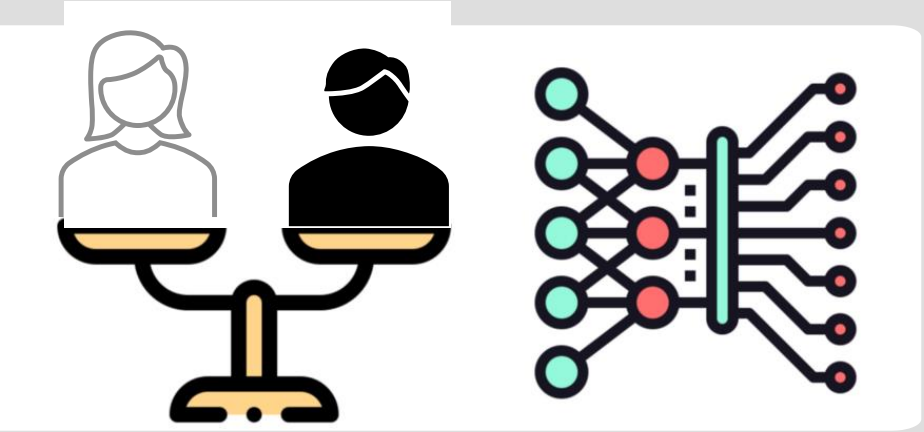
The Alan Turing Institute

Fairness of data, algorithms, and models

Dr. AbdulRahman Alsewari



What is Algorithm Fairness?

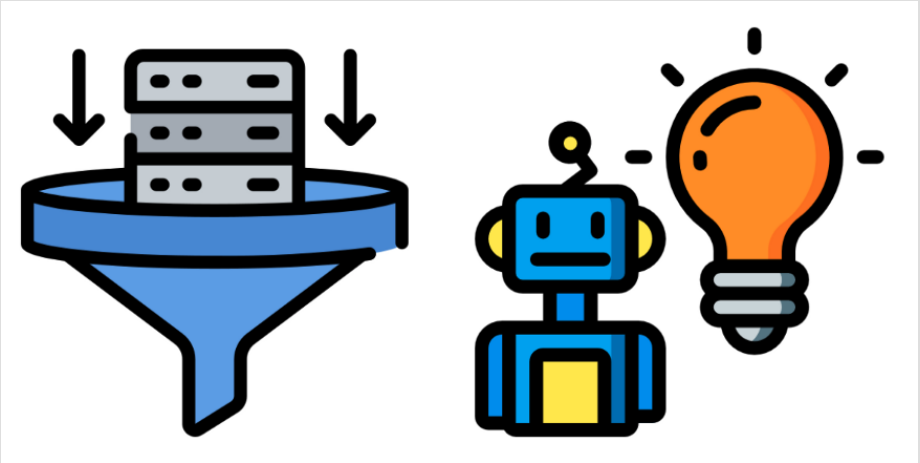
An introduction to the field that aims at understanding and preventing unfairness in machine learning

Pessach, D. and Shmueli, E., (2020), *Algorithmic fairness*. https://arxiv.org/abs/2001.09784

https://towardsdatascience.com/what-is-algorithm-fairness

What is algorithm fairness?

In machine learning, the terms algorithm and model are used interchangeably. To be precise, algorithms are mathematical functions like linear regression, random forests or neural networks. Models are algorithms that have been trained on data. Once trained, a model is used to make predictions which can help automate decisions. These decisions can include anything from diagnosing a patient with cancer to accepting mortgage applications.



Algorithm fairness

- Is the field of research aimed at understanding and correcting biases like these. It is at the intersection of machine learning and ethics.
  - Specifically, the field includes:
    - Researching the causes of bias in data and algorithms
    - Defining and applying measurements of fairness
    - Developing data collection and modelling methodologies aimed at creating fair algorithms
    - Providing advice to governments/corporates on how to regulate machine learning

GO beyond Data

– It is also important to understand that approaches to fairness are not only quantitative. This is because the reasons for unfairness go beyond data and algorithms. The research will also involve understanding and addressing the root cause of unfairness.

Why is algorithm fairness important?

- As mentioned, machine learning models are being used to make important decisions. The consequences of incorrect predictions could be devastating for an individual. If the incorrect predictions are systematic then entire groups could suffer. To understand what we mean by this, it will help to go over a few examples such as:
  - Apple recently launched a credit card — Apple Card
  - Amazon automate recruitment system.
  - COMPAS, an algorithm used by the American criminal justice system.

Why is algorithm fairness important?

- Apple recently launched a credit card — Apple Card
- You can apply for the card online and you are automatically given a credit limit. As people started to use this product, it was found that women were being offered significantly lower credit limits than men. This was even when the women were of a similar financial position (and credit risk). For example, Apple co-founder, Steve Wozniak, said he was offered a credit limit 10 times higher than his wife

Amazon automate recruitment system.

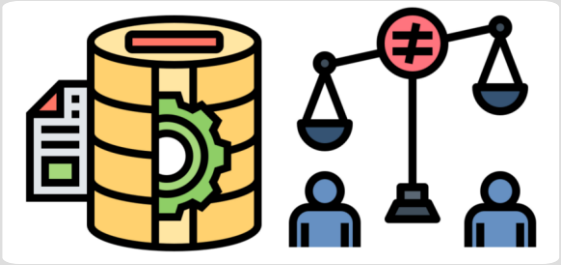
– Another example is a system used by Amazon to help automate recruitment. Machine learning was used to rate the resumes of new candidates. To train the model, Amazon used information from historically successful candidates. The issue is that, due to the male dominance of the tech industry, most of these candidates were male. The result was a model that did not rate resumes in a gender-neutral way. It actually went as far as penalising the word “woman” (e.g. Captain of the **woman’s** soccer team).

COMPAS

– Models can also discriminate based on race. COMPAS was an algorithm used by the American criminal justice system to predict if a defendant was likely to re-offend. An incorrect prediction (i.e. false positive) could result in the defendant being falsely imprisoned or having to face a longer prison sentence. It was found that the false positive rate was twice as likely for black offenders than white offenders. That is black offenders were twice as likely to be **incorrectly** labelled as potential re-offenders.

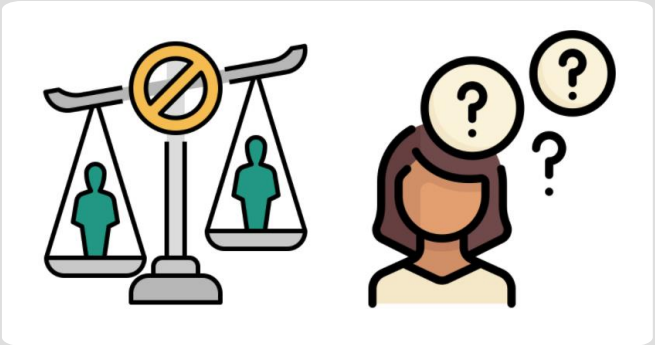
The reasons for unfairness

– Clearly, they are bad but how do we even end up with unfair algorithms? Algorithm fairness is actually a bit of a misleading term. Algorithms, by themselves, are not inherently biased. They are just mathematical functions. By training one of these algorithms on data, we obtain a machine learning model. It is the introduction of biased data that will lead to a biased model. That being said our choices around algorithms can still amplify those biases.



5 Reasons your Model is Making Unfair Predictions

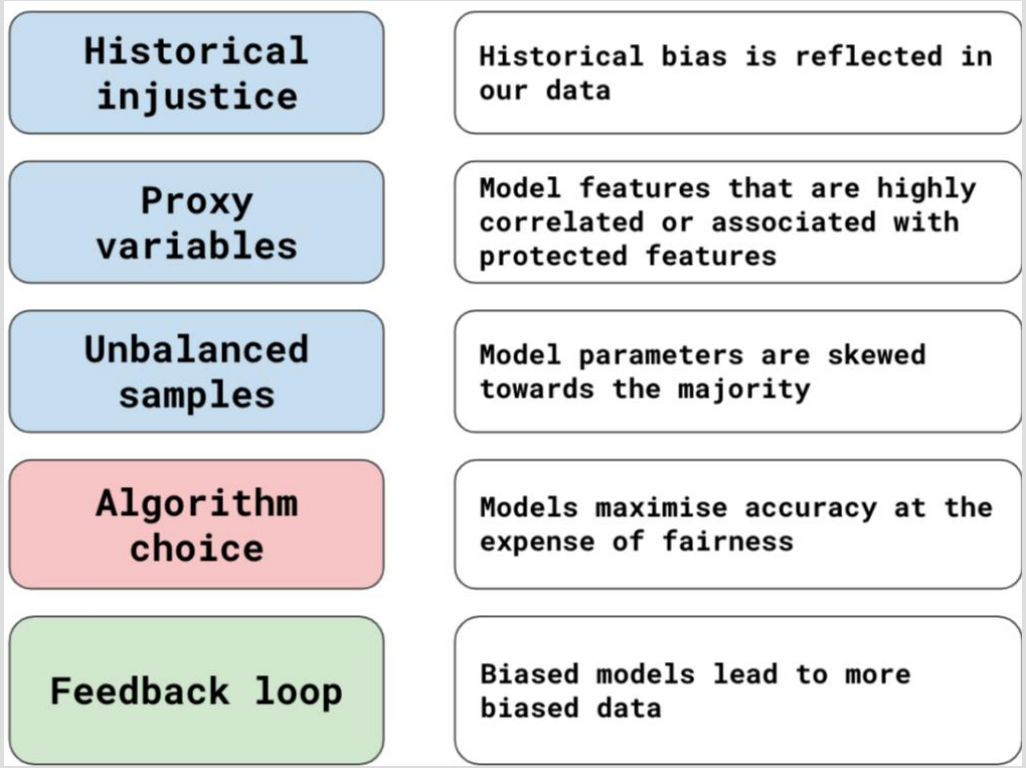
- Common sources of bias — **historical bias, proxy variables, unbalanced datasets, algorithm choices** and **user interaction**



Bias vs Fairness

– Before we dive into that, let’s clarify a few things. “**Bias**” can be a confusing term. It is used across many fields to refer to a range of issues. In ML, bias is any systematic error made during model development. These errors can be a result of incorrect assumptions or mistakes in code. They can be introduced during data collection, feature engineering, or training. Even the way you deploy your model can introduce bias.

5 Reasons your Model is Making Unfair Predictions



5 Reasons your Model is Making Unfair Predictions

Reason 1: historical injustice

In the past, certain groups have been discriminated against. This can happen intentionally through laws or unintentionally through unconscious bias. Whatever the reason, this historical bias can be reflected in our data. Models are designed to fit data. This means any historical bias in data can be reflected in the model itself.

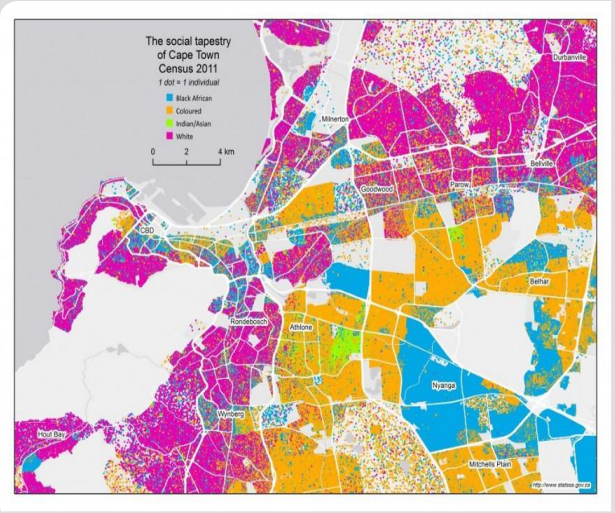
Historical bias can manifest in different ways. Data may be more likely to be missing or recorded incorrectly for minority groups. The target variable may reflect decisions that were unfair. This is more likely when the target variable is based on a subjective human decision. A recent example of this comes from a model developed by Amazon used to help automate recruitment.

5 Reasons your Model is Making Unfair Predictions

Reason 2: proxy variables

– We mentioned **protected variables** above. These are model features that represent sensitive characteristics like race or gender. Using these features can result in an unfair model. Often, normal features can be correlated or associated with protected features. We call these proxy variables. A model that uses a proxy variable can effectively be using a protected variable to make decisions.

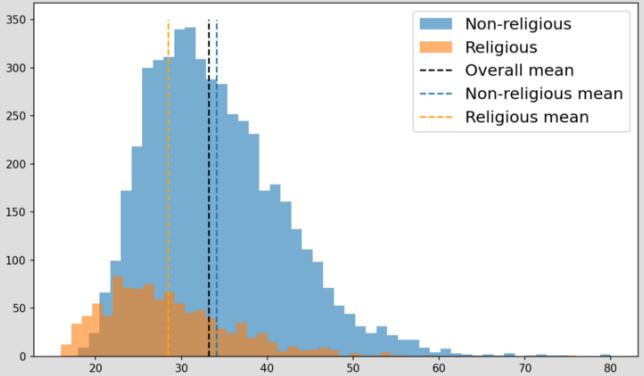
– For example, in South Africa is very predictive of your race. This is due to the country's history of racial segregation. You can see what we mean in the map of Cape Town. The city of Cape Town is still divided on racial lines. This means that using someone's zip code in a model could result in racial discrimination.



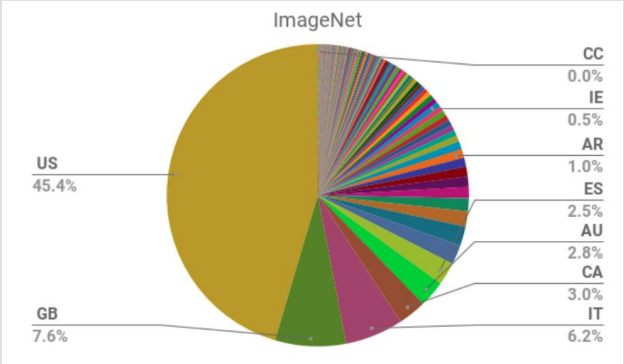
5 Reasons your Model is Making Unfair Predictions

Reason 3: unbalanced samples

- Suppose we calculate the mean age of marriage in a population. We get a value of 33. Looking at Figure, we can see that this value is not representative of all subgroups of the population. The average age of the religious group, 28, is much lower. Due to the size of the non-religious group the average has been skewed. In other words, the population average is much closer to that of the non-religious group.
- We can see an example of a skewed dataset in Pie chart. ImageNet is a large dataset used to train image recognition models.



example of how population average can be skewed (https://towardsdatascience.com/algorithm-fairness-sources-of-bias)



geodiversity of ImageNet dataset (Source: S. Shankar, et. al)

5 Reasons your Model is Making Unfair Predictions

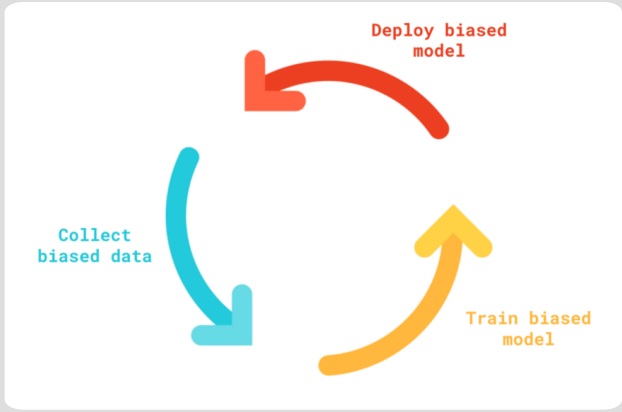
Reason 4: algorithm choice

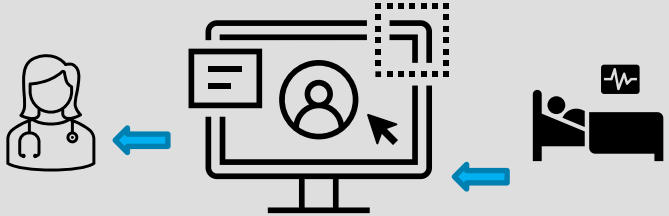
– The previous 3 sources of bias have all involved issues with data. Our choices around the algorithms we use can also contribute to fairness issues. For one, some algorithms are less interpretable than others. This can make it harder to identify the source of bias and correct it. Another major factor is the objective of a model. That is what the model is trained to do.

5 Reasons your Model is Making Unfair Predictions

Reason 5: user feedback loop

– The last source of bias has to do with how we interact with models. Once trained, a model will be deployed. As users interact with the model we will collect more data to train future versions of the model. The way users interact with the model can introduce bias. Existing bias in models can lead to further bias in new data.





We conclude from **5 Reasons your Model is Making Unfair Predictions**

- So, before deploying a model,
  - We need to measure bias, identify its source and correct it.
  - Find where bias comes from.
  - Knowing the source is key.
  - It will help us decide on the best solution to correct it.

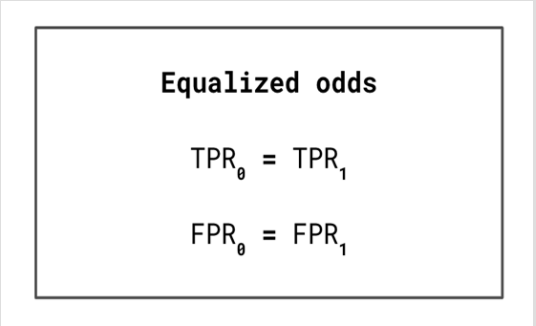
Analysing and measuring unfairness

- Many of algorithm fairness research aims at developing methods to analyse and measure unfairness. This can involve analysing **data** for the potential reasons for unfairness mentioned above. It also involves measuring unfairness in **model predictions.**
- Definitions of fairness.

  –

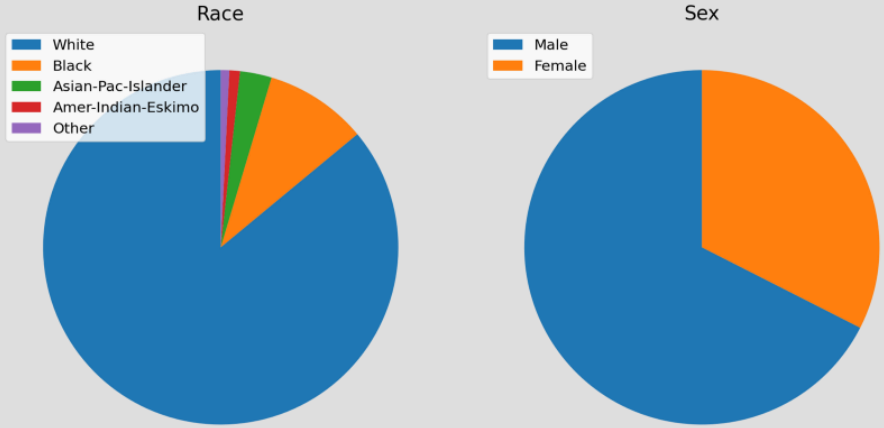
  We can measure fairness in predictions by applying different definitions of fairness. Most of the definitions involve splitting the population into privileged (e.g. male) and unprivileged (e.g. female) groups. We then compare the groups using evaluation metrics. For example, under the **equalized odds** definition we require the true positive rates and false positive rates of the two groups to be equal. A model with significantly different rates is considered unfair. Other definitions include **equal opportunity** and **disparate impact**.

  –



Exploratory fairness analysis

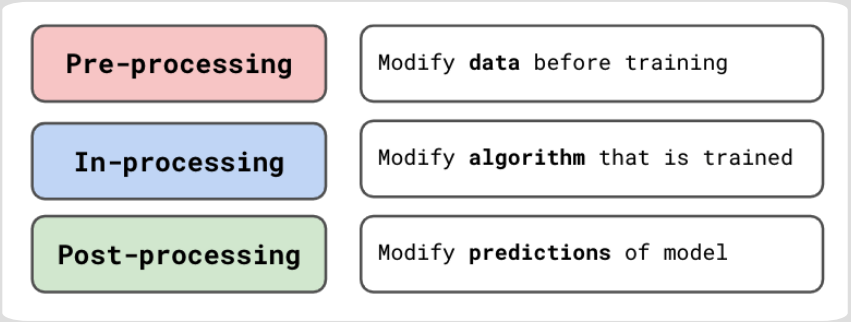
– Assessing fairness does not start when you have your final model. It should also be a part of your exploratory analysis. In general, we do this to build some intuition around our dataset. So, when it comes to modelling, you have a good idea of what results to expect. Specifically, for fairness, you want to understand what aspects of your data may lead to an unfair model.



https://towardsdatascience.com/analysing-fairness-in-machine-learning-with-python-96a9ab0d0705

Correcting and preventing unfairness

– If we discover that our model is unfair we would naturally want to correct it. Various quantitative approaches have been developed. They can be divide into:



– This depends on what stage during the model development they are applied. For example, we can adjust the cost function of a regression model to consider fairness. This would be considered an in-processing method.

https://towardsdatascience.com/approaches-for-addressing-unfairness-in-machine-learning-a31f9807cf31

Interpretability and fairness

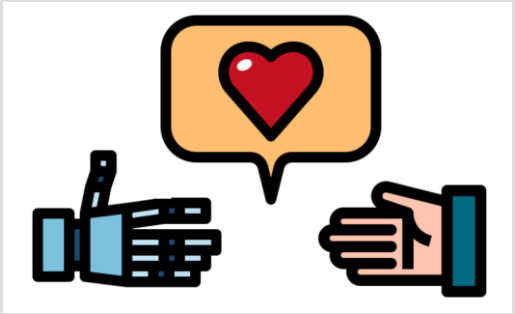
Interpretability and fairness seem to go hand in hand. Interpretability involves understanding how models make predictions. Fairness involves understanding if predictions are biased towards certain groups. These characteristics are consistently mentioned together in responsible AI frameworks and ML conferences. However, interpretability does not necessarily imply fairness.

3 reasons why interpretable models are more likely to be fair

Reason 1: easier to identify the reason for unfairness

Reason 2: easier to correct bias

Reason 3: they're both about building trust



https://towardsdatascience.com/the-relationship-between-interpretability-and-

FAIRNESS-RELATED DATASETS

– In this section, it reviews the most commonly used datasets in the literature of algorithmic fairness.

- ProPublica risk assessment dataset
- Adult income dataset
- German credit dataset
- Ricci promotion dataset
- Mexican poverty dataset
- Diabetes dataset
- Heritage health dataset
- The College Admissions dataset
- The Bank Marketing dataset
- The Loans Default dataset
- The Dutch Census dataset
- The Communities and Crimes dataset

Dataset Name	Domain	# Records	Sensitive Attributes	Target Attributes
ProPublica	Criminal risk assessment	6,167	Race; Gender	Whether an inmate has recidivated (was arrested again) in less than two years after release from prison
Adult	Income	48,842	Age; Gender	Whether an individual earns more or less than 50,000$ per year
German	Credit	1,000	Gender; Age	Whether an individual should receive a good or bad credit risk score
Ricci	Promotion	118	Race	Whether an individual receives a promotion
Mexican poverty	Poverty	183	Young and old families; Urban and rural areas	Poverty level of households
Diabetes	Health	100,000	Race	Whether a patient will be readmitted
Heritage health	Health	147,473	Age	Whether an individual will spend any days in the hospital in the next year
College Admissions	College Admissions	20,000	Gender; Race	Whether a law student will pass the bar exam
Bank Marketing	Marketing	41,188	Age	Whether the client subscribed to a term deposit service
Loans Default	Loans	30,000	Gender	Whether a customer will default on payments
Dutch Census	Census	189,725	Gender	Whether an individual holds a highly prestigious occupation
Communities and Crimes	Crime	1,994	Percentage of African-American population	For each community, the number of violent crimes per 100,000 individuals