# Designing Against Bias: Identifying and Mitigating Bias in Machine Learning and AI

David J Corliss[1]

[1] Peace-Work, Plymouth MI 40170, USA

**Abstract.** Bias in machine learning algorithms is one of the most important ethical and operational issues in statistical practice today. This paper describes common sources of bias and how to develop study designs to measure and minimize it. Analysis of disparate impact is used to quantify bias in existing and new applications. New open-source packages such as Fairlearn and AI Fairness 360 Toolkit quantify bias by automating the measurement of disparate impact on marginalized groups, offering great promise to advance the mitigation of bias. These design strategies are described in detail with examples. Also, a comparison algorithm can be developed that is designed to be fully transparent and without features subject to bias. Comparison to this bias-minimized model can identify areas as bias in other algorithms.

**Keywords:** Bias Mitigation, Machine Learning, AI, Disparate Impact, Fairlearn.

## 1 Introduction: Bias in Machine Learning and AI

### 1.1 Overview

One of the driving purposes of using of developing artificial intelligence and machine learning algorithms was to improve the fairness of processes that were relied on that had previously relied on human judgment the idea was that by taking the human element out of a decision process that the process was supposed to become more fair unfortunately experience has proven that this process often hasn't worked out AI machine learning and AI have often failed their promise of developing more fair processes. This is a motivating factor in investigating these weaknesses and failures of machine learning and AI, identifying root causes, and developing mitigation strategies and tools to minimize their impact.

### 1.2 Important Examples

One well-known example of algorithm failure is the COMPAS algorithm used by the criminal justice systems in several states ([1] Kirkpatrick 2017). The *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS) is an algorithm predicting risk of recidivism developed by the justice tech company Northpointe and now owned by Equivant. The algorithm has been used to inform decisions to permit or deny bail, set bail amounts, in sentencing, and by parole boards. In 2016, an

investigation by ProPublica ([2] Angwin, [3] Larson et al. 2016) found persons of color were much more likely than whites for a false positive High Risk classification. At the same time, whites were found to have a much higher likelihood of false *negative* than blacks. Despite
serious and well-documented failures, algorithms of this type with bias issues
continue to be used ([4] Thomas, Pontón-Núñez 2022).

Another example is the use of algorithms to screen resumes, where gender bias has sometimes been found. In 2018, the news media agency Reuters reported the tech company Amazon chose not to deploy a resume screening algorithm it had developed after finding evidence of gender bias. As reported by Reuters ([5]Dastin 2018), the company planned to eliminate problems such as gender bias and other concerns by implementing an algorithm and hiring the very highest scores without further review. By contrast, the algorithm was found to replicate a bias it was intended to address.

## 2 Types of Bias and Their Root Causes

### 2.1 Overview

### 2.2 Confirmation Bias

This familiar form of bias results from the human tendency to acquire and retain information, analyze data, and develop explanations that confirm their preexisting beliefs ([6] Nickerson 1998). Confirmation bias is a well-established phenomenon in the field of psychology and cognitive science, and has been extensively studied and documented in a wide range of domains. This phenomenon is so robust that it has been observed in diverse cultures, ages, and educational levels.

This bias is thought to be driven by a number of factors, including cognitive dissonance, motivated reasoning, and the need for consistency ([7] Lewandowsky and Cook 2013). In data science, confirmation bias can adversely impact thoroughness in considering and investigating issues of algorithm performance.

### 2.3 Sampling Bias

Sampling bias results from errors in the manner in which data are selected from a population, especially in the case of data used to train algorithms. This type of bias produces a sample that is not representative of the population from which it is drawn. This can occur when the sample is not selected randomly or when certain groups within the population are underrepresented or overrepresented in the sample. As a result, the sample may not accurately reflect the characteristics or opinions of the population as a whole. This bias can lead to inaccurate conclusions and generalizations about the population based on the sample ([8] Lavrakas 1993).

A biased, unrepresentative sample can occur in several different ways. In some cases, study samples are taken from all people who volunteer to provide data. This problem, called Convenience Sampling, is especially common in social media surveys

([9] Groves et a. 2004). Another common sampling error is a sample evenly representative of the population as a whole, with the result that small subsets of the population lack sufficient examples to train the algorithm accurately.

An example of this failing to over-sample small sub-populations can be found in some voice recognition systems. In some cases, these algorithms are found to more accurately is easier for them to recognize a voice and correctly interpret speech from certain demographic segments rather than others. If the samples of people used to train the algorithm are not oversampled for small subsets of the population, poor algorithm performance can result.

An example of this is described by A. Najibi ([10] 2020) and a team from Harvard University, which evaluated the accuracy of facial recognition technologies from several leading different companies. This study found facial recognition for all the products was most accurate for lighter skinned males and least accurate for darker skinned females. This was the result of not properly oversampling people with darker skin in the training set, who constitute a small portion of the US population.

### 2.4 **The History Problem**

The history problem results from training an algorithm where the labeling of the data is taken from previous human decisions. While the intention may have been to develop an algorithm to reduce bias in a human, subjective process, use of labeled data from previous biased human decisions merely trains the algorithm to replicate the human bias.

In cases such as these, the term “Prejudice” is often used in literature (e.g. [11] Emspak 2016), as it often results from personal prejudices of the people who labeled the historical data being included in the training set for the new algorithm. The failure of both the COMPAS recidivism algorithm and the Amazon resume screening system are examples of this. However, this problem can result from causes unrelated to prejudice – for example, an algorithm for detecting quality variations in manufacturing. Consequently, the author will recommend use of the term History Problem for all instances of algorithm failure resulting from the error of using historical subjective human labeling of training data.

### 2.5 **The Spaghetti Problem**

This issue arises where hundreds or even thousands of potential predictors are present. While this most often happens in NLP classification algorithms, the key feature is the very large number of potential features, such as can be found in genetic testing, high volume sensor data, and other situations. In such cases, a large number of candidate predictors may not be carefully screened for bias. This will be termed the “Spaghetti Problem”, from the aphorism “anything that sticks” in the belief of some that spaghetti will stick to a wall when fully cooked. The uncritical acceptance of predictors without carefully testing individual candidate features for bias can result in the inclusion of biased features in the algorithm.

An example of this is found in the Amazon resume screening algorithm, where

candidates were rated lower by the algorithm if the word “softball” was found in the resume ([12] Rodriguez-Villa 2022). As women are much more likely to play softball in college than men, the inclusion of the term as a predictor in the algorithm with a negative impact on the score contributed to a bias against women job candidates.

### 2.6 **Lack of Transparency**

While not a source of bias in and of itself, a lack of algorithm transparency can greatly complicate testing, identification and confirmation of potential bias, and mitigating the effects. One of the most important problems with the COMPAS was the implementation of a black box algorithm due to withholding of information deemed proprietary by the government supplier who developed it. The features included in the COMPAS algorithm were withheld from the people using the algorithm and from the people whose lives were significantly affected by its use.

## 3 **Bias Mitigation**

### 3.1 **Measuring Bias With Disparate Impact**

Disparate Impact is often used to investigate the difference in the impact of an event between different sub-populations, such as by race, gender, age etc. For example, one investigation of the COVID-19 pandemic initial wave (through June 30, 2020) resulted in higher mortality rates (Figure 1) for several marginalized populations such as BIPOC ([13] Corliss 2021).

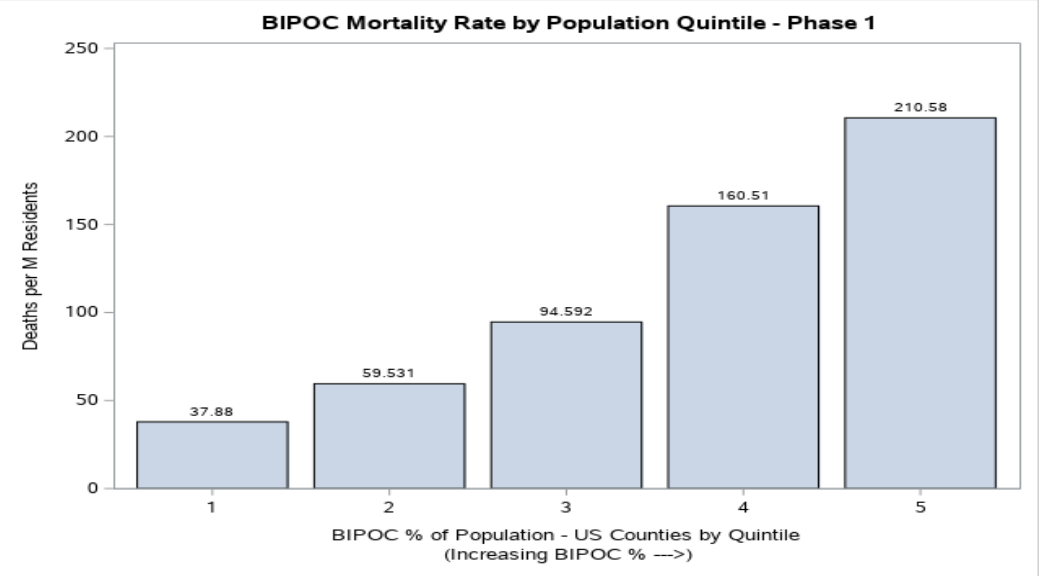


**Fig. 1.** COVID-19 Mortality - Wave 1: March 1 - June 30, 2020), US Counties by Quintile

Disproportionate impact on marginalized groups can be measured using relative risk ([14]Heilbrun 2023), log odds, or odds ratios of a given event occurring within each population subset (Table 1). Comparison of these metrics across population subsets provide useful metrics for quantifying bias ([13] Corliss 2021).

| Characteristic or Risk Factor | Odds Ratio |
|---|---|
| Black / African American | 10.1 |
| Cardiovascular Disease | 9.3 |
| Chronic Lung Disease | 5.9 |
| Prison Populations ([15] Saloner et al. 2020) | 5.5 |
| Indigenous | 3.3 |
| Poverty (% Below Poverty Line) | 2.9 |
| High Population Density | 1.9 |

**Table. 1.** COVID-19 Wave 1 Odds Ratios for U.S. Population Subjects

### 3.2 **Biased Minimized Comparison Algorithm**

One means for the quantification and mitigation of algorithm bias is a Bias-Minimized Comparison Algorithm (BMCA). This is a secondary algorithm in which careful scrutiny excludes any factors which might contribute to unwanted bias. While these models will generally not perform as well as production algorithms containing more factors, the disparate impact between the BMCA and the production algorithm will quantify the amount of bias in the factors added in the production algorithm. This can be used to evaluate variations in a proposed algorithm to inform the final selection of features.

The BMCA method is especially useful for wide datasets which can be subject to the “Spaghetti Problem”. Where it can be difficult to screen hundreds of potential predictors, use of a BMCA will help quantify the overall amount and impact bias introduced by a group of potential features.

The BMCA methodology is applicable in cases of intrinsic bias, where some bias naturally exists in a model. For example, age may be a real, non-prejudicial factor to some degree in a model but should not be overstated. A Bias-Minimized Comparison Algorithm can estimate the amount of intrinsic bias in a population and minimize the impact.

### 3.3 **Open Source Packages for Measuring and Mitigating Bias**

Recent developments in data science have produced a number of open source packages facilitating the measurement of bias in machine learning algorithms and AI. Prominent among these are AI Fairness 360 from IBM ([16] Bellamy et al. 2019) and Fairlearn from Microsoft. Fairlearn is an open-source Python library developed by Microsoft Research that seeks to enable the development of machine learning models that are fair and transparent ([17] Owen 2022). The library provides a suite of algorithms and metrics for evaluating and mitigating bias in machine learning models, and it is designed to work seamlessly with popular machine learning frameworks such as scikit-learn and PyTorch.

Fairlean focuses on calculation and comparison of model Accuracy (correct prediction rate) and Sensitivity (true positive rate). It supports calculation of confusion matrices (Figure 2) to display bias characteristics. The Fairlearn library also includes visualization tools for comparing different model versions to optimize accuracy while minimizing bias. One especially useful plot compares performance and bias characteristics for different versions of a model concept (Figure 3). This allows model developers to test modeling methods, hypertuning the model, and test individual fields for potential impact from bias. Use of this plot supports decisions to maximize model accuracy while minimizing bias,

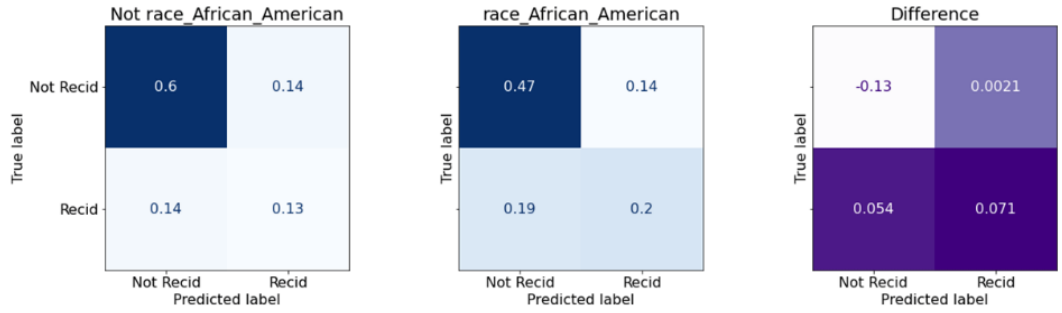


**Fig. 3.** Confusion Matrices for African-American defendants vs. others, with difference, for a Fairlearn-adjusted model. From Owen, S 2022.

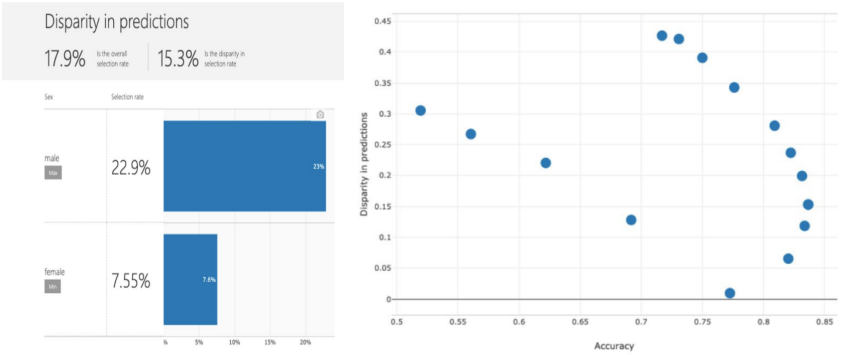


**Fig. 4.** Fairlean: Disparity in predictors and a visual comparison of different model versions to facilitate maintaining algorithm accuracy while minimizing bias. On the right, different model versions are compared to maximizing accuracy while minimizing bias. From Owen 2022.

## 4 Summary

Accurate measurement of bias in machine learning and AI algorithms is critical to the success of these algorithms. Measurement can be accomplished using log odds or odds ratio comparison of model outcomes for population subsets. Log odds is more often used in statistical contexts, while comparison using odds ratios is often better understood and hence more effective with non-technical audiences found in business, policy discussions, and the general public. The new, open-source python package Fairlearn shows great promise for quantifying and mitigating the impact of bias in machine learning and AI.

Best practices for minimizing bias include:

- Parsimonious Models
- Screen all predictors for bias
- Transparent Methods, not Black Box
- Develop the model using new outcomes screened for bias - not past decisions
- Test for disparate impact on at-risk groups using relative risk or odds ratios
- Open Source the data and algorithm

## Acknowledgements

Thanks are due to Nancy Brucken and Brandy Sinco for several helpful comments on measurements methods and presentation of the Disparate Impact paper. Thanks are due to Lance Hielbrun, Harvey Qu, and Karry Roberts for comments on work in ethical preobl;ems in analytics presented to th8e Detroit Chapter of the American Statistical Association and especially to Lance Heilbrun for the

recommendation of relative risk as a preferred metric for evaluating disparate impact.

## References


1. Kirkpatrick, K., "It's not the algorithm, it's the data". Communications of the ACM. 60 (2): 21–23 (2017)
2. Angwin, J., Larson, J.; "Machine Bias". ProPublica (2016)
3. Larson J., Mattu S., Kirchner L., Angwin J.; *How We Analyzed the COMPAS Recidivism Algorithm,* ProPublica (2016)
4. Thomas, C., Pontón-Núñez, A.; " Automating Judicial Discretion: How Algorithmic Risk Assessments in Pretrial Adjudications Violate Equal Protection. Minnesota Journal of Law & Inequality. 40 (2): 5 (2022).
5. Dastin, J., "*Amazon scraps secret AI recruiting tool that showed bias against women*", Reuters, October 2018
6. Nickerson, R. S., "*The Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*", Review of General Psychology (1998)
7. Lewandowsky, S., Cook, J.; "*Why People Don't Believe in Climate Change*", Scientific American (2013)
8. Lavrakas, P., "*Sampling Bias and Data Quality*", Public Opinion Quarterly (1993)
9. Groves, R., Couper, M., Lepkowski, J., Singer, E, Tourangeau, R.; "*Nonresponse Bias in Household Surveys*", Handbook of Survey Research (2004)
10. Najibi, A., "*Racial Discrimination in Face Recognition Technology*", Gender Shades Project, Harvard (2020)
11. Emspak, J., "*How a Machine Learns Prejudice*", Scientific American (2016)
12. Rodriguez-Villa, F., "*Ethical AI: Mitigating Bias*", AdeptAI, (2022) https://www.adept-id.com/news/ethical-ai22022, last accessed 2023/1/15
13. Corliss, D. J.; 2021), "*Disproportional Impact of COVID-19 on Marginalized Communities*", Proc. SAS Global Forum 2021 https://communities.sas.com/t5/SAS-Global-Forum-Proceedings/Disproportional-Impact-of-COVID-19-on-Marginalized-Communities/ta-p/726372
14. Heilbrun, L., personal communication, January 2023
15. Saloner B, Parish K, Ward JA, DiLaura G, Dolovich S. COVID-19 Cases and Deaths in Federal and State Prisons. JAMA. 2020;324(6):602–603. doi:10.1001/jama.2020.12528
16. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. and Nagar, S., 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), pp.4-1.
17. Owen, S., "*Mitigating Bias in Machine Learning With SHAP and Fairlearn*", Databricks (2022) https://www.databricks.com/blog/2022/09/16/mitigating-bias-machine-learning-shap-and-fairlearn.html