# Bias and Fairness in Machine Learning

Abhishek Tiwari

Citation: *A. Tiwari*, "Bias and Fairness in Machine Learning", Abhishek Tiwari, 2017. doi:10.59350/7tdcs-4q330

Published on: July 03, 2017

In AI and machine learning, the future resembles the past and bias refers to prior information. There has been a growing interest in identifying the harmful biases in the machine learning. Often these harmful biases are just the reflection or amplification of human biases which algorithms learn from training data. Some training data sets such as text, medical, criminal, educational, financial etc. are more susceptible to human biases compared to others. For example, weather data is little or not impacted by human bias.

## Harmful biases

In machine learning, algorithmic biases are new kinds of bugs. These bugs generically referred as *unwarranted associations*. Such bugs can be harmful to both people and businesses. Tramer et al. [1] argue we should proactively check for unwarranted associations, debug, and fix them with the same rigor as we do to other security and privacy bugs. There are concerns that harmful biases often keep alive the prejudice and unfairness. For instance, biases present in the word embedding (i.e. which words are closer to she than to he, etc.) trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. Most of these biases are implicit and hard to recognize. With some work and mix of good ethical practices these hidden biases can be uncovered, fixed, and learnt from.

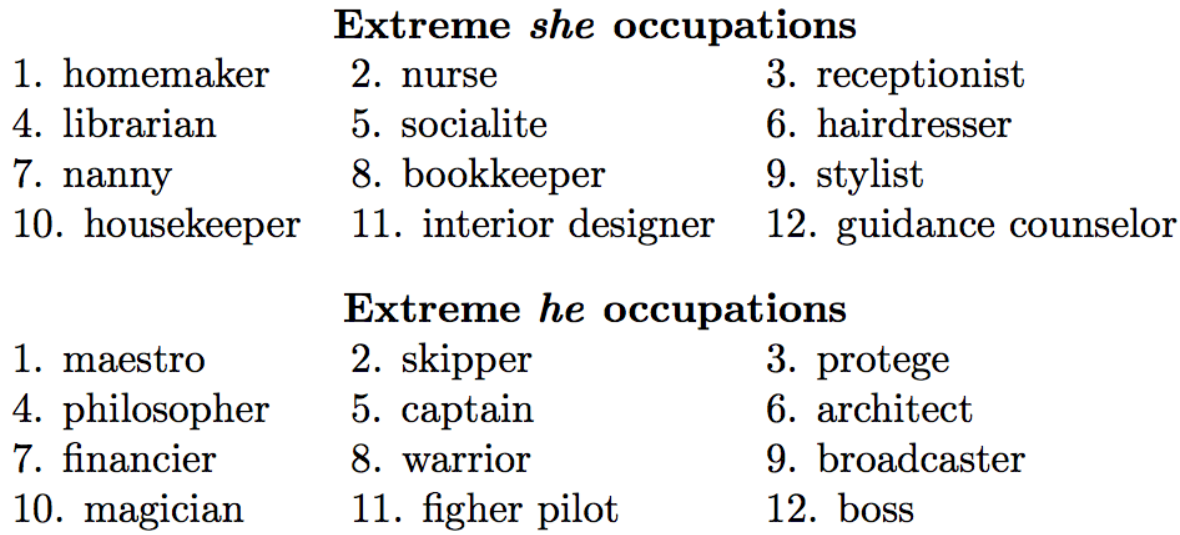


**Figure 1:** Gender biases in word embeddings - occupations as projected on to the she vs he

Although not comprehensive, following list highlights some of the well-known unwarranted associations,

[1] FairTest: Discovering Unwarranted Associations in Data-Driven Applications

- Google's image tagger was found to associate racially offensive labels with images of black people
- Wall Street Journal investigators showed that Staples' online pricing algorithm discriminated against lower-income people
- Black people were more likely to be assessed as having a higher risk of recidivism when using commercial prediction tools such as COMPAS
- An insurance company that used machine learning to workout insurance premiums involuntarily discriminated against elderly patients
- A credit card company used tracking information to personalize offers steering minorities into higher rates

## Fairness

The goal of fairness [2] in machine learning is to design algorithms that make fair predictions across various demographic groups. It is important to differentiate between *outcomes fairness* and the *process fairness*. Process fairness [3] relies on the use of appropriate features in order to make fair decisions. If undesirable features are used for prediction, then despite the use of unbiased data outcomes can be unfair. For instance, it can be fair to use criminal history of an individual as an input feature in the decision-making process, but unfair to use family criminal history of the individual in the question. Obviously, removing an undesirable feature improves process fairness, it may also lead to reduced accuracy or lower outcome fairness [4] - a little cost to pay when humanity is at the stake.

Machine learning algorithms particularly supervised learning methods can be unfair for several reasons,

- Data might encode existing biases (For instance, Caliskan et al.[5] demonstrated that how language itself contains human-like biases)
- Data collection circular dependency (For instance, to get a credit card you need credit history, to have a credit history you need credit card)
- Different populations with different properties (SAT score might correlate with label differently in populations that employ SAT tutors)
- Use of protected attributes such as race, color, religion, gender, disability, or family status
- Less data (by definition) about minority populations (For instance, Asian prefer cash transaction which results in less data related to credit history)

[2] Fairness Through Awareness
[3] The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making
[4] On the relation between accuracy and fairness in binary classification and Algorithmic decision making and the cost of fairness
[5] Semantics derived automatically from language corpora necessarily contain human biases

In recent years, there has been a lot of research work on how machine learning algorithms can ensure fairness. These include dealing with biases through unawareness/blindness, awareness/lipschitz property [6] [7], individual fairness, statistical parity/group fairness, counterfactual fairness [8], demographic parity/disparate impact [9], preference-based [10], and equality of opportunity [11] [12].

## Measuring unfairness or how to detect the biases?

Hardt et al. suggest that measuring unfairness is a lot easier than when compared to discovering or proving fairness.

### Association tests

Tramer et al. [13] proposed unwarranted associations (UA) framework - a scalable and statistically rigorous methodology for the discovery of unfair, discriminatory, or offensive user treatment in machine learning algorithms. UA framework methodology and incorporates the three core investigation primitives - testing, discovery, and error profiling. FairTest - a tool which implements UA framework - enables the investigation of associations between algorithmic outcomes and sensitive user attributes (such as race or gender). FairTest also provides debugging capabilities.

Inspired from Implicit Association Test (IAT), Caliskan et al. used the association tests to measure the association between two sets of words for possible bias and prejudice in the text corpus.

### Perturbation tests

Basic mechanics is quite simple - if a small perturbation to an input feature dramatically changes the output of the model then the model is sensitive to the feature. For example, FairML[14][15] uses the orthogonal projection of input as a perturbation scheme.

[6] Fairness Through Awareness
[7] Fairness through Awareness Slides
[8] Counterfactual Fairness
[9] Certifying and Removing Disparate Impact
[10] From Parity to Preference-based Notions of Fairness in Classification
[11] Equality of Opportunity in Supervised Learning
[12] Attacking discrimination with smarter machine learning
[13] FairTest: Discovering Unwarranted Associations in Data-Driven Applications
[14] FairML is a python toolbox auditing the machine learning models for bias.
[15] FairML: Auditing Black-Box Predictive Models

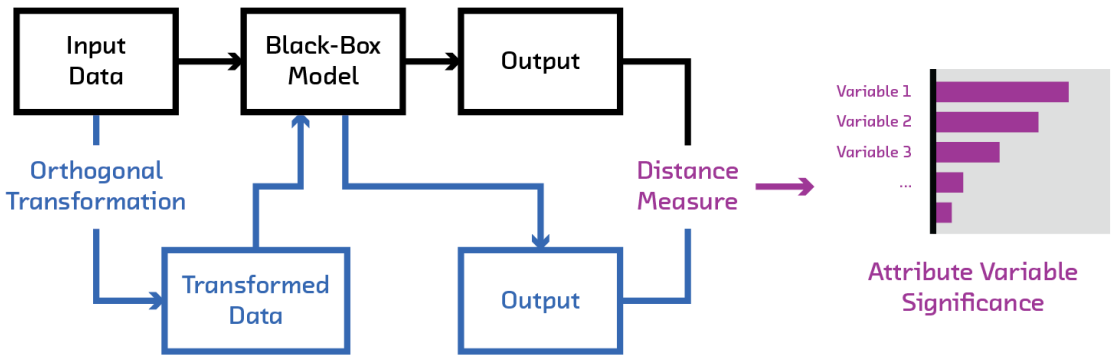


**Figure 2:** Using FairML assess the model's fairness or discriminatory extent. FairML orthogonally projects the input to measure the dependence of the predictive model on each attribute. Image credits fastforwardlab.com

## Achieving fairness or how to fix the biases?

There are three different strategies to reduce and even eliminate biases,

- Pre-processing - eliminating any sources of unfairness in the data before the algorithm is formulated.
- In-processing - making fairness adjustments as part of the process by which algorithm is constructed.
- Post-Processing - after the algorithm is applied, its performance is adjusted to make it fairer.

These three strategies can be combined depending on accuracy/fairness requirements.

### Debiasing data

If the biases are represented in data, there are ways to remedy them either by neutralizing (hard debiasing) or equalizing (soft debiasing). Bolukbasi et al. [16] proposed a method to *debias* the vector space. They applied this approach to remove both direct and indirect gender biases in word embeddings. This approach requires algebraic formulation of bias. This approach can be considered similar to fairness through blindness/unawareness. It does solve the issue of biases only to an extent, but it has an important limitation prejudice can creep back in through proxies.

[16] Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings


**Figure 3:** Debiasing Word Embeddings

Bolukbasi et al. used following approach to remove gender associations,

> The first step, called Identify gender subspace, is to identify a direction (or, more generally, a subspace) of the embedding that captures the bias. For the second step, we define two options: Neutralize and Equalize or Soften. Neutralize ensures that gender neutral words are zero in the gender subspace. Equalize perfectly equalizes sets of words outside the subspace and thereby enforces the property that any neutral word is equidistant to all words in each equality set. For instance, if {grandmother, grandfather} and {guy, gal} were two equality sets, then after equalization babysit would be equidistant to grandmother and grandfather and also equidistant to gal and guy, but presumably closer to the grandparents and further from the gal and guy. This is suitable for applications where one does not want any such pair to display any bias with respect to neutral words