Machine Learning Fairness Primer

CSE 474/574 – Spring 2020

1 - What is fairness?

According to the OED, fairness is “impartial and just treatment or behavior without favoritism or discrimination”. At face value this seems to succinctly capture the spirit of the term, but closer evaluation reveals an endless supply of ambiguity. For example, what qualifies as “just” treatment? The answer will obviously vary from field to field, but it is just as likely to vary from person to person as well. The definition assumes some objective standard by which to base our judgements and behaviors, but in practice this standard is ultimately supplied by inescapably subjective agents, each of whom have their own desires, values, and assessments of reasonable behavior.

The information revolution has led to a rise in “technologist” solutions, which propose outsourcing these questions to computational machines. These machines are cold, unfeeling, and rely purely on logic to reach their conclusions and decisions. Thus freed from the inevitable biases plaguing their human creators, the zeitgeist of the late 20th and early 21st century was that machines, and particularly applications in machine learning, would quickly be able to solve all of our societal problems in a way that was just and fair.

Machine learning algorithms generally function through the extraction and manipulation of mathematical encodings of patterns present within a set of data. Often these patterns are too subtle or too obscure to be recognizable by humans. From this viewpoint, and with the clarity of retrospect, it’s quite easy to see why the technologist answer falls short of the desired goal. Bias and discrimination are patterns of behavior, and even when not overtly expressed they are still deeply interwoven into the operation of human society. It follows that data acquired from measurements of human interactions retain these patterns, and machine learning algorithms trained on these data end up displaying the same disparate judgements.

As machine learning engineers and data scientists we desire for our creations to be better than us, for them to reflect the society we want as opposed to the society we live in. This in turn further complicates our definition of fairness into something hypocritical; we are attempting to derive justice from ourselves, who have proven to be an inherently unjust source. Various measures have been proposed to achieve these ends, and they have met with varying measures of success and acknowledgement from the legal and ethical communities. Ultimately, however, we have been unable to concretely define this notion in a way that is applicable and measurable across different contexts and situations.

The best we have done is to define proxy values - a slew of mathematical measurements and statistical analyses that can show specific biases in decision making. As with many engineering tasks, we’ve successfully broken the problem into multiple smaller, more approachable tasks. The debate now seems to center on which of these proxies are most important to fair decision making. We will see that some solutions satisfy certain conditions while entirely ignoring others. Which of these can be called “fair”? The answer is still unclear, and subject to intense scrutiny. More reasonable questions are “what kinds of bias are inherent in data?”, “how can we measure bias?” and “how can we program machines to make ethical decisions?”. The remainder of this document explores some of the existing answers to these questions and presents their respective merits and limitations.

2 - How can we measure bias?

Before we can get into the different statistical measurements it's important to introduce the notation that we'll be using. For the purposes of this project we're exploring bias in supervised machine learning problems. These problems all consist of:

- A dataset **D**, which consists of n samples of feature vector **X**
- A label value for each data point, **Y**
- A prediction value for each data point, **Ŷ**

Additionally, in the measurement of bias we'll also be considering one or more *protected attributes* (also known as *sensitive features*), represented by **A**. In real-world problems **A** will often be a feature that carries legal ramifications for decisions based around it – for example, race, age, gender, or sexual orientation.

For the purposes of explanation, we'll consider a toy example of loan approval scores. In many actual financial situations loans are approved or denied by running an applicant's data through an automated software system. This system uses various metrics to predict whether the applicant will successfully pay back the loan or default on their payments. We will consider two affiliations, pinks and purples, as our protected attribute. Thus in this problem our above notations represent:

- **D** – the set of all loan applicants
- **X** – a vector of features describing each applicant, of the form {*age, credit_score, affiliation}*
- **Y** – a label corresponding to whether the applicant defaulted on their loan or not. **Y**=0 represents a default, while **Y**=1 represents a successful repayment.
- **Ŷ** – our model's prediction of the default status of each applicant, derived from **X**. Values follow the same scheme as **Y**
- **A** – the value of the protected attribute for each applicant. In this case either **A** = pink or **A** = purple

We now go on to explain the different statistical metrics we'll be using to understand bias in our model. It's worth noting that each of these metrics is applicable to any arbitrary grouping of the data. So we could determine the true positive rate of the total population, the true positive rate of applicants with pink affiliations, or any other combination.

For the sake of measuring bias, it is most useful to compare these metrics as measured across different groups of the protected attribute. For example, if we compare the true positive rates of pink affiliated people and purple affiliated people, a significant difference could signify bias in our model.

Note that these metrics can be separated into two broad categories: those that are conditioned on label values, and those that are conditioned on prediction values. The question of which set of metrics should be more strongly considered when making "fair" decisions is hotly contested, as demonstrated in the varying opinions available in the supplementary material.

True Positives

The number of true positives represent the amount of applicants in a group who were predicted to successfully repay their loan and actually did so. In general terms, this is the number of entities of the group with $\hat{Y}=1$ and $Y=1$.

True Negatives

The number of true negatives represents the amount of applicants in a group who were predicted to default and actually did default. In general terms, this is the number of entities of the group with $\hat{Y}=0$ and $Y=0$.

False Positives

The number of false positives represents the amount of applicants in a group who were predicted to repay their loan when they actually defaulted. In general terms, this is the number of entities of the group with $\hat{Y}=1$ and $Y=0$.

False Negatives

The number of false negatives represents the amount of applicants in a group who were predicted to default when they actually repaid their loan. In general terms, this is the number of entities of the group with $\hat{Y}=0$ and $Y=1$.

False Positive Rate (FPR) and True Negative Rate (TNR)

The false positive rate (FPR) is the number of false positives in the group divided by the total number of applicants who defaulted within the group. It can also be expressed as the probability of a prediction of “repaid” given a label of “default”.

$$\begin{aligned} \text{FPR} &= \text{False Positives} / \text{Labelled Negatives} \\ &= \text{False Positives} / (\text{True Negatives} + \text{False Positives}) \\ &= P\,(\hat{Y}=1 \mid Y=0) \end{aligned}$$

Or, if considering a group defined by membership in a protected class *a*:

$$= P\,(\hat{Y}=1 \mid Y=0, A=a)$$

The true negative rate, also known as the *specificity*, is the complement of the false positive rate and can be computed easily by:

$$\text{TNR} = 1 – \text{FPR}$$

$= P(\hat{Y}=0 \mid Y=0)$

False Negative Rate (FNR) and True Positive Rate (TPR)

The false negative rate (FNR) is the number of false negatives in the group divided by the total number of applicants who repaid within the group. It can also be expressed as the probability of a prediction of “default” given a label of “repaid”.

FNR = False Negatives / Labelled Positives

= False Negatives / (True Positives + False Negatives)

$= P(\hat{Y}=0 \mid Y=1)$

Or, if considering a group defined by membership in a protected class *a*:

$= P(\hat{Y}=0 \mid Y=1, A=a)$

The true positive rate (TPR), also known as the *sensitivity* or *recall*, is the complement of the false positive rate, and can be computed easily by:

TPR = 1 – FNR

$= P(\hat{Y}=1 \mid Y=1)$

Positive Predictive Value (PPV)

The positive predictive value (PPV), also known as the *precision*, is the number of true positives in the group divided by the total number of applicants predicted to repay within the group. It can also be expressed as the probability of a label of “repaid” given a prediction of “repaid”.

PPV = True Positives / Predicted Positives

= True Positives / (True Positives + False Positives)

$= P(Y=1 \mid \hat{Y}=1)$

Or, if considering a group defined by membership in a protected class *a*:

$= P(Y=1 \mid \hat{Y}=1, A=a)$

Negative Predictive Value (NPV)

The negative predictive value (NPV) is the number of true negatives in the group divided by the total number of applicants predicted to default within the group. It can also be expressed as the probability of a label of “default” given a prediction of “default”.

PPV = True Negatives / Predicted Negatives

= True Negatives / (True Negatives + False Negatives)

$= P\ (Y=0 \mid \hat{Y}=0)$

Or, if considering a group defined by membership in a protected class *a*:

$= P\ (Y=0 \mid \hat{Y}=0, A=a)$

False Omission Rate (FOR)

The false omission rate (FOR) is the number of false negatives in the group divided by the total number of applicants predicted to default within the group. It can also be expressed as the probability of a label of “repaid” given a prediction of “default”.

FOR = False Negatives / Predicted Negatives

= False Negatives / (True Negatives + False Negatives)

$= P\ (Y=1 \mid \hat{Y}=0)$

Or, if considering a group defined by membership in a protected class *a*:

$= P\ (Y=1 \mid \hat{Y}=0, A=a)$

False Discovery Rate (FDR)

The false discovery rate (FDR) is the number of false positives in the group divided by the total number of applicants predicted to repay within the group. It can also be expressed as the probability of a label of “default” given a prediction of “repaid”.

FDR = False Positives / Predicted Positives

= False Positives / (True Positives + False Positives)

$= P\ (Y=0 \mid \hat{Y}=1)$

Or, if considering a group defined by membership in a protected class *a*:

$= P\ (Y=0 \mid \hat{Y}=1, A=a)$

Confusion Matrix

All of these statistical metrics can be combined into a matrix, known as the *confusion matrix* or *error matrix.* Ironically, this matrix can often provide more intuitive understandings of what the values actually reflect in the data.

	Actual – Positive	Actual - Negative
Predicted – Positive	True Positive (TP) True Positive Rate (TPR) Positive Predictive Value (PPV)	False Positive (FP) False Positive Rate (FPR) False Discovery Rate (FDR)
Predicted – Negative	False Negative (FN) False Negative Rate (FNR) False Omission Rate (FOR)	True Negative (TN) True Negative Rate (TNR) Negative Predictive Value (NPV)

F-score / F1 score / F-measure

The F-score, also known as the *harmonic mean,* is a measurement of a binary classification's accuracy. It relies on the precision and recall of the results, and is given by the following formula:

$$F_1 = 2\frac{precision * recall}{precision + recall}$$

In our example, imagine if 70% of applicants in our dataset had defaulted, and 30% had not. Using a traditional accuracy measurement would not reflect this disparity, and could be misleading. If our model was simply predicting that every applicant would default it would report an accuracy of 70%, which is obviously not representative of the task at hand. Using the F-score can provide a more intuitive measurement in cases like this.

ROC Curves

A classifier's ROC (Receiver operating characteristic) curve plots the TPR against the FPR at various threshold settings. It can be used to determine an optimal threshold, and thus an optimal model, independently of the cost context and class distribution. The area under the curve (AUC) is often used as a measure of a model's overall performance.

3 - Evaluation Methods and Fairness Constraints

There are many different constraints, proposed by various sources, that can be enforced to ensure certain methods of “fairness”. One important point to note is that for all of these cases the optimal constrained solution differs from the optimal unconstrained solution, as proven in https://arxiv.org/pdf/1701.08230.pdf. This implies that these measures of fairness come at the cost of accuracy, and marks an important tradeoff to be considered when designing machine learning systems.

These techniques can be split into many different categorizations. For our purposes we have chosen to separate them based on where they are applied in the data stream. Note that this list is not exhaustive – instead, it contains what we consider to the be the most relevant for the problem at hand. Also note that many of these techniques go by various different names. We have tried to include the alternatives when possible.

All of the postprocessing techniques used in PA3 rely on thresholding, allowing them to be applied to any model that outputs a binary decision as a real valued number between 0 and 1. Any value above or equal to the threshold is considered a 1, and any value below the threshold is considered a 0. Different threshold values for different protected groups can be applied in order to enforce various fairness metrics.

An Important Note:

The pros and cons listed alongside some of these methods are neither definitive or exhaustive. They reflect the author’s own observations, and are mostly oversimplified for the sake of brevity. The supplementary readings will provide considerably more insight about the reported benefits and shortcomings of these methods, and should be consulted before reaching any conclusions.

Preprocessing - Fairness through Unawareness (Blindness)

Fairness through unawareness is achieved by not including the protected class in the feature vector during training. In our example, we would not include *affiliation* in the feature vector **X***,* instead training our model on only age and credit score.

> ***Pros*** – at face value, the sensitive attribute is not impacting the results.
>
> ***Cons*** – models can still discriminate with regards to the protected class through the use of redundant encodings. For this example, if age has any correlation with affiliation then the model can still implicitly learn about the sensitive attribute.

Preprocessing - Causal discrimination

A classifier satisfies this definition if it produces the same classification for any two subjects with the exact same attributes (besides the sensitive attribute). In our example this implies that a pink applicant, aged 35, with a 720 credit score should be classified in the exact same way as a purple applicant, also aged 35 and with a 720 credit score.

In theory, this can be achieved by altering the training data set in the following way:

For each member of a protected class, add another entry with all of the same information but with a different value for the protected attribute. Add a label for this new data point, with the same value as the label for the original data point.

> ***Pros*** – Ostensibly removes biases from the data before training, allowing the decision to be made entirely independent of the sensitive attribute.
>
> ***Cons*** – In practice, this definition is very rarely satisfied by the final classifier produced by this method. It entirely neglects additional patterns and encodings that can influence predictions.

Training algorithm modification techniques

While training time modification techniques exist, as of yet they are not as popular as the other methods. One such technique involves incorporating fairness criteria into logarithmic loss minimization in the form of a minimax game. More info can be found from the source: https://arxiv.org/pdf/1903.03910.pdf

Postprocessing - Single Threshold

Single threshold is the simplest post-processing technique, and is usually the default method for thresholding real-valued data. It consists in choosing a single threshold and applying it to all of the classifications.

> ***Pros*** – No loss to accuracy due to additional constraints, and simplest to implement. All groups are held the same decision "standard", independent of sensitive attributes.
>
> ***Cons*** – No considerations towards statistical measurements of fairness or underlying biases within the data.

Postprocessing - Equal opportunity

A classifier satisfies this definition if all sensitive groups have equal true positive rates. This method strives to ensure non-discrimination over only the "privileged" outcome. In our example, this means that all people who would actually pay back their loan should have an equal opportunity to actually get a loan in the first place. This condition can also be visualized as a horizontal line spanning across the ROC curves of all protected attributes.

This can be achieved by choosing different thresholds for each group, allowing it to be used as a post-processing method. In statistical terms, equal opportunity means that the probability of a positive prediction given an actual positive label and any value *a* should be equal for all possible values of *a.* In our binary problem, this can be simplified to:

$$P\ (\hat{Y}=1 \mid Y=1, A=0) == P\ (\hat{Y}=1 \mid Y=1, A=1)$$

Pros – forces the designer of the model to create a good classifier for every sensitive group. In effect, this transfers the burden of uncertainty from the protected class to the decision maker.

Cons – can slightly reduce overall accuracy, as no predictor can have a higher TPR than the worst performer of the lot.

Postprocessing - Predictive Parity

A classifier satisfies this definition if all sensitive groups have equal PPV, which mathematically implies that such a classifier will also have equal FDRs. In other words, predictive parity requires that correct positive predictions be independent of the protected attribute. In our example enforcing predictive parity ensures that the fraction of correct positive predictions will be the same for both affiliations.

Again, this can be achieved by picking different thresholds for each sensitive group. In statistical terms, the resulting classifier must satisfy the following equation:

$$P(Y=1 \mid \hat{Y}=1, A=0) == P(Y=1 \mid \hat{Y}=1, A=1)$$

Pros – Reaches an equitable percentage of correct positive predictions, enforcing a sense of accuracy without significantly reducing overall performance.

Cons – Entirely ignores all measurements conditioned on label values, such as true positive rate, false positive rate, true negative rate, and false negative rate.

Postprocessing - Demographic Parity (disparate impact)

This notion of fairness requires that a positive prediction be independent of the protected attribute. This can also be interpreted as all sensitive groups having equal predicted positive rates. Using our previous notation:

$$P\ (\hat{Y}=1 \mid A=a) \text{ must be the same for all } a$$

This resulting probability can be set to any value. For our example choosing values that are too high or too low can exasperate the disparate impact towards or against certain groups. A good starting value is the probability that any given applicant will actually be approved, P (Y=1).

In practice, it can be quite difficult to get the probabilities to line up exactly. In practice it is common to introduce a value epsilon, which denotes the allowed tolerance in the disparity of positive prediction rates across sensitive groups.

We can also rearrange this definition to get a single numerical value, known as the *disparate impact*, that gives a fairness ratio based on predicted probabilities. This requires recognizing one class as a “privileged class” or “reference class” and comparing all other classes against it. This can either be done in a binary style (by grouping all non-reference classes into one), or by taking the average of each non-reference class individually compared to the reference class.

$$DI = P\ (\hat{Y}=1 \mid A\neq 1)\ /\ P\ (\hat{Y}=1 \mid A=1)$$

Where A=1 designates the reference class

Pros – Tends to reduce variance in false positive rates and false negative rates across different classes. If all ground truths were equal (all affiliations had equal probabilities to default on a loan), this could create an optimally fair classifier

Cons – can end up selecting qualified applicants from one demographic but unqualified applicants from another, as long as the percentage of acceptances match. If the target value Y is even slightly correlated with the sensitive attribute A (which in practice, it almost always is), this method completely excludes the ideal predictor, $\hat{Y} = Y$.

Postprocessing - Maximum Profit (or maximum accuracy)

A maximum profit strategy refers to a classifier which maximizes financial gain (or minimizes loss) with no extra considerations towards fairness. This obviously requires that the financial gains and losses of correct and incorrect classifications are known. With no financial information, this method becomes maximum accuracy.

When considering maximum accuracy this method seeks to find thresholds that correctly classify as many data points as possible. When financial information is considered this becomes a little more complex, as true positives, true negatives, false positives and false negatives all may carry differ weights towards the overall financial performance.

Pros – As suggested, accuracy (or profit) is maximized

Cons – This method operates independently of any additional notions of fairness. In practice, it often tends to actually exasperate existing disparities in the data.

4 – Two Opposing Ethical Frameworks

In the study of practical ethics two main frameworks dominate the field; utilitarianism and deontology. These ideas both have their own proponents and detractors, and each of them can provide an interesting approach towards potential applications in machine learning.

This section will provide very simple summaries of each framework, as well as show a practical example of how they might be applied in a machine learning problem.

4.1 – Utilitarianism

Utilitarianism was first developed by English philosopher Jeremy Bentham, and then later refined by fellow English philosopher John Stuart Mill. The central tenet of utilitarianism is that decisions should be made based on the amount of overall happiness or benefit they provide. The ethical decision in any dilemma is one that provides the greatest *utility* for the greatest number, hence the name.

One benefit of utilitarianism is that values the happiness of all humans equally. No one is privileged when the goal is to maximize total utility. However, there is a big issue in the quantification of utility; who's to say where the greatest benefit lies? If you say you'll enjoy my property twice as much as me, does that mean I'm ethically obligated to give my things to you? Obviously the subjectivity of human experience makes this a difficult practical question to answer. For our purposes we'll only consider standard utilitarianism, in which all agents gain or lose the same amount of utility under the same circumstances.

Utilitarianism as a human decision making strategy was famously explored in depth through Phillipa Foot's "trolley problems" (https://en.wikipedia.org/wiki/Trolley_problem). In the basic formulation, a runaway trolley is hurtling down the track, and if nothing is done it will kill 5 workmen. However, you are standing next to a lever that, if pulled, can divert the trolley onto a different track in which only 1 workman will die.

The majority of people presented with this problem tend to choose to kill the 1 man to save the 5, which is the utilitarian answer. However, countless variations have been proposed that change the way the answers fall. What if the trolley could only be stopped by pushing a fat person in front of it? What if the 1 person on the alternate track was your grandmother? In these variations people tend to overwhelmingly choose the non-utilitarian answer, showing that it certainly isn't a universal human approach to human decision making.

Variations of the trolley problem have also been applied to autonomous cars, which will be explored in section 4.3

4.2 – Deontology

Deontology was a philosophy developed and championed by Immanuel Kant. The name stems from the Greek *deon*, meaning duty or obligation. Similarly, deontology concerns itself with ethical decisions as intrinsic moral duties. Kant provides a long list of what qualifies something as a duty, but the central notion is known as the categorical imperative.

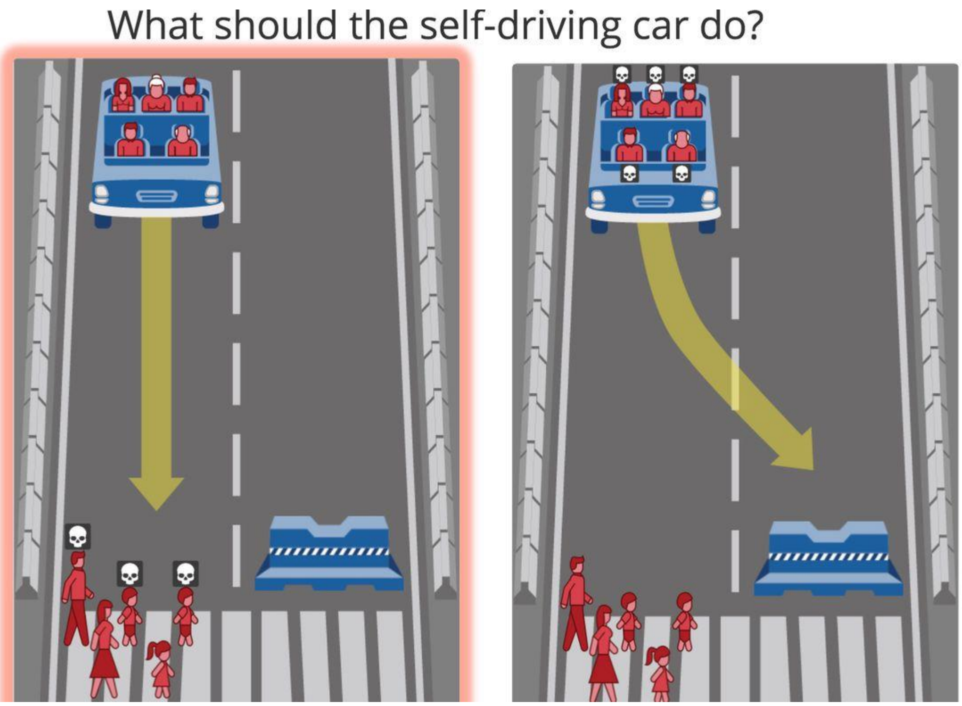
The categorical imperative states that a proper moral law must be unconditional and absolute for all agents, regardless of their intents or motives. Variations of this idea often show up in traditional teachings under names such as “the golden rule”.

A downside of this approach stems from this doctrine of universality. A famous example, recognized by Kant himself, is known as “the inquiring murderer”. (http://www.cabrillo.edu/~cclose/docs/Kant_Inquiring%20Murderer.pdf) There are many variations, but the main idea is this; you are at home with a friend when a man comes to your door. He asks if your friend is in your home – when you ask why, he says he intends to murder him/her. Under deontology you are not allowed to lie to the man, since lying goes against the categorical imperative. So you must say that your friend is indeed in your home, and thus allow them to be murdered.

There have been plenty of accounts and examples arguing over this situation, but at the very least it shows that the notion of unwavering duties prescribed by deontology also has flaws.

4.3 – Ethical Frameworks Applied to Autonomous Vehicles.

Let's imagine a variation of the trolley problem.



To maintain similarity, let's assume that there is only one person in the car. The autonomous vehicle has two options – crash into the barrier, and kill the driver, or crash into the 5 pedestrians and kill them instead. Which is the better option?

Utilitarianism is clear-cut – the car should act in such a way that maximizes total utility. In this situation this means minimizing the number of deaths, so the car should crash into the barrier. However, this does introduce some uncertainty into the equation – what if there were 5 people in the vehicle and one in the crosswalk? The decision made by the vehicle would be different and thus unpredictable in order to ensure a similar numerical outcome. Less people will die in this single situation, but stakeholders will be unable to properly plan for risk in the future because the decision will always be variable

Deontology is a bit more difficult a first glance, because one must determine where the duty lies. Do autonomous vehicle manufacturers have a duty to ensure the safety of the people who purchase and operate their vehicles? One might argue this point, especially since it holds many similarities to existing laws about vehicle safety features. On the other hand, do vehicle manufacturers have a duty to protect the welfare of the general public who may be exposed to their products?

Either one of these can be a valid argument, but the real point is that regardless of which is chosen the decision will be absolute. From the time that the duty is determined all parties will be aware that in similar situations a vehicle will either always crash and kill the driver or always keep driving and kill the pedestrians. This static, universal decision making removes uncertainty and allows anybody involved to make plans to mitigate their risk.

5 – Common Types of Statistical and Cognitive Biases in Data Analysis

It is not an unreasonable precaution to approach all data sets as containing some sort of bias. Data is collected by humans or human-built machines, and analyzed similarly. The underlying data points are created by situations and decisions mired in human cognitive biases and heuristics, ultimately giving rise to data with a complex interplay of multiple sources of noise, error, and bias.

This section lists a number of common types of statistical and cognitive biases present in data acquisition and analysis. Note that this list is nowhere near exhaustive, and some terminology/classifications are used differently in different sources. The information in this document attempts to capture the most common usages and definitions, although there is considerable overlap between many of the categories.

In statistics a data sample is considered biased if it is unrepresentative of the population parameter being estimated. This definition will serve as a benchmark for viewing how the following errors can lead to undesirable qualities in data and model performance.

Selection Bias

Selection bias occurs when data points are selected for analysis in some non-random way. There are a few subtypes which further characterize the type of non-random selection and the stage in which the error is introduced.

Sampling Bias

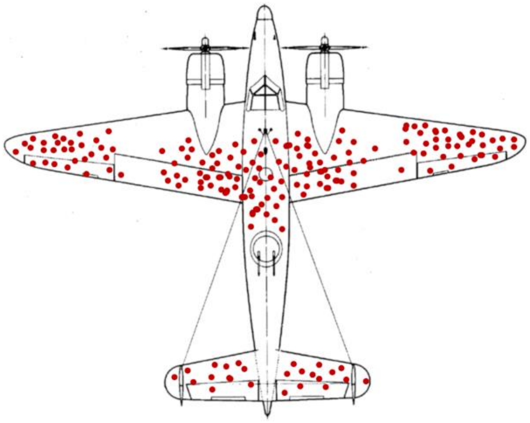
Sampling bias is often classified as a subtype of selection bias, although it is sometimes considered its own type of bias. Sampling bias refers specifically to the methodology used to collect data samples, whereas selection bias more broadly refers to how the samples are selected for analysis.

Sampling bias occurs when a data sample is obtained in some non-random way. This results in some members of the population being more or less likely to be included than the underlying distribution would imply. For example, imagine a population represented by a normal distribution. If you only took samples in the domain lying ±1 standard deviation from the mean you would be committing a sampling bias, and your sample would not accurately represent the total population.

Survivorship Bias

This type of selection bias occurs when only some portion of the data makes it past some critical selection threshold. The rest of the data are not visible, and thus can be mistakenly overlooked during analysis.

Possibly the most famous example of overcoming survivorship bias comes from World War II. Statistician Abraham Wald was tasked with choosing which parts of planes to reinforce to minimize losses.



Given the following distribution of received gunfire, common sense dictated that the most logical places for reinforcement were those that had received the most gunfire. However, Wald cleverly noted that planes which were shot in other places, such as the propellers or engine, did not survive their missions and were thus not properly factored into the data sample. He argued that it was these unrepresented critical places which would benefit the most from reinforcement, undoubtedly saving many lives with his decision.

Base Rate Fallacy / Base Rate Neglect / Base Rate Bias

As discussed during the coverage of Bayesian statistical analysis, the base rate fallacy refers to the tendency of people to ignore general information (such as the prior probability of an event occurring) and instead focus on specific evidence when estimating outcomes.

Conjunction Fallacy

This refers to an estimation error in which people assume that specific conditions are more probable than a single general one. This can be considered a variation of the base rate fallacy, in that the larger prior probability of the general option is overshadowed by additional information.

Here’s an example: you are working in Africa, and you know that the wildlife in Africa can sometimes be dangerous. Given this knowledge, what has a higher chance of occurring?

1. You are attacked by a cat
2. You are attacked by a lion

Many people incorrectly pick option 2, weighing heavily on the description. Lions live in Africa, and lions are known to occasionally be dangerous to humans. However, a lion is a type of cat – the set of lions is a smaller subset of the set of all cats. Option 2 is therefore a conjunction, containing option 1 as a conjunct, and by that definition it must have a smaller likelihood of occurring.

Interestingly, expert statisticians perform no better at conjunction estimation tasks than the average person.

Response Bias

Response bias refers specifically to faulty practices in the surveying and collection of human data via survey, interview or query. This can include dishonest or inaccurate responses, or biases introduced from the survey delivery or environment.

Question Order Bias

This type of bias refers to different reactions based on the ordering of questions. Different inquiries, especially those regarding moral or social norms, can prime respondents to respond differently about subsequent questions.

Framing Effect

The framing effect refers to the tendency of people to respond to questions differently depending on whether the options are presented in a positive or negative frame.

The original study by Tversky and Kahneman asked people to choose between two possible treatment options for 600 people infected with a deadly disease. Participants were split into two groups, each of which received the options both framed in either a positive or negative way.

Framing	Treatment A	Treatment B
Positive	"Saves 200 lives"	"A 33% chance of saving all 600 people, 66% possibility of saving no one."
Negative	"400 people will die"	"A 33% chance that no people will die, 66% probability that all 600 will die."

Respondents chose Treatment A 72% of the time when presented with positive framing, but only 22% percent of the time when presented with negative framing. In general, subsequent studies have found that people are open to assuming more risk when options are framed negatively. On the other hand, people are generally more risk averse when the same options are framed positively.

Confirmation Bias

Confirmation bias refers to a cognitive bias introduced by people during data analysis. If a researcher has some preexisting notion of what patterns they expect to find in their data, they may inadvertently introduce a selection bias by cherry-picking data points that support their conclusion.

Of course, this cherry-picking can happen purposefully too. This is the backbone of pseudoscience – finding data to explicitly *confirm* a hypothesis instead of accurately and robustly *testing* a hypothesis.

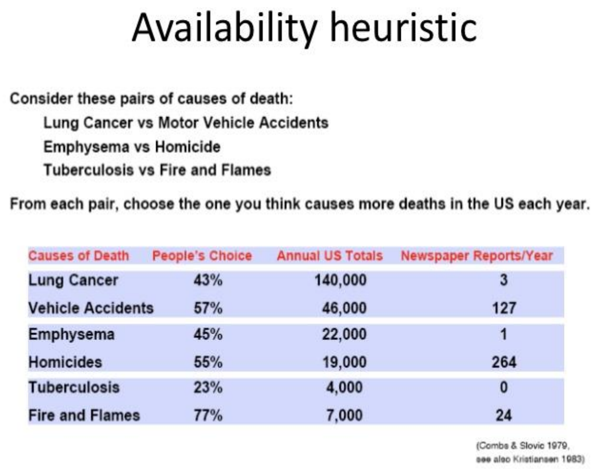
Detection Bias

Detection bias occurs when there is a systematic difference between groups in how an outcome is determined. For example, overweight and obese people are statistically more likely to have type-II diabetes. This knowledge could lead physicians to more readily suggest testing for diabetes in overweight patients, leading to an overrepresentation of that group.

The above example overlaps with selection bias. However, there are some situations where an outcome may be more or less difficult to recognize in certain groups. Depression is a common comorbidity with many other mental illnesses, and because of its outward behavioral changes it is often more easily recognized than other symptoms. This makes it difficult to accurately diagnose more complex afflictions of which depression may just be a single symptom, leading to an underrepresentation of people will those illnesses.

Availability Heuristic/Availability Bias

The availability heuristic is a cognitive bias in which people tend to over-assume the likelihood of an event or consequence based on how easy it is to recall data instances related to that event. A set of very illustrative examples concern people's judgements of the likelihoods of various causes of death. In many ways this is similar to the base rate fallacy – the difference is that instead of ignoring the base rate, people often misjudge it based on readily available instances.



Availability heuristic

Consider these pairs of causes of death:

Lung Cancer vs Motor Vehicle Accidents
Emphysema vs Homicide
Tuberculosis vs Fire and Flames

From each pair, choose the one you think causes more deaths in the US each year.

Causes of Death	People's Choice	Annual US Totals	Newspaper Reports/Year
Lung Cancer	43%	140,000	3
Vehicle Accidents	57%	46,000	127
Emphysema	45%	22,000	1
Homicides	55%	19,000	264
Tuberculosis	23%	4,000	0
Fire and Flames	77%	7,000	24

(Combs & Slovic 1979,
see also Kristiansen 1983)

Social Biases

Social biases, also sometimes referred to as societal biases or attribution errors, refer to any systematic biases that arise from human social interactions. This is a broad class of cognitive and statistical biases that account for many disparities in both human decision making and ML systems.

Prejudice Bias

Prejudice bias refers to the systematic skewing of any data distribution that incorporates data resulting from the application of cultural stereotypes. For example, the American criminal justice system has long been accused of disproportionately targeting African American males. The stereotypes surrounding the group lead to higher arrest rates, which in turn disproportionately increases the criminality base rate. This can in turn increase the believability of the initial stereotypes, creating a self-fulfilling prophecy of bias.

Group Attribution Error

The group attribution error refers to the tendency of people to believe that the characteristics and actions of an individual are representative of a group as a whole. This type of generalization leads to the birth of many stereotypes, which in turn can influence the actions of in-group and out-group members and reinforce the initial bias.

Labeling Bias / Label Bias

The labeling bias refers to expectations that arise of a person, group, or data instance given a particular label. For example, if you were told a man named Tom works as a librarian, your likelihood of assuming certain characteristics about Tom would increase. You may assume he is quiet, educated, bookish, well-organized, and probably wears glasses.

A similar problem is a common occurrence in machine learning models. A system may come to associate a label with a single feature (or a disproportionately small set of features) while ignoring other discriminating characteristics.

Representativeness Heuristic

This heuristic is the reverse of the labeling bias – it refers to an over-assumption of how likely a data point is to be a member of a group by considering how similar that data point is to a typical member of that group.

You are told that Tom is a quiet, educated, bookish, well-organized, glasses-wearing man. You are then asked what Tom's likely occupation is – librarian, or waiter? People are much more likely to assume that Tom is a librarian, entirely ignoring the fact that any given person is statistically much more likely to be a waiter or waitress. (In the U.S. there are an estimated 166,164 librarians, compared to over 2.6 million waiters/waitresses) This is a specific type of base rate fallacy, in which the prior probability is ignored in favor of specific characteristics that seem typical for a member of a certain group.

5 - Additional Resources

The following is a list of research papers and articles that further explore some of the concepts covered in this document. Many differing viewpoints are showcased, often endorsing one method while downplaying others. This wide variety of opinions should be strongly considered before reaching any definitive decisions about what constitutes fairness in a given context.

A comparative study of fairness-enhancing interventions in machine learning:

https://arxiv.org/pdf/1802.04422.pdf

Algorithmic decision making and the cost of fairness:

https://arxiv.org/pdf/1701.08230.pdf

Equality of Opportunity in Supervised Learning:

https://arxiv.org/pdf/1610.02413.pdf

Fairness Definitions Explained:

http://fairware.cs.umass.edu/papers/Verma.pdf

Fairness for Robust Log Loss Classification:

https://arxiv.org/pdf/1903.03910.pdf

“Fair” Risk Assessments: A Precarious Approach for Criminal Justice Reform

https://scholar.harvard.edu/files/bgreen/files/18-fatml.pdf

MIT Moral Machine test lab for autonomous vehicles

http://moralmachine.mit.edu/

Judgement Under Uncertainty – Heuristics and Biases

https://www.socsci.uci.edu/~bskyrms/bio/readings/tversky_k_heuristics_biases.pdf