. We attempt to start that process and show several interesting results, critically evaluate them, and describe some key takeaways.

We build on Bolukbasi et al. and show bias in the occupation word set with new directions. In order to find the most extreme occupations of one target set, we create a direction by adding all of the vectors for words in that target set and then subtract all of the vectors for the words in the opposite target set. We then find the top words that are closest to that new vector direction. We do this using *gensim*, which is a popular word2vec Python library. We use the model trained on the same corpus as Bolukbasi et al., which is the Google News set. We uncover, quantify, and qualitatively describe the biases present using the occupation word set. We first generate directions using group titles like *whites*, *blacks*, *straights*, and *gays*. Then we generate target directions based off of the most common first and last names in those groups. We show biases that result from these directions both qualitatively using lists of the most extreme occupations as projected on to those directions and quantify these biases with the DirectBias score.

	whites	**blacks**	**whites**	**Latinos**	**whites**	**minorities**
1	guitarist	alderman	headmaster	congressman	butler	parliamentarian
2	waiter	councilman	butler	paralegal	dad	advocate
3	skipper	congressman	civil servant	councilman	footballer	deputy
4	monk	attorney	barrister	educator	socialite	chancellor
5	cabbie	historian	skipper	senator	crooner	legislator
6	drummer	educator	inventor	ballplayer	teenager	lawyer
7	adventurer	advocate	butcher	pollster	sailor	employee
8	maestro	administrator	warden	attorney	adventurer	undersecretary
9	chef	paralegal	industrialist	undersecretary	maestro	envoy
10	vocalist	comic	colonel	alderman	ballerina	lawmaker

Figure 7 The top ten occupations along three different directions: whites-blacks, whites-Latinos, and whites-minorities. The occupation set was the 319 occupations used by Bolukbasi et al. The target sets were, respectively, ["whites", "white", "Caucasian"] to ["blacks", "black", "African American"], ["whites", "white", "Caucasian"] to ["Latinos", "Latino", "Hispanic", "Latina"], and ["whites"] to ["minorities"]. A list of the 319 occupations can be found in Appendix 3.

There are some key differences between establishing a clean racial direction, for example *whites-blacks*, and establishing a gender direction. Gender can be indicated in many different ways including pronouns, nouns, names, orthography, and gender-specific words at large. For example, we have words like *he*, *man*, *father*, *businesswoman*, *lesbian* and *prince* (see Appendix 2 for a full list of 218 gender-specific words). However, there are not nearly as many words that indicate racial difference. Additionally, gender directions are much easier to establish because gender is indicated in some form – through pronouns, names, explicit gender markers, etc. – in almost every sentence. However, race is indicated nearly as often. Because there are so many more occurrences of gender markers than that of racial markers, there should probably also be a much stronger gender direction than a racial direction.

Originally, we created the racial direction using target sets of group titles. For example, we initially created the *whites-blacks* direction using the sets {*whites, white, Caucasian*} and {*blacks, black, African American*}. One would originally think that the *whites* set would be associated with higher paying, more prestigious jobs and the *minorities*, *blacks*, or *Latinos* sets would be associated with lower paying, less prestigious jobs. Bolukbasi et al. support this and find that the most extreme *whites* occupations are *parliamentarian, advocate, deputy, chancellor, legislator*, and *lawyer*, whereas the most extreme *minorities* occupations are *butler, footballer, socialite,* and *crooner*.

Interestingly, we see almost the exact opposite bias in Figure 7 then what we would expect and what is noted in Bolukbasi et al. The most obvious reason for this is first-order bias, which refers to how natural language processing algorithms learn connections between words based off of their direct juxtaposition. For example, one common practice in NLP is to process words in bigrams or n-grams where the *n* terms before a word are stored, so that the algorithm could use their counts to calculate the probability of a word coming after the *n* previous words. We would expect that words that are often used together in the same sentence or description would be highly correlated in a word embedding model. However, this is not always the case. For example, because of first-order bias, we would initially think that a word like *nurse* would be much closer related to *male* than *female* because the term *male nurse* is several times more frequent than *female nurse* (Bolukbasi et al., 2016). However, we see that word embedding is sometimes capable of overcoming first-order bias as *nurse* is still much closer to *female* than it is to *male*. This is probably because there are many other signifiers of gender that would imply that the majority of nurses are actually female, i.e. while *male nurse* is much more frequent than *female nurse,* female gender pronouns and names are probably correlated with the word *nurse* more often than their male parallels. On the other hand, in the case of *whites, minorities, blacks,* and *Latinos,* we see that word embedding was probably not able to overcome first order bias where terms like *black congressman* are probably much more frequent than *white congressman*.

We try the same method as described above with other groups as well. We see in Figure 8, that first order bias still seems present along a *straight-gay* direction, but does not change the extreme occupations along the *Christian-Jew* or *Christian-Muslim* directions from what we would expect given historic bias. Interestingly, the top ten most extreme *gay* occupations are almost all religious in nature. This makes sense as religion and homosexuality are often written about together, even though they are usually at odds.

	straight	gay	Christian	Jew	Christian	Muslim
1	maestro	pastor	evangelist	rabbi	parishioner	cleric
2	tycoon	rabbi	pastor	mobster	missionary	shopkeeper
3	cinematographer	missionary	missionary	nanny	pastor	cab driver
4	N/A	priest	solicitor general	violinist	organist	cabbie
5	N/A	bishop	administrator	waiter	evangelist	lawmaker
6	N/A	preacher	skipper	cab driver	priest	jurist
7	N/A	evangelist	vice chancellor	cellist	bishop	taxi driver
8	N/A	archbishop	manager	composer	soloist	butcher
9	N/A	chaplain	preacher	narrator	pianist	diplomat
10	N/A	comedian	principal	pianist	counselor	lyricist

Figure 8 The top ten occupations along three different directions: straight-gay, Christian-Jew, and Christian-Muslim. The occupation set was the 319 occupations used by Bolukbasi et al. The target sets were, respectively, ["straight", "straights", "heterosexual", "hetero"]- ["gay", "gays", "queer", "homosexual", "homo"], ["Christianity", "Christian", "Christians"]- ["Judaism", "Jewish", "Jews", "Jew"], ["Christianity", "Christian", "Christians"]- ["Islam", "Muslim", "Muslims"]. N/A means that no more occupations from our set were found amongst the top 100,000 closest word vectors to the direction.

First order bias seems to change what we expect substantially and at times may seem even in favor of the minority group, like for occupations and the *whites-minorities* direction. However, we see from Figure 9 that there is still significant direct bias in each target direction. Figures 7-9 imply that while we may not get the exact biases that we were expecting, the words used in each target direction are still encoded with bias. So, even if the word *minorities* does not necessarily capture the way that minorities have historically experienced bias, it is still biased in ways that can be hurtful.

In order to overcome first order bias, we repeat some of these tests using common names associated with each group as the set of target words, which has precedence from Caliskan et al. This method will exclude noise from the second meanings of homonyms in the target set. For example, the word *white* can refer to a person's race or refer to the color. We note from Figure 10 significant qualitative bias between historically black names like *Jamal* and *Lakisha* and historically (European-American) white names like *Brad* and *Emily* (see Appendix 4 for a

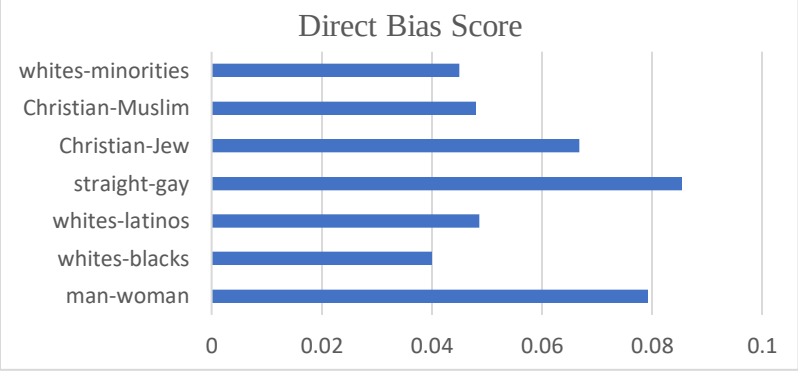

Figure 9 DirectBias score of each target direction against the occupation set. Values can be found in Appendix 14.

	white male & female first names	black male & female first names	white male first names	black male first names	white female first names	black female first names
1	architect	artiste	adjunct professor	shopkeeper	architect	artiste
2	consultant	shopkeeper	consultant	taxi driver	historian	cab driver
3	director	taxi driver	director	artiste	consultant	taxi driver
4	historian	cab driver	adventurer	cleric	planner	shopkeeper
5	adjunct professor	preacher	freelance writer	cab driver	director	preacher
6	inventor	boxer	architect	preacher	naturalist	singer
7	naturalist	laborer	inventor	gangster	screenwriter	boxer
8	planner	barber	investment banker	laborer	physicist	maid
9	adventurer	singer	author	barber	inventor	laborer
10	programmer	cleric	manager	boxer	archaeologist	barber

Figure 10 The top ten occupations along three different directions: white-black (male & female first names), white-black (male first names), and white-black (female first names). The occupation set was the 319 occupations used by Bolukbasi et al.

full list of names). This further substantiates the claim made in Caliskan et al. that having African-American names can be very damaging in the interview and hiring process. It is easy to imagine that hiring algorithms that use word embedding will encode this bias into their results. Additionally, it is interesting to note that we do not see substantial differences between the *white-black (male & female first names)* direction, the *white-black (male first names)* and *white-black (female first names)* directions, i.e. we do not see that the most extreme *black female first names* are significantly more biased than those for *black male & female first names*. These results may imply that there are no significant negative effects from intersectionality here, where intersectionality is defined as the "study of intersecting social categories – such as race, gender, and social class – with which an individual identifies (Guittar et al., 2015). Alternatively, this may imply that the racial direction here is strong and paralleled in both male and female names.

We also look at the same effects for Hispanic names. We do this using common Hispanic names like *Mateo, Santiago, Sofia*, and *Isabella* (see Appendix 4 for a full list of names). We see from Figure 11 that many historic biases in the *white-Hispanic* direction are reflected in word embedding. We note once again that there are similar extreme occupations across male & female, only male, and only female directions. These biases can be very harmful. For example, if someone searches the word *mobster* on Google, then Hispanic names may come up more often because of word embedding, which would then further reinforce the bias that mobsters are Hispanic. We see similar results to the *whites-Hispanic* direction in Appendix 6, where common Arab names like "Mohammed" and "Omar" are used to create a *whites-Arab* direction.

	white male & female first names	Hispanic male & female first names	white male first names	Hispanic male first names	white female first names	Hispanic female first names
1	freelance writer	mobster	surveyor	house-keeper	freelance writer	mobster
2	surveyor	house-keeper	barrister	janitor	author	artiste
3	author	saint	solicitor	nun	sportswriter	saint
4	vice chancellor	waiter	professor emeritus	taxi driver	curator	waiter
5	professor emeritus	taxi driver	author	saint	vice chancellor	boxer
6	barrister	janitor	stock-broker	mobster	guidance counselor	ballerina
7	investment banker	ballerina	vice chancellor	assassin	writer	laborer
8	curator	laborer	freelance writer	waiter	librarian	taxi driver
9	sportswriter	artiste	historian	priest	professor emeritus	house-keeper
10	editor	priest	investment banker	ballerina	columnist	bodyguard

Figure 11 The top ten occupations along three different directions: white-Hispanic (male & female first names), white-Hispanic (male first names), and white-Hispanic (female first names). The occupation set was the 319 occupations used by Bolukbasi et al.

	White last names	Hispanic last names	White last names	Asian last names
1	surveyor	major leaguer	preacher	monk
2	vice chancellor	undersecretary	chaplain	researcher
3	solicitor	archbishop	sportsman	doctoral student
4	architect	infielder	dad	artiste
5	philanthropist	ballplayer	firebrand	housewife
6	headmaster	priest	trooper	violinist
7	mathematician	housekeeper	mediator	assistant professor
8	barrister	mobster	ballplayer	cellist
9	inventor	congressman	coach	taxi driver
10	adventurer	nun	handyman	professor

Figure 12 The top ten occupations along two different directions: white last names - Hispanic last names and white last names-Asian last names. The occupation set was the 319 occupations used by Bolukbasi et al.

We also look at sets of last names to see if similar biases exist. In the list of extreme Hispanic occupations, we see many sports positions, which we did not see in the list of extreme

occupations for Hispanic first names. This implies some combination of historic bias and the fact that athletes are referred to by their last name. Additionally, we see historic bias in the list of extreme Asian occupations. The multiple occupations in academia like *researcher* and *assistant professor* in the list of extreme Asian occupations suggest a correlation between academia in general and Asian names.

Additionally, we rate the DirectBias score of these directions in order to quantify these results. We show the results for directions that were generated using lists of names. We note that these DirectBias scores are lower than the DirectBias scores for the gender direction and some of the target directions generated using only group title names like the *straights-gays* direction. However, these DirectBias scores are still significant and some are actually higher than their parallel DirectBias scores calculated using group title names. For example, all of the DirectBias scores of target directions that are generated using historically African American names are higher than the DirectBias score of the *whites-blacks* target direction generated using {"whites", "white", "Caucasian"} and {"blacks", "black", "African American"}.

We also note that many of the directions that we created using group names cannot easily be recreated using personal first and last names. For example, there aren't stereotypical first or last names for gay individuals. Additionally, we note that when creating the target directions, we do not mean to conflate different, disparate communities. For example, we do not try to conflate the Hispanic and Latino communities into one entity or all of the various Asian communities into one entity. We set up the directions that we did in order to try and create the strongest directions that we could.

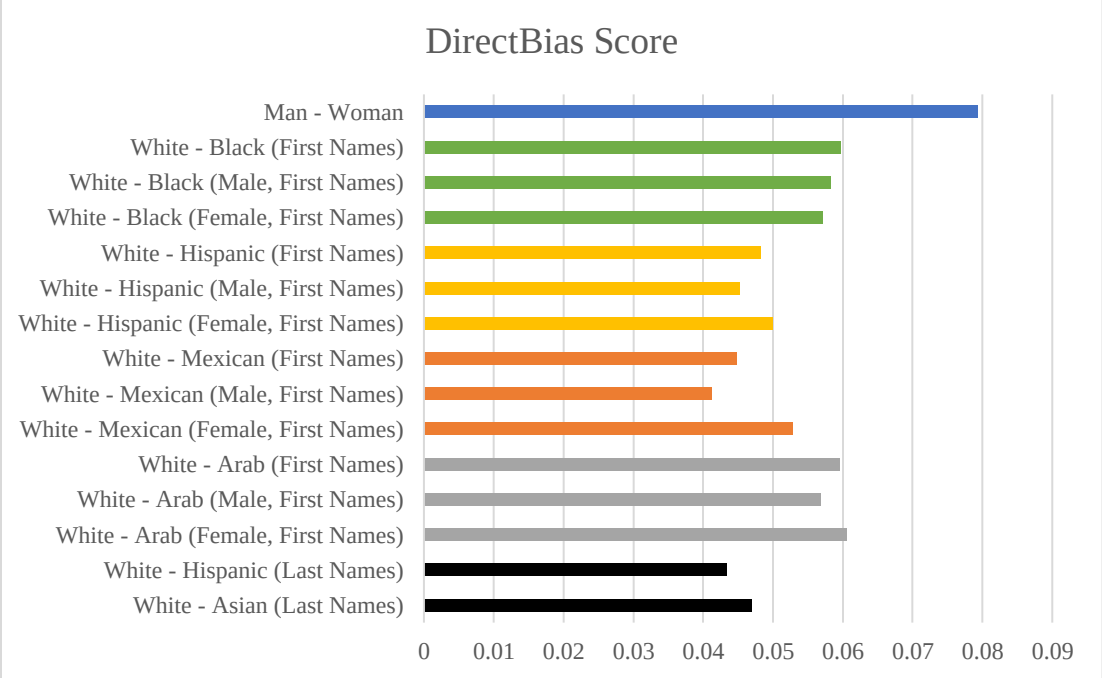

Figure 13 DirectBias score of each target direction against the occupation set. The same occupation set and target words were used as in previous examples. The list of values can be found in Appendix 11.

New Findings Using Adjectives

We also expand upon the metrics used to quantify bias in word embedding. Occupations, analogies, and pleasant versus unpleasant terms have all been used. We add adjectives to the list of sets that we test for bias. We do this because there are many known adjectives and a larger set size could mean less noise. Additionally, looking at the most extreme adjectives along a direction can give a more colorful, qualitative description of bias. Lastly, we can split adjectives into sets of positive, neutral and negative words. The percentage of each type of adjective can then serve as a new way to compare across different directions.

The adjective list that we use includes 590 adjectives. The list breaks down into about 35.93% positive adjectives, 17.97% neutral adjectives, and 46.01% negative adjectives. The list is taken from a resource on the MIT Ideonomy webpage, which includes 638 primary personality types, but our list cuts words or terms that do not appear in the word2vec model. Positive adjectives seem to almost always be used in a positive way and negative adjectives seem to be almost always used in a negative way; for example, the positive list includes words such as *admirable* and *attractive* and the negative list includes words such as *abrasive* and *careless*. The neutral list seems to be made up of words that could be neither good nor bad or are often used in both good and bad ways. For example, the list includes words like *busy* and *emotional*. However, it is often unclear why some words are included in the neutral list like *dominating* or *glamorous*, which seem to be used more in either negative or positive ways respectively. We note that while some of the classification between positive, negative, and neutral may seem arbitrary, most of the list seems truthful to our use of adjectives. We also note that the list is not evenly split between the three categories and the negative set is larger than the positive and neutral set. The full set of adjectives can be found in Appendix 7-9.

We find the most extreme adjectives projected onto a direction in a similar way to how we found the most extreme occupations. We first test along a *man-woman* direction. We find that, unsurprisingly, the most extreme adjectives for women are romantic, sexual, and relate to appearance. Whereas the most extreme adjectives for men are more varied and less shallow.

	Man	**Woman**
1	irascible	vivacious
2	phlegmatic	sexy
3	loquacious	prim
4	stolid	maternal
5	charismatic	sensual
6	ascetic	glamorous
7	cerebral	romantic
8	pugnacious	cute
9	honorable	submissive
10	miserly	libidinous

Figure 14 The top ten extreme adjectives along the man-woman direction. The list of words to create this direction is the same as the list used in Bolukbasi et al.

We then test along the *white-black*, *white-Hispanic*, and *white-Arab* target directions using both male and female first names. We see in Figure 15 that there is clear bias amongst adjectives along these target directions. We note that there are positive and negative adjectives in every list. This bias does not only hurt the disadvantaged minority group in the target direction; rather, this bias affects both sides of each target direction. Additionally, we note how some of the most extreme adjectives could have very serious effects; for example, *unpatriotic* is close to black names. Interestingly, the word *oppressed* appears twice on the extreme non-white lists and the words *submissive* or *obedient* appears on each list. This could mean that word embedding has captured some of context of the power imbalance that is frequently written about between white and non-white groups. We can deduce from Figure 16, that there are similar biases in the most extreme adjectives as projected onto the *white-Hispanic* and *white-Asian* directions that were generated using last names. Word embedding captures historic bias and popular stereotypes

	White first names	Black first names	White first names	Hispanic first names	White first names	Arab first names
1	firm	oppressed	droll	authoritarian	personable	religious
2	obsessive	peaceful	kind	sensual	retiring	oppressed
3	enthusiastic	submissive	firm	submissive	droll	peaceful
4	quirky	maternal	misguided	melancholic	solid	ascetic
5	retiring	vivacious	helpful	romantic	predatory	obedient
6	iconoclastic	unpatriotic	unsentimental	punctual	exciting	incorruptible
7	irascible	lyrical	blunt	effeminate	steely	barbaric
8	meticulous	ungrateful	folksy	familial	unsentimental	insulting
9	impractical	barbaric	callous	mystical	neat	escapist
10	pompous	sensual	dishonest	ascetic	admirable	ungrateful

Figure 15 The top ten extreme adjectives along three different directions: white-black, white-Hispanic, and white-Arab. All three directions are generated using the most common first names in the respective communities.

	White last names	Hispanic last names	White last names	Asian last names
1	stolid	authoritarian	folksy	authoritarian
2	firm	oppressed	retiring	regretful
3	urbane	familial	fiery	insincere
4	folksy	neglectful	old fashioned	repressed
5	idiosyncratic	submissive	hateful	asocial
6	abrasive	sensual	foolish	individualistic
7	natty	repentant	excitable	ascetic
8	regimental	noncompetitive	meddlesome	effeminate
9	iconoclastic	disrespectful	cantankerous	self-reliant
10	droll	N/A	miserly	unappreciative

Figure 16 The top ten extreme adjectives along two different directions: white-Hispanic and white-Asian. The directions are generated using the most common last names in the respective communities. N/A means that no more adjectives from our set were found amongst the top 100,000 closest word vectors to the direction.

along both directions. For example, *asocial* and *repressed* appear in the most extreme Asian adjectives and these are sometimes used as stereotypes about Asian-Americans.

Furthermore, we calculate the DirectBias scores along multiple directions in Figure 17. We list all of the DirectBias scores and also split up the various groups by gender in Appendix 10. We see that every target direction contains significant bias with the *white-black* direction being the most biased. We note that for many directions the DirectBias scores are quite level across the different categories of positive, neutral, and negative adjectives. There are some cases where one adjective set has a DirectBias that is much larger than the other sets. This is possibly due to the fact that there is a larger percentage of those words that are close to the direction. For example, the DirectBias score for neutral adjectives on the *white-Arab* direction is much larger than the other DirectBias scores on the *white-Arab* direction. At the same time, if we look at Figure 18, we also notice that there is large percentage of neutral adjectives on the *white-Arab* direction than in the reference set.

Lastly, we calculate the percentages of each type of adjective in Figure 18 and Appendix 13. We see that along certain directions, there are greater percentages of positive, neutral, and / or negative adjectives than the reference set. This could be because the frequency of the words in each set may not be equal. However, the variance amongst the percentages for each set along different directions would suggest that these percentages do capture some valuable information like helping explain some of the DirectBias scores. It is important to note that these percentages do not capture all of the bias in a direction. Even if a hundred percent of the adjectives close to a target along a direction are positive, there could still be significant bias. For example, women along the *man-woman* direction have a slightly higher percentage of positive adjectives than that

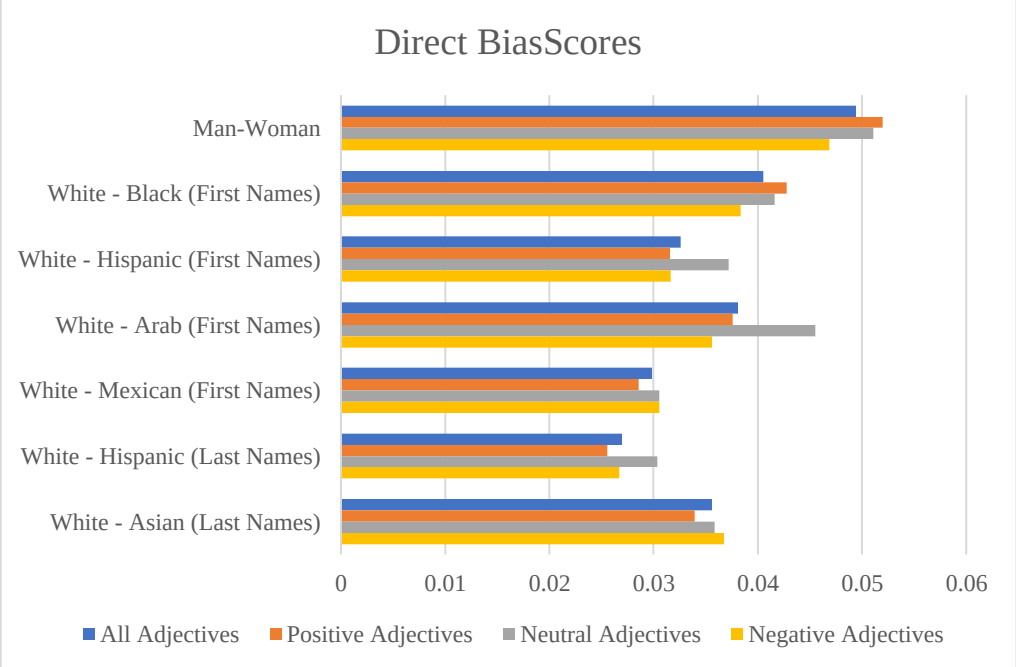

Figure 17 DirectBias score of each target direction against the adjective set. The blue row for each set includes all of the adjectives, whereas the next three rows are broken down into positive, neutral, and negative adjectives. The list of values can be found in Appendix 12.

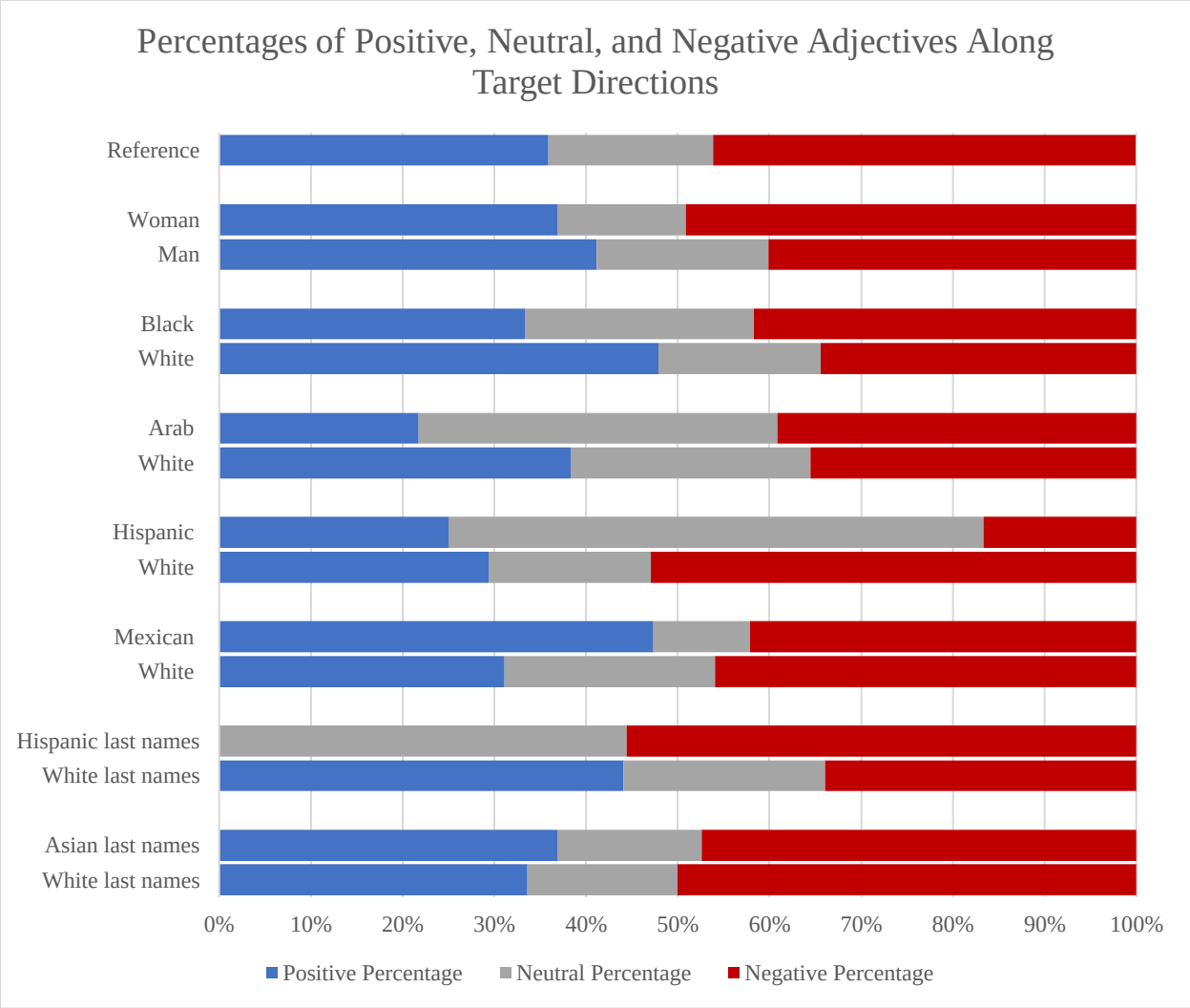

Figure 18 *This table shows the various percentages of each type of adjective along each direction for each target group. The whole list of values can be found in Appendix 13.*

in the reference set. However, as we have seen in the top ten most extreme adjectives along this direction in Figure 14, many of the adjectives close to women are mostly related to aesthetics. These extreme women adjectives clearly sexualize and objectify women and, while theoretically positive, they are actually biased in a negative way. Therefore, this implies that simple quantitative measures do not fully capture the bias inherent in word embedding. The combination of qualitative measures, like looking at the most extreme words in a set, and quantitative measures, like bias scores and percentages of positive versus neutral versus negative words are crucial to fully understanding bias.

Conclusion

We have explored definitions of fairness and bias in algorithms at large and in word embedding specifically. We also explored many previously unexplored biases that exist in word embedding. We found some startling results like how the most extreme *Hispanic* occupation is *mobster* and how the most extreme *woman* adjectives are *vivacious* and *sexy*. Additionally, we

showed how the most extreme adjectives and the percentages of positive, neutral, and negative adjectives for each target group can qualitatively and quantitatively help describe bias along a direction. Clearly, these biases need to be addressed.

We see that recently there has been a lot of work to debias word embedding along the gender direction. We see that Schmidt tries to removes the whole gender direction as a way to get rid of bias, but at the same time also removes important relationships like *man* to *woman* (Schmidt 2015). Additionally, we see that Bolukbasi et al. try to retain important gender relationships while eliminating stereotypes by removing the gender relationship from gender neutral words; however, because some gendered words are homonyms with other meanings, they also present a method to "soften" or reduce the stereotypes, while maintaining other meanings (Bolukbasi et al., 2016). Lastly, we see that some people try and correct for gender bias using corpus level constraints or lemmatization (Zhao et al., 2017; McCurdy et al., 2017).

While this recent work is making strides to debias gender in word embedding, these methods have not been applied to the other biases that we show in this paper. In future work, these algorithms should be applied to the biases that arise due to race, ethnicity, and other identity markers. Additionally, more work should be done to quantify and describe these biases. For example, rating analogies is an important indicator of bias, but it was too hard to do in this paper because of the lack of resources needed to crowd source ratings. This work should be done not only for transparency, but also to help motivate others to work on the crucial task of debiasing word embedding further.

Appendix

1. Sets of pleasant and unpleasant terms from Caliskan et al.

    • Pleasant: *caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation*.

    • Unpleasant: *abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison*.

2. A list of 218 gender-specific words taken from Bolukbasi et al. sorted by frequency:

    he, his, her, she, him, man, women, men, woman, spokesman, wife, himself, son, mother, father, chairman, daughter, husband, guy, girls, girl, boy, boys, brother, spokeswoman, female, sister, male, herself, brothers, dad, actress, mom, sons, girlfriend, daughters, lady, boyfriend, sisters, mothers, king, businessman, grandmother, grandfather, deer, ladies, uncle, males, congressman, grandson, bull, queen, businessmen, wives, widow, nephew, bride, females, aunt, prostate cancer, lesbian, chairwoman, fathers, moms, maiden, granddaughter, younger brother, lads, lion, gentleman, fraternity, bachelor, niece, bulls, husbands, prince, colt, salesman, hers, dude, beard, filly, princess, lesbians, councilman, actresses, gentlemen, stepfather, monks, ex-girlfriend, lad, sperm, testosterone, nephews, maid, daddy, mare, fiancé, fiancée, kings, dads, waitress, maternal, heroine, nieces, girlfriends, sir, stud, mistress, lions, estranged wife, womb, grandma, maternity, estrogen, ex-boyfriend, widows, gelding, diva, teenage girls, nuns, czar, ovarian cancer, countrymen, teenage girl, penis, bloke, nun, brides, housewife, spokesmen, suitors, menopause, monastery, motherhood, brethren, stepmother, prostate, hostess, twin brother, schoolboy, brotherhood, fillies, stepson, congresswoman, uncles, witch, monk, Viagra, paternity, suitor, sorority, macho, businesswoman, eldest son, gal, statesman, schoolgirl, fathered, goddess, hubby, stepdaughter, blokes, dudes, strongman, uterus, grandsons, studs, mama, godfather, hens, hen, mommy, estranged husband, elder brother, boyhood, baritone, grandmothers, grandpa, boyfriends, feminism, countryman, stallion, heiress, queens, witches, aunts, semen, fella, granddaughters, chap, widower, salesmen, convent, vagina, beau, beards, handyman, twin sister, maids, gals, housewives, horsemen, obstetrics, fatherhood, councilwoman, princes, matriarch, colts, ma, fraternities, pa, fellas, councilmen, dowry, barbershop, fraternal, ballerina

3. The list of 319 occupation words used.

    accountant, acquaintance, actor, actress, adjunct professor, administrator, adventurer, advocate, aide, alderman, alter ego, ambassador, analyst, anthropologist, archaeologist, archbishop, architect, artist, artiste, assassin, assistant professor, associate dean, associate professor, astronaut, astronomer, athlete, athletic director, attorney, author, baker, ballerina, ballplayer, banker, barber, baron, barrister, bartender, biologist, bishop, bodyguard, bookkeeper, boss, boxer, broadcaster, broker, bureaucrat, businessman, businesswoman, butcher, butler, cab driver, cabbie, cameraman, campaigner, captain, cardiologist, caretaker, carpenter, cartoonist, cellist, chancellor,

chaplain, character, chef, chemist, choreographer, cinematographer, citizen, civil servant, cleric, clerk, coach, collector, colonel, columnist, comedian, comic, commander, commentator, commissioner, composer, conductor, confesses, congressman, constable, consultant, cop, correspondent, councilman, councilor, counselor, critic, crooner, crusader, curator, custodian, dad, dancer, dean, dentist, deputy, dermatologist, detective, diplomat, director, disc jockey, doctor, doctoral student, drug addict, drummer, economics professor, economist, editor, educator, electrician, employee, entertainer, entrepreneur, environmentalist, envoy, epidemiologist, evangelist, farmer, fashion designer, fighter pilot, filmmaker, financier, firebrand, firefighter, fireman, fisherman, footballer, foreman, freelance writer, gangster, gardener, geologist, goalkeeper, graphic designer, guidance counselor, guitarist, hairdresser, handyman, headmaster, historian, hitman, homemaker, hooker, housekeeper, housewife, illustrator, industrialist, infielder, inspector, instructor, interior designer, inventor, investigator, investment banker, janitor, jeweler, journalist, judge, jurist, laborer, landlord, lawmaker, lawyer, lecturer, legislator, librarian, lieutenant, lifeguard, lyricist, maestro, magician, magistrate, maid, major leaguer, manager, marksman, marshal, mathematician, mechanic, mediator, medic, midfielder, minister, missionary, mobster, monk, musician, nanny, narrator, naturalist, negotiator, neurologist, neurosurgeon, novelist, nun, nurse, observer, officer, organist, painter, paralegal, parishioner, parliamentarian, pastor, pathologist, patrolman, pediatrician, performer, pharmacist, philanthropist, philosopher, photographer, photojournalist, physician, physicist, pianist, planner, plastic surgeon, playwright, plumber, poet, policeman, politician, pollster, preacher, president, priest, principal, prisoner, professor, professor emeritus, programmer, promoter, proprietor, prosecutor, protagonist, protégé, protester, provost, psychiatrist, psychologist, publicist, pundit, rabbi, radiologist, ranger, realtor, receptionist, registered nurse, researcher, restaurateur, sailor, saint, salesman, saxophonist, scholar, scientist, screenwriter, sculptor, secretary, senator, sergeant, servant, serviceman, sheriff deputy, shopkeeper, singer, singer songwriter, skipper, socialite, sociologist, soldier, solicitor, solicitor general, soloist, sportsman, sportswriter, statesman, steward, stockbroker, strategist, student, stylist, substitute, superintendent, surgeon, surveyor, swimmer, taxi driver, teacher, technician, teenager, therapist, trader, treasurer, trooper, trucker, trumpeter, tutor, tycoon, undersecretary, understudy, valedictorian, vice chancellor, violinist, vocalist, waiter, waitress, warden, warrior, welder, worker, wrestler, writer

4. Full list of names used to create target directions:
   woman =
   ["woman", "girl", "she", "mother", "daughter", "gal", "female", "her", "herself", "Mary"]
   man =
   ["man", "boy", "he", "father", "son", "guy", "male", "his", "himself", "John"]
   (Source: Bolukbasi et al. 2016)

   white names =
   ["Todd", "Neil", "Geoffrey", "Brett", "Brendan", "Greg", "Matthew", "Jay", "Brad", "Emily", "Anne", "Jill", "Allison", "Laurie", "Sarah", "Meredith", "Carrie", "Kristen"]
   white male names =
   ["Todd", "Neil", "Geoffrey", "Brett", "Brendan", "Greg", "Matthew", "Jay", "Brad"]

white female names =
["Emily", "Anne", "Jill", "Allison", "Laurie", "Sarah", "Meredith", "Carrie", "Kristen"]
white last names =
["Smith", "Johnson", "Miller", "Adams", "Jones", "Williams", "Davis", "Anderson", "Wilson", "Martin", "Taylor", "Moore", "Thompson", "Lewis", "Clark", "Thomas", "Hall", "baker", "Nelson", "Allen", "Harris"]
(Source: Mongabay.com)

black names =
["Rasheed", "Tremayne", "Kareem", "Darnell", "Tyrone", "Hakim", "Jamal", "Leroy", "Jermaine", "Aisha", "Keisha", "Tamika", "Lakisha", "Tanisha", "Latoya", "Kenya", "Latonya", "Ebony"]
black male names =
["Rasheed", "Tremayne", "Kareem", "Darnell", "Tyrone", "Hakim", "Jamal", "Leroy", "Jermaine"]
black female names =
["Aisha", "Keisha", "Tamika", "Lakisha", "Tanisha", "Latoya", "Kenya", "Latonya", "Ebony"]
(Source: Bolukbasi et al. 2016)

Hispanic names =
["Sofia", "Isabella", "Valentin", "Camila", "Valeria", "Luciana", "Maria", "Catalina", "Martina", "Mateo", "Santiago", "Matias", "Sebastian", "Alejandro", "Diego", "Joaquin", "Tomas", "Felipe"]
Hispanic female names =
["Sofia", "Isabella", "Valentin", "Camila", "Valeria", "Luciana", "Maria", "Catalina", "Martina"]
Hispanic male names =
["Mateo", "Santiago", "Matias", "Sebastian", "Alejandro", "Diego", "Joaquin", "Tomas", "Felipe"]
(Source: Babycenter.com)
Hispanic last names =
["Garcia", "Rodriguez", "Martinez", "Hernandez", "Lopez", "Gonzalez", "Perez", "Sanchez", "Ramirez", "Torres", "Flores", "Rivera", "Gomez", "Diaz", "Reyes", "Morales", "Cruz", "Ortiz", "Gutierrez", "Chavez"]
(Source: Mongabay.com)

Asian last names =
["Nguyen", "Lee", "Kim", "Patel", "Tran", "Chen", "Wong", "Le", "Yang", "Wang", "Chang", "Chan", "Pham", "Li", "Park", "Singh", "Lin", "Liu", "Wu", "Huang"]
(Source: Mongabay.com)

Mexican names =
["Jose", "Juan", "Miguel", "Francisco", "Alejandro", "Pedro", "Manuel", "Carlos", "Ricardo", "Maria", "Juana", "Alejandra", "Leticia", "Josefina", "Rosa", "Teresa", "Martha", "Gloria"]

Mexican male names =
["Jose", "Juan", "Miguel", "Francisco", "Alejandro", "Pedro", "Manuel", "Carlos", "Ricardo"]
Mexican female names =
["Maria", "Juana", "Alejandra", "Leticia", "Josefina", "Rosa", "Teresa", "Martha", "Gloria"]
(Source: Babycenter.com)

Arab first names =
["Mohammed", "Omar", "Ahmed", "Ali", "Youssef", "Abdul", "Abdullah", "Yasin", "Hamza", "Mariam", "Jana", "Malak", "Salma", "Nour", "Lian", "Mayar", "Fatima", "Sara"]
Arab male first names =
["Mohammed", "Omar", "Ahmed", "Ali", "Youssef", "Abdul", "Abdullah", "Yasin", "Hamza"]
Arab female first names =
["Mariam", "Jana", "Malak", "Salma", "Nour", "Lian", "Mayar", "Fatima", "Sara"]
(Source: Babycenter.com)

5. Extreme Mexican occupations:

	white male & female first names	**Mexican male & female first names**	**white male first names**	**Mexican male first names**	**white female first names**	**Mexican female first names**
1	barrister	laborer	barrister	janitor	hooker	laborer
2	screenwriter	janitor	solicitor	laborer	screenwriter	housekeeper
3	hooker	housekeeper	surveyor	housekeeper	swimmer	janitor
4	solicitor	taxi driver	publicist	taxi driver	captain	taxi driver
5	sportswriter	shopkeeper	screenwriter	undersecretary	comic	shopkeeper
6	comic	housewife	vice chancellor	waiter	investment banker	housewife
7	writer	waiter	editor	policeman	fighter pilot	saint
8	publicist	undersecretary	stockbroker	boxer	sportswriter	worker
9	swimmer	boxer	sportswriter	shopkeeper	writer	boxer
10	skipper	worker	headmaster	ballplayer	skipper	barber

Figure 19 Figure 17 The top ten occupations along three different directions: white male & female first names - Mexican male & female first names, white male first names- Mexican male first names, and white female first names- Mexican female first names. The occupation set was the 319 occupations used by Bolukbasi et al.

6. Extreme Arab occupations:

	white male & female first names	Arab male & female first names	white male first names	Arab male first names	white female first names	Arab female first names
1	organist	shop-keeper	organist	shop-keeper	organist	shop-keeper
2	sportswriter	cleric	sportswriter	cleric	patrolman	artiste
3	naturalist	taxi driver	professor emeritus	taxi driver	trooper	taxi driver
4	patrolman	artiste	cinematographer	artiste	sports-writer	laborer
5	professor emeritus	laborer	biologist	civil servant	naturalist	cleric
6	historian	civil servant	naturalist	housewife	curator	civil servant
7	trooper	housewife	hooker	cab driver	freelance writer	housewife
8	athletic director	cab driver	historian	laborer	guidance counselor	cab driver
9	biologist	parliamen-tarian	patrolman	parliament-tarian	historian	parliament-tarian
10	hooker	lawmaker	architect	lawmaker	athletic director	minister

Figure 20 The top ten occupations along three different directions: white male & female first names - Arab male & female first names, white male first names-Arab male first names, and white female first names-Arab female first names. The occupation set was the 319 occupations used by Bolukbasi et al.

7. The list of positive adjectives from http://ideonomy.mit.edu/essays/traits.html

accessible, active, adaptable, admirable, adventurous, agreeable, alert, amiable, appreciative, articulate, aspiring, athletic, attractive, balanced, benevolent, brilliant, calm, capable, captivating, caring, challenging, charismatic, charming, cheerful, clean, clearheaded, clever, colorful, compassionate, conciliatory, confident, conscientious, considerate, constant, contemplative, cooperative, courageous, courteous, creative, cultured, curious, daring, debonair, decent, decisive, dedicated, deep, dignified, directed, disciplined, discreet, dramatic, dutiful, dynamic, earnest, ebullient, educated, efficient, elegant, eloquent, empathetic, energetic, enthusiastic, esthetic, exciting, extraordinary, fair, faithful, farsighted, firm, flexible, focused, forceful, forgiving, forthright, freethinking, friendly, gallant, generous, gentle, genuine, good-natured, gracious, hardworking, healthy, hearty, helpful, heroic, honest, honorable, humble, humorous, idealistic, imaginative, impressive, incisive, incorruptible, independent, individualistic, innovative, inoffensive, insightful, insouciant, intelligent, intuitive, invulnerable, kind, knowledge, leisurely, liberal, logical, lovable, loyal, lyrical, magnanimous, masculine,

mature, methodical, meticulous, moderate, modest, neat, objective, observant, open, optimistic, orderly, organized, original, painstaking, passionate, patient, patriotic, peaceful, perceptive, perfectionist, personable, persuasive, playful, polished, popular, practical, precise, principled, profound, protean, protective, providential, prudent, punctual, purposeful, rational, realistic, reflective, relaxed, reliable, resourceful, respectful, responsible, responsive, reverential, romantic, rustic, sage, sane, scholarly, scrupulous, secure, selfless, self-reliant, sensitive, sentimental, serious, sexy, sharing, shrewd, simple, skillful, sober, sociable, solid, sophisticated, spontaneous, sporting, stable, steadfast, steady, stoic, strong, studious, suave, subtle, sweet, sympathetic, systematic, tasteful, thorough, tidy, tolerant, tractable, trusting, uncomplaining, understanding, upright, urbane, venturesome, vivacious, warm, winning, wise, witty, youthful

8. The list of neutral adjectives from http://ideonomy.mit.edu/essays/traits.html

   aggressive, ambitious, amusing, artful, ascetic, authoritarian, boyish, breezy, businesslike, busy, casual, cerebral, chummy, circumspect, competitive, complex, confidential, conservative, contradictory, crisp, cute, deceptive, determined, dominating, dreamy, driving, droll, dry, earthy, effeminate, emotional, enigmatic, experimental, familial, folksy, formal, freewheeling, frugal, glamorous, guileless, hurried, hypnotic, iconoclastic, idiosyncratic, impassive, impersonal, impressionable, intense, invisible, irreligious, irreverent, maternal, mellow, modern, moralistic, mystical, neutral, noncommittal, noncompetitive, obedient, old-fashioned, ordinary, outspoken, paternalistic, physical, placid, political, predictable, preoccupied, private, progressive, proud, pure, questioning, quiet, religious, reserved, restrained, retiring, sarcastic, sensual, skeptical, smooth, soft, solemn, solitary, stern, stolid, strict, stubborn, stylish, subjective, surprising, soft, tough, unambitious, unceremonious, unchanging, undemanding, unfathomable, unhurried, uninhibited, unpatriotic, unpredictable, unsentimental, whimsical

9. The list of negative adjectives from http://ideonomy.mit.edu/essays/traits.html

   abrasive, abrupt, agonizing, aimless, airy, aloof, amoral, angry, anxious, apathetic, arbitrary, argumentative, arrogant, artificial, asocial, assertive, barbaric, bewildered, bizarre, bland, blunt, boisterous, brittle, brutal, calculating, callous, cantankerous, careless, cautious, charmless, childish, clumsy, coarse, cold, colorless, complacent, compulsive, conceited, condemnatory, conformist, confused, contemptible, conventional, cowardly, crafty, crass, crazy, criminal, critical, crude, cruel, cynical, decadent, deceitful, delicate, demanding, dependent, desperate, destructive, devious, difficult, dirty, disconcerting, discontented, discouraging, discourteous, dishonest, disloyal, disobedient, disorderly, disorganized, disputatious, disrespectful, disruptive, dissolute, dissonant, dogmatic, domineering, dull, egocentric, enervated, envious, erratic, escapist, excitable, expedient, extravagant, extreme, faithless, false, fanatical, fanciful, fatalistic, fawning, fearful, fickle, fiery, fixed, flamboyant, foolish, forgetful, fraudulent, frightening, frivolous, gloomy, graceless, grand, greedy, grim, gullible, hateful, haughty, hedonistic, hesitant, hidebound, highhanded, hostile, ignorant, imitative, impatient, impractical,

imprudent, impulsive, inconsiderate, incurious, indecisive, indulgent, inert, inhibited, insecure, insensitive, insincere, insulting, intolerant, irascible, irrational, irresponsible, irritable, lazy, libidinous, loquacious, malicious, mannered, mawkish, mealy-mouthed, mechanical, meddlesome, melancholic, meretricious, messy, miserable, miserly, misguided, mistaken, monstrous, moody, morbid, naive, narcissistic, narrow, narrow-minded, natty, neglectful, neurotic, nihilistic, obnoxious, obsessive, obvious, odd, offhand, opinionated, opportunistic, oppressed, outrageous, paranoid, passive, pedantic, perverse, petty, phlegmatic, plodding, pompous, possessive, predatory, prejudiced, presumptuous, pretentious, prim, procrastinating, profligate, provocative, pugnacious, puritanical, quirky, reactionary, reactive, regimental, regretful, repentant, repressed, resentful, ridiculous, rigid, ritualistic, rowdy, ruined, sadistic, sanctimonious, scheming, scornful, secretive, sedentary, selfish, self-indulgent, shallow, shortsighted, shy, silly, sloppy, slow, sly, sordid, steely, stiff, strong-willed, stupid, submissive, superficial, superstitious, suspicious, tactless, tasteless, tense, thoughtless, timid, transparent, treacherous, trendy, troublesome, unappreciative, uncaring, uncharitable, unconvincing, uncooperative, uncreative, uncritical, unctuous, undisciplined, unfriendly, ungrateful, unhealthy, unimaginative, unimpressive, unlovable, unpolished, unprincipled, unrealistic, unreflective, unreliable, unrestrained, unstable, vacuous, vague, venal, venomous, vindictive, vulnerable, weak, weak-willed, willful, wishful, zany

10.

Target Direction	All Adjectives' DirectBias Score	Positive Adjectives' DirectBias Score	Neutral Adjectives' DirectBias Score	Negative Adjectives' DirectBias Score
white male first names-Mexican male first names	0.028471034	0.028950512	0.028971203	0.027904244
white female first names-Mexican female first names	0.033313593	0.031097822	0.033696434	0.034892804
White male first names-Black male first names	0.038233074	0.040928353	0.041451358	0.03488999
White female first names-Black female first names	0.040834544	0.041651519	0.041075813	0.040104647
White male first names-Hispanic male first names	0.030598116	0.030317761	0.037527799	0.028141567
White female first names-Hispanic female first names	0.037546066	0.034272838	0.037766342	0.040012225
White male first names-Arab male first names	0.036660751	0.036996789	0.041534599	0.03451739
White female first names-Arab female first names	0.04069712	0.03786991	0.048648457	0.03983123

Figure 21 DirectBias score of each target direction against the adjective set. The first column of DirectBias scores includes all of the adjectives, whereas the next three columns are broken down into positive, neutral, and negative adjectives.

11.

Target Direction	**DirectBias Score**
white male & female first names - black male & female first names	0.059723043
white male first names - black male first names	0.058271542
white female first names - black female first names	0.057107878
white male & female first names - Hispanic male & female first names	0.048283132
white male first names - Hispanic male first names	0.045172556
white female first names - Hispanic female first names	0.049903152
white male & female first names - Mexican male & female first names	0.044727039
white male first names - Mexican male first names	0.041221998
white female first names - Mexican female first names	0.052790098
white male & female first names - Arab male & female first names	0.05959575
white male first names - Arab male first names	0.056844909
white female first names - Arab female first names	0.060595805
white last names - Hispanic last names	0.043413032
white last names - Asian last names	0.046879433

Figure 22 DirectBias score of each target direction against the occupation set. The same occupation set and target words were used as in previous examples.

12.

Target Direction	All Adjectives' DirectBias Score	Positive Adjectives' DirectBias Score	Neutral Adjectives' DirectBias Score	Negative Adjectives' DirectBias Score
White first names-Black first names	0.040540138	0.042759361	0.041640974	0.038385495
White first names-Hispanic first names	0.032626744	0.031600599	0.037230704	0.031649269
White first names-Mexican first names	0.029838708	0.028584323	0.030552701	0.030540768
White first names-Arab first names	0.03806921	0.037558254	0.045500821	0.035598634
White last names-Hispanic last names	0.026947994	0.025543527	0.030357262	0.026726573
White last names-Asian last names	0.035581341	0.033949944	0.035860648	0.03674505

Figure 23 DirectBias score of each target direction against the adjective set. The first column of DirectBias scores includes all of the adjectives, whereas the next three columns are broken down into positive, neutral, and negative adjectives. The list of values can be found in Appendix 12.

13. The percentages of positive, neutral, and negative adjectives across each direction for each target group.

	Positive Percentage	Neutral Percentage	Negative Percentage
Reference	35.93%	17.97%	46.01%
Woman	36.84%	14.04%	49.12%
Man	41.18%	18.72%	40.11%
Black	33.33%	25.00%	41.67%
White	47.92%	17.71%	34.38%
Arab	21.74%	39.13%	39.13%
White	38.32%	26.17%	35.51%
Hispanic	25.00%	58.33%	16.67%
White	29.41%	17.65%	52.94%
Mexican	47.37%	10.53%	42.11%
White	31.08%	22.97%	45.95%
Hispanic last names	0.00%	44.44%	55.56%
White last names	44.07%	22.03%	33.90%
Asian last names	36.84%	15.79%	47.37%
White last names	33.64%	16.36%	50.00%
Mexican Male	36.84%	15.79%	47.37%
White Male	36.84%	21.05%	42.11%
Mexican Female	39.39%	12.12%	48.48%
White Female	31.33%	14.46%	54.22%
Black Male	20.00%	30.00%	50.00%
White Male	50.59%	18.82%	30.59%
Black Female	33.33%	33.33%	33.33%
White Female	47.03%	16.76%	36.22%
Hispanic Male	22.73%	36.36%	40.91%
White Male	41.11%	24.44%	34.44%
Hispanic Female	28.57%	35.71%	35.71%
White Female	25.93%	13.33%	60.74%
Arab Male	29.63%	22.22%	48.15%
White Male	48.89%	25.56%	25.56%
Arab Female	16.67%	50.00%	33.33%
White Female	32.00%	19.20%	48.80%

Figure 24 This table shows the various percentages of each type of adjective along each direction for each target group.

14. DirectBias score of each target direction against the occupation set.

Target Direction	DirectBias Score
man-woman	0.080186774
whites-blacks	0.040086738
whites-Latinos	0.04859312
straight-gay	0.085444394
Christian-Jew	0.066810632
Christian-Muslim	0.048058418
whites-minorities	0.045046315

Figure 25 DirectBias score of each target direction against the occupation set.

Work Cited

"Baby Names." BabyCenter, www.babycenter.com/baby-names.

Banarjee, Amitav. "Hypothesis Testing, Type I and Type II Errors." Industrial Psychiatry Journal, vol. 18.2, 2009, pp. 127–131.

Barry-Jester, Anna Maria, et al. "The New Science of Sentencing." The Marshall Project, 8 Nov. 2017, www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing.

Bolukbasi, Tolga, et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." 2016.

Byrnes, Nanette. "How Do You Feel about Artificial Intelligence?" MIT Technology Review, MIT Technology Review, 20 May 2016, www.technologyreview.com/s/600996/artificial-intolerance/.

Caliskan, Aylin, et al. "Semantics derived automatically from language corpora contain human-like biases." 2017.

Chakraborty, Tuhin, et al. "Reducing gender bias in word embeddings." 2016.

Clifton, David A. "Machine learning for healthcare technologies - an introduction." Machine Learning for Healthcare Technologies, 2016, pp. 1–6., doi:10.1049/pbhe002e_ch1.

Copeland, Michael. "The Difference Between AI, Machine Learning, and Deep Learning? | NVIDIA Blog." The Official NVIDIA Blog, 3 Aug. 2017, blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/. Accessed 3 Sept. 2017.

Corbett-Davies, Sam, et al. "Algorithmic Decision Making and the Cost of Fairness." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 17, 2017, doi:10.1145/3097983.3098095.

Couch, Christina. "Ghosts in the Machine." PBS, Public Broadcasting Service, 25 Oct. 2017, www.pbs.org/wgbh/nova/next/tech/ai-bias/.

Dahal, Govinda, et al. "Challenges in measuring gender and minorities." Https://Unstats.un.org/Unsd/Gender/Rome Dec2007/Docs/2.3 Me.Pdf, 2017.

Dobrev, Dimiter. "A Definition of Artificial Intelligence." ArXiv, 19 Jan. 2004, doi:https://arxiv.org/pdf/1210.1568.pdf. Accessed 3 Sept. 2017.

Domingos, Pedro. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Books, 2015

Diakopoulos, Nicholas, and Michael Koliska. "Algorithmic Transparency in the News Media." Digital Journalism (2016): 10. Web.

Dwork, Cynthia, et al. “Decoupled classifiers for fair and efficient machine learning.” 21 July 2017.

Guittar, Stephanie, and Nicholas Guittar. “Intersectionality.” Elsevier, vol. 12, 2015.

Hajian, Sara, and Josep Domingo-Ferrer. “A Methodology for Direct and Indirect Discrimination Prevention in Data Mining.” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 7, 2013, pp. 1445–1459., doi:10.1109/tkde.2012.72.

Heuer, Hendrik. “Text comparison using word vector representations and dimensionality reduction.” 2015.

Hunt, D. Bradford. Redlining, Encyclopedia of Chicago. 2005.

Jernigan, Carter, and Behram F.t. Mistree. “Gaydar: Facebook friendships expose sexual orientation.” First Monday, vol. 14, no. 10, 2009, doi:10.5210/fm.v14i10.2611.

Joseph, Matthew, et al. “Fairness in Learning: Classic and Contextual Bandits.” 2016.

Kleinberg, Jon, et al. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” 19 Sept. 2016, doi:arXiv:1609.05807.

McCurdy, Katherine, and Serbetc ¸i Og˘uz. “Grammatical gender associations outweigh topical gender bias in cross linguistic word embeddings.” Babbel, 14 June 2017.

Mikolov, Tomas, et al. “Distributed Representations of Words and Phrases and their Compositionality.” 2013.

Miller, Claire Cain. “Can an Algorithm Hire Better Than a Human?” The New York Times, The New York Times, 25 June 2015, www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html.

Mongabay. “Most Common First Names and Last Names.” Mongabay.com, 5 Nov. 2002, names.mongabay.com/.

Pedreschi, D., et al. “Discrimination-Aware Data Mining.” ACM Int’l Conf. Knowledge Discovery and Data Mining, 2008, pp. 560–588.

Rawls, John. A Theory of justice. Belkinap Press of Harvard University Press, 1999.

Richards, Whitman. “Ideonomy.” Ideonomy, ideonomy.mit.edu/.

Romei, Andrea, and Salvatore Ruggieri. “A multidisciplinary survey on discrimination analysis.” The Knowledge Engineering Review, vol. 29, no. 05, Mar. 2013, pp. 582–638., doi:10.1017/s0269888913000039.

Rudin, Cynthia. “Predictive Policing: Using Machine Learning to Detect Patterns of Crime.” Wired, Conde Nast, 6 Aug. 2015, www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime.

Russell, Stuart J., and Peter Norvig. Artificial intelligence a modern approach. Prentice Hall, 2003.

Schmidt, B. Rejecting the gender binary: a Vector-Space operation. 2015, bookworm.benschmidt. org/posts/2015-10-30-rejecting-the-gender-binary.html.

Selbst, Andrew D. “Disparate Impact in Big Data Policing.” SSRN Electronic Journal, 2017, doi:10.2139/ssrn.2819182.

Simonite, Tom. “Study Suggests Google's Ad-Targeting System May Discriminate.” MIT Technology Review, MIT Technology Review, 7 July 2015, www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/.

Zhao, Jieyu, et al. “Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints.” 2017.

Žliobaitė, Indrė. “Measuring discrimination in algorithmic decision making.” Data Mining and Knowledge Discovery, vol. 31, no. 4, 2017, pp. 1060–1089., doi:10.1007/s10618-017-0506-1.

Addendum

Economic Analysis

Algorithmic bias, and specifically bias in word embedding, can have a large impact on the economic activity of specific organizations, the economic prospects of different groups of people, and the economy as a whole. These different economic effects are related; however, we will look at each of them separately. Additionally, these biases can affect various industries differently. In fact, the second half of the paper highlights the biases present in word embedding which can influence different occupations and bias their industries towards certain groups of people. This addendum to the paper will not go over these extensive results. Altogether the business case for analyzing the effects of algorithmic bias is strong and varied.

First, we look at how algorithmic bias can affect the economic activity of specific organizations. In section *Algorithmic Bias*, on pages 5-7, we see the many different types of algorithmic bias. We see that they can be used to discriminate against an array of people for different reasons. Ultimately, these biases will result in far less diversity in a company's personnel and clientele. If algorithms are used to help hire for a company, then the company's personnel makeup will be less diverse. We saw earlier in the paper that companies will continue to hire from the same group of people; in the United States, this will mean that large companies will continue to hire rich, straight, white and protestant men. On the other hand, according to the European Commission Directorate-General for Employment, Industrial Relations and Social Affairs, we see that diversity can provide two principal types of economic benefits. Firstly, diversity can strengthen long term "value-drivers," or the assets that allow companies to be competitive, generate cash flows and satisfy shareholders (European Commission, 2003). For example, diversity of personnel can strengthen human, organizational, and knowledge capital. Additionally, diversity can help generate short and medium term opportunities to improve cash flows with return-on-investment benefits by reducing costs, opening up new markets, and improving performance in existing markets (European Commission, 2003). For example, more female workers can better position a company to sell to a female demographic. Additionally, if algorithms are used to help decide who a company should serve, then the company will have a much less diverse set of clientele. This can inhibit a company's growth into new markets, prevent a company from recognizing new trends, and work in an increasingly global environment.

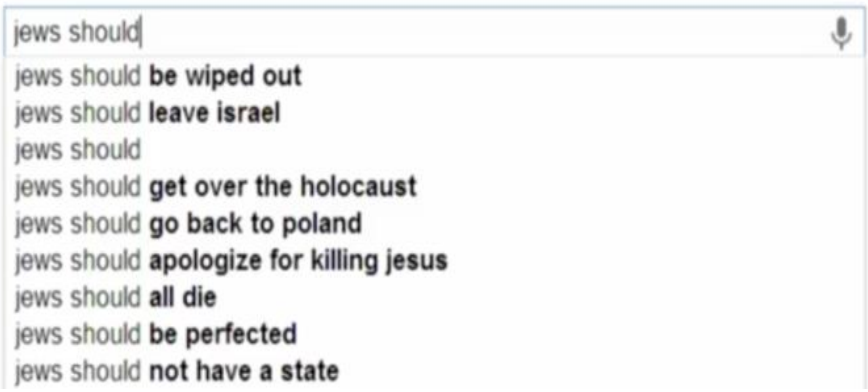

Figure #1: Screengrab: Kate Crawford's "The Trouble With Bias" at NIPS 2017 (Fussel, 2017).

Specific organizations can also be hurt by the negative consequences of algorithms and algorithmic bias. For example, Microsoft's Tay Twitter chatbot and its quick transformation into a bigoted racist caused a significant setback in public perception of AI. Within twenty-four hours, the chatbot had to be taken offline and people were very displeased with Microsoft. These biases can be found in many places; looking at Figure 1, it is obvious that Jewish people using Google would be offended by the proposed search suggestions and may harbor negative feelings towards Google because of it. Moreover, we have seen an increase in public, political, and academic awareness of algorithmic bias in 2017, so much so that the New York City Council recently passed what may be the US' first AI transparency bill (Ip, 2018; Fussel, 2017). Companies must now be much more warry of the effects of algorithmic bias because of the increased awareness, far reaching effects, and large audiences that these algorithms have.

Economically, algorithmic bias can affect different groups of people and the companies who serve them. We see on pages 5-7 the different types of algorithmic bias that can affect different groups of people. We saw from this section that people from certain geographic areas can be discriminated against using *redlining*. This can prevent geographic mobility and decrease the ability of companies to find new geographic markets. We also see that groups that have been the victim of bias and oppression in the past can experience even more bias with algorithmic bias. This can occur because of the bias types *historical discrimination*, *encoding existing bias*, and *data collection feedback loops* referenced on pages 5-7. These types of bias prevent social mobility. Moreover, a lack of social mobility has been shown to slow the economy and curb economic growth (Reeves, 2016). Additionally, we see that specific groups of people can be targeted from the bias types *explicit discrimination*, *scarcity of minority population data*, and *sensitive attribute as proxy* referenced on pages 5-7. These bias types will further prevent social mobility and will also stop companies from understanding, serving, and benefiting from new groups of people.

Additionally, it was shown in sections *Gender Bias in Word Embedding* on pages 14-17 and *New Findings using Occupations* on pages 18-24 that word embedding can discriminate against certain groups of people. Because certain occupation words are closer to certain names within different groups of people, members of certain groups are more likely to come up in searches and be sought out for some jobs. This type of bias can hurt the members of certain groups and limit the diversity of personnel and cliental in those occupations. Additionally,

sections *Other Biases in Word Embedding* on pages 17-18 and *New Findings using Adjectives* on pages 24-27, show that different groups will be discriminated against in searches and research with adjectives and descriptions. This may hurt the ability of companies to properly understand their demographics, market to the right people, and create the best products.

Algorithmic bias will also affect various industries differently. Industries that rely more on algorithms and deal with different constituent groups of people will be more strongly affected by algorithmic bias. Additionally, companies that rely on search and other Natural Language Processing tasks will be more sensitive to the biases inherent in word embedding. Credit Suisse has released a report highlighting how much algorithmic bias will affect fintech or the financial technology industry. Historical bias is teaching algorithms to favor certain groups, like men over women, even though the other groups are actually better suited for that financial service. The news industry is being affected in very strong ways by these biases. Facebook and other mediums for news distribution may provide news selectively to different groups, which may create "echo chambers," news pockets, and bad press or condemnation. We also see that algorithmic bias may increase the inequalities between racial, social and economic divides in healthcare (Hart, 2017). This risk comes from the existing biases in healthcare data due to current inequalities. Additionally, the biases present in drug trials may be exacerbated by algorithmic bias, which means that women, the elderly, and minorities may be more likely to suffer detrimental side effects from new medications and procedures (Hart, 2017). Algorithmic bias also opens doors to biased diagnoses. This will only further build systems that negatively affect the way that certain groups receive healthcare.

Algorithmic bias and the bias in word embedding will have large economic effects. In fact, many economists are working towards understanding algorithmic bias using economic models. We have shown throughout this paper and in this addendum that algorithmic bias is dangerous for individual businesses, members of various groups in society, and the economy as a whole. Therefore, the study of the economic effects of algorithmic bias warrants study and focus.

Work Cited for Addendum

European Commission Directorate-General for Employment, Industrial Relations and Social Affairs. *THE COSTS AND BENEFITS OF DIVERSITY A Study on Methods and Indicators to Measure the Cost-Effectiveness of Diversity Policies in Enterprises.* Centre for Strategy and Evaluation Services, Oct. 2003, www.coe.int/t/dg4/cultureheritage/mars/source/resources/references/others/17%20-%20Costs%20and%20Benefits%20of%20Diversity%20-%20EU%202003%20ExSum.pdf.

Credit Suisse. “Sponsored: Tech emerges as a contributor to gender bias.” Quartz, Quartz, 10 Nov. 2017, qz.com/1121150/algorithmic-bias-a-new-fintech-challenge/.

Fussell, Sidney. “AI Professor Details Real-World Dangers of Algorithm Bias [Corrected].” Gizmodo, Gizmodo.com, 8 Dec. 2017, gizmodo.com/microsoft-researcher-details-real-world-dangers-of-algo-1821129334.

Hart, Robert. “If you're not a white male, artificial intelligences use in healthcare could be dangerous.” Quartz, Quartz, 10 July 2017, qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/.

Ip, Chris. “In 2017, society started taking AI bias seriously.” Engadget, 30 Jan. 2018, www.engadget.com/2017/12/21/algorithmic-bias-in-2018/.

Reeves, Richard V. “The Economic Case for Social Mobility.” Brookings, Brookings, 28 July 2016, www.brookings.edu/opinions/the-economic-case-for-social-mobility/.

Vincent, James. *Twitter Taught Microsoft's Friendly AI Chatbot to Be a Racist Asshole in Less than a Day*. The Verge. The Verge, 24 Mar. 2016