# *Assignment: Fairess in Machine Learning*

*Miranda Lao*

*Fall 2019*

### Learning Objectives

- Gain some familiarity with the subject of fairness in machine learning.
- Familiarize with confusion matrix terminology.
- Understand the limitations and use cases of commonly-used fairness metrics.
- Consider viewing impossibility theorems surrounding fairness as trolley problems.
- Practice auditing models for bias.

### Notice

The content of this assignment is heavily inspired by a FAT/ML Tutorial given by Arvind Narayanan. Feel free to watch the tutorial (55 minutes) to reinforce the material in this assignment (or is it the other way around?).

## 1 *Motivation and Context*

The dictionary site Merriam-Webster defines fairness as: "fair or impartial treatment : lack of favoritism toward one side or another". When we talk about STEM subjects, there is often an assumed level of objectivity or lack of bias, exemplified by the popular idiom, "the numbers don't lie." However, we've seen in previous discussions (gender and privacy, debate between Propublica and Northpointe surrounding the COMPAS recidivism algorithm) that the outcomes of algorithmic decision making can and do unfairly benefit some people over others. As the use of machine learning and algorithmic decision making becomes increasingly widespread, calls for fairness, accountability, and transparency (including those made by the eponymous organization, FAT/ML) in these systems have also emerged. But what exactly is fairness in machine learning?

There exist many definitions and competing metrics of fairness given by the computer science and statistics communities alone, not to mention discussions and definitions of fairness given in philosophy, law, economics, and game theory. Every definition comes with a context, and appropriate applications for use. This assignment will attempt to give some of the common language used when discussing notions of fairness in machine learning, as well as some of the challenging questions that technologists face when implementing fairness in their systems.

## 2 *Confusion (Matrices) and Impossibility (Theorems)*

Though the title of this section is not-so-encouraging, we'll actually be revisiting ideas that were introduced in a previous assignment. Understanding these ideas is important in understanding some commonly used metrics of fairness in machine learning.

### *Confusion Matrices*

A confusion matrix is a 2x2 table that helps in interpreting the performance of an algorithm, named for how it helps to identify whether or not the system is confusing two classes of objects. Many of the terms used surrounding a confusion matrix may sound familiar; some of them were described in the Module 2 Assignment 3 reading and companion notebook.

The rows of a confusion matrix give the predictions for each of the two classes, while the columns of the confusion matrix give the actual status for each of the same two classes. For example, if an algorithm is predicting between the number of cats and not-cats in a dataset, the confusion matrix would look as follows:

| | **Actual Cat** | **Actual Not-Cat** |
|---|---|---|
| **Predicted Cat** | True Postive (TP) | False Positive (FP) |
| **Predicted Not-Cat** | False Negative (FN) | True Negative (TN) |

These definitions are relatively straightforward. We call the total number of actual positives the **condition positive (P)**, where $P = TP + FN$. The total number of actual negatives is the **condition negative (N)**, where $N = FP + TN$.

A variety of combinations of these 4 quadrants (TP, FP, FN, and TN) are found in the building blocks widely used fairness metrics in machine learning. These building blocks include:

- **True Positive Rate (TPR):**
$$TPR = \frac{TP}{P}$$
- **True Negative Rate (TNR):**
$$TNR = \frac{TN}{N}$$
- **False Positive Rate (FPR):**
$$FPR = \frac{FP}{P}$$
- **False Negative Rate (FNR):**
$$FNR = \frac{FN}{P}$$
- **Positive Predictive Value (PPV):**
$$PPV = \frac{TP}{TP + FP}$$

- **Negative Predictive Value (NPV):**

$$NPV = \frac{TN}{TN + FN}$$

- **False Discovery Rate (FDR):**

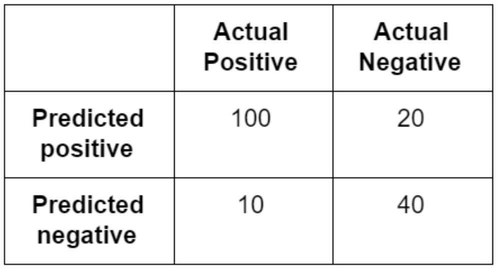
$$FDR = \frac{FP}{TP + FP}$$

- **False Omission Rate (FOR):**

$$FOR = \frac{FN}{TN + FN}$$

## Exercise 1 (20 minutes)

(a) For the following confusion matrix, compute each of the confusion matrix metrics mentioned in the section above (TPR, TNR, etc.)



| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted positive** | 100 | 20 |
| **Predicted negative** | 10 | 40 |

(b) Each of these metrics has an inverse pair (for example TPR = 1 - FNR). Find the three other inverse pairs.

(c) Consider the following scenarios: imagine you are creating a model to determine whether someone should be brought in for an additional cancer screening based upon initial test results. Which metric would you care most about minimizing? What if instead you were creating a model to determine whether a defendant should be prosecuted for a low level offense?

*Impossibility Theorems*

We previously encountered the notions of independence, separation, and sufficiency when thinking about fairness in ML in an earlier assignment (Module 2 Assignment 3). These are different definitions of **group fairness**, which focuses on whether outcomes differ systematically between different demographic groups (for example, people grouped by race, gender, age, disability, etc.). Framed in confusion matrix terms,

- **independence** is when the ratio between P and N is equal across different groups.

- **separation** is when TPR and FNR are equal across different groups.
- **sufficiency** is when PPV and NPV are equal across different groups.

All of these can be framed as variations on **statistical bias**: the difference between expected value and true value. This definition is perhaps reminiscent of the notion of "accuracy" in the model. However, a model that is "accurate", or "not statistically biased" and therefore fair under one metric may not be fair under a different metric. For example, the COMPAS recidivism algorithm was not statistically biased in respect to rearrest (satisfied sufficiency). However, as pointed out by ProPublica, it did not satisfy the separation definition of fairness. In fact, these notions of fairness cannot be reconciled, in general.

The impossibility theorems center on the idea that as soon as a sensitive attribute is not independent from what we are trying to predict, these notions become mutually exclusive. In less abstract terms, this is true whenever prevalence is not observed equally across groups. For example, in the COMPAS recidivism dataset, rearrests were not observed equally between black and white defendants. Consequently, both separation and sufficiency cannot be satisfied in this case, meaning that there is no way to satisfy the fairness requirements of both Propublica and Northpointe.

### External Resource(s) (Optional, 30 minutes)

Read the section Relationships Between Criteria (ends right before Inherent Limitations of Observational Criteria) from Chapter 2 of Fairness and Machine Learning. This section gives several proofs showing that in general, a system cannot satisfy all three of these criteria. Though this may be a bit hard to follow, it is worth checking out to confirm the impossibility theorems for yourself.

### *Trolley Problems and... UOCD?*

Though it might be frustrating to have reasonable metrics of fairness defined in ways that directly compete with one another, one way of framing the core idea of the impossibility theorems is that tradeoffs are inherent in decision making. A popular thought experiment in ethics frames this tension as the trolley problem.

The trolley problem is described as follows: A runaway trolley is headed for 5 people who are lying incapacitated on the trolley's main track. You are standing next to a lever that controls a switch. If you pull the lever, the trolley will be redirected onto a side track, and the five people on the main track will be saved. However, there is a single person lying on the side track. You have two options:

1. Do nothing and allow the trolley to kill the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

In ethics, the two options are usually presented as a difference in utilitarianism and deontological views: in the utilitarian view, it is better to run over one person for the

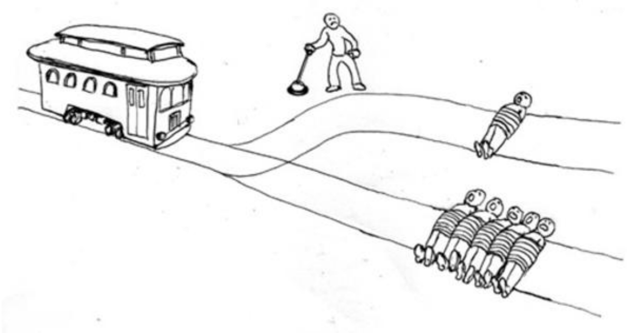


Figure 1: A popular meme depicting the trolley problem.

greater good; whereas in the deontological view, the situation is morally wrong either way, and active participation in pulling the lever makes you partially responsible in a situation where otherwise no one would be responsible.

Generalized, we can think of the decision to choose between competing fairness metrics as different tracks on the trolley system: there is no right answer, and any answer depends inherently on an individual's own values. At Olin, we dedicate a substantial amount of our curriculum to considering the values of different stakeholders in a product. One way of deciding on an appropriate fairness metric may be through prioritizing the values of different stakeholders, as different stakeholders may prefer different fairness criteria.

For example, consider a recidivism algorithm, not unlike COMPAS. A judge using the algorithm to affect her decision may wonder, "of those labelled high-risk, how may will recidivate", and therefore prioritize sufficiency metrics. The defendant may be most concerned with the probability that they will be classified as high-risk, prioritizing independence. Others might wonder if the algorithm treats people from different demographic groups similarly, prioritizing separation. The design of the algorithm and fairness metric used would then rest with the designer of the algorithm and which stakeholder values they prioritized. Indeed, they are standing in front of a switch to a trolley problem.

Many definitions and proposed notions of fairness in machine learning tend to assume a utilitarian framework: how do I optimize a given loss function? This assumed framework and resulting discussion might be justified by focusing on the optimization aspect, because this generally is the most mathematically robust way to describe a "best" solution. Ultimately, however, the underlying challenge to all this is how to make algorithmic systems support human values, and focusing on reframing ethical problems as optimization problems may not be the most productive way to do so. Perhaps it is time for practitioners of machine learning (and technologists in general) to start placing a greater focus on conversations in ethics and philosophy, as their work starts to increasingly interact and intervene directly in society.

Figure 2: In any case, we would want to avoid a multi-track drifting situation.

### Exercise 2 (15 minutes)

Write down your responses to the following questions:

- In reference to the trolley problem, do you find yourself leaning towards one view over another (pull the lever/leave the lever alone)? Can you justify your decision?
- There are many variations on the trolley problem, some of which might be more meaningful to you than others. In one, the problem is reframed as follows: A runaway trolley is headed for 5 innocent civilians who are incapacitated on the trolley's track. The villain responsible for tying the five people down stands on an overpass, directly over the trolley's path. If you push the villain down, the trolley will kill the villain, but this will cause it to stop before it reaches the civilians. Do you push the villain? Did this change anything for you?
- Alternatively, what if the roles of the incapacitated people and the person standing on the overpass are switched: the 5 incapacitated are villains, and the person standing on the overpass is a well-respected hero? Does this change anything for you? Why or why not?
- Another variation on the trolley problem instead changes the context: You are a surgeon, with 5 patients in desparate mortal need of different organ transplants. One healthy person comes in to see you, and by some miracle you see that their organs are compatible for each of the 5 respective other patients. You can use the healthy person's organs to save the other 5; however, the healthy person will inevitably die. No one will ever know what you have done. Would you do it?

## *3 Six Questions of Fairness in Algorithmic Decision Making*

Now we've seen some ways in which human values are inevitably intertwined with creating machine learning algorithms, which often in turn are used to help guide and scale up human decisions. There are, of course, many more questions that might arise in developing such an algorithm.

### *Question 1: Is this a case of disparate impact or disparate treatment?*

Disparate impact refers to policies and practices that adversely affect one demographic group more than another, even though these policies and practices are by themselves neutral. For example, if a library is only accessible by means of a long flight of stairs, the elderly and physically disabled will be more adversely affected than those who are not part of these groups, even though the stairs, viewed in isolation, are neutral.

On the other hand, disparate treatment refers to unequal treatment of demographic groups through practices and policies that actively discriminate against some demographics. For example, if a bouncer stood guard at a library and actively turned away those who they deemed were elderely or physically disabled, this would be an example of disparate treatment.

In practice, these differences may be much more subtle. Sometimes it may be hard to identify whether an algorithm results in disparate impact or is a case of disparate treatment through proxies. In using a fairness algorithm to call out biases in another algorithm, subtleties are lost. Our legal system can work through the tension between disparate impact and disparate treatment on a case-by-case basis. However, because machine learning inherently relies on classification, its outputs are broad generalizations, ignoring the details of an individual case for the bigger picture. Of course, the legal process is notoriously slow, while a major advantage of algorithms is their ability to produce an output quickly.

### *Question 2: Is blindness important in my algorithm?*

In human decision making, blindness is an important fairness metric. For example, it may be important for a recruiter to be given a list of resumes without names, thereby eliminating any racial bias that may be expressed through the names. However, the same importance of blindness may not hold in machine learning. A study by Hardt et al. showed that, in predicting who would default based on credit scores, an algorithm that incorporated blindness toward an individual protected attribute (in this case, race) performed essentially the same as an algorithm that maximized for profit (no fairness constraints). One reason this may be because bias in machine learning is a side-effect of maximizing accuracy. Another is that machine learning algorithms are much better at picking up proxies in the dataset. Therefore, in a situation where blindness is enough for human decision making to be considered fair, the same may not hold true for an algorithm.

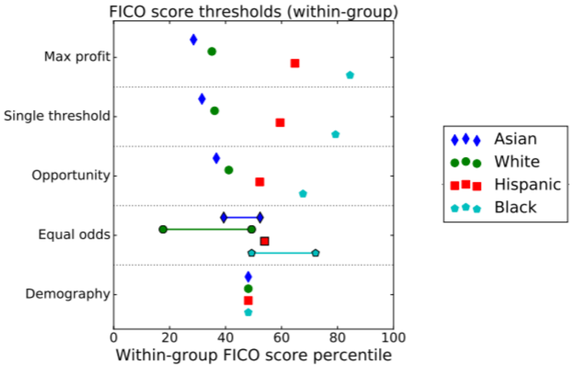


Figure 3: The figure from Hardt et al. that exemplifies this notion. Max profit has no fairness constraints, and will pick for each group the threshold that maximizes profit. This is the score at which 82% of people in that group do not default. Race blind requires the threshold to be the same for each group. Hence it will pick the single threshold at which 82% of people do not default overall.

## *Question 3: What if I focus on individual fairness rather than group fairness?*

Some researchers have tackled the problem of fairness in machine learning through the lens of individual fairness, rather than group fairness. **Individual fairness** is the notion that similar individuals should be treated similarly. A paper by Dwork et al. uses this notion. A primary hurdle in using individual fairness is determining a **distance metric** between individuals, or a way to quantify how different or similar any two individuals in the dataset are. This has close ties to the idea of **diversity**, where instead you are optimizing for some average distance between any two selected individuals. Dwork et al. suggest that this distance metric should be determined by legislators or society as a whole, absolving machine learning practitioners of some responsibility, though ultimately this distance metric is required if they want to operate under this notion of fairness.

## *Question 4: Can a randomized classifier be fair?*

An intuitive response to this might be no: for example, an algorithm that randomly recommended defendants for longer prison terms would probably be considered wildly unfair. Dwork et al. have a counterpoint to this: there are cases where ONLY randomized classifiers can be fair. Using the example of an algorithm that recommended defendants for longer prison terms, if the system is deterministic for a certain threshold of recidivism score (say, 50%), then defendants that were extremely similar would receive vastly different outcomes (say, 49 vs 51%), and therefore not satisfy individual fairness (where similar individuals should be treated similarly). This is exemplified in the image below: a partially randomized classifier between some given thresholds would then average out recommendations across a population, giving similar individuals a closer overall outcome.

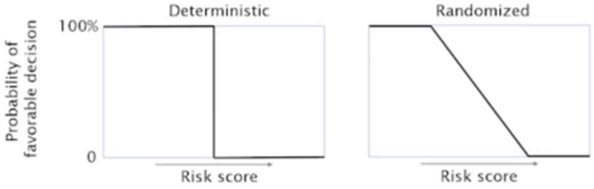


Figure 4: A comparison between a deterministic and a randomized algorithm for a given outcome.

### *Question 5: How much should we trust the decision maker's intentions?*

When making an algorithm to enhance human decision making, how you view the end-user may fundamentally change how you write your algorithm. An end-user might be viewed in one of two camps:

1. The decision maker is cold and uncaring; the algorithm's compliance with the fairness definition is the only thing preventing discriminatory behavior.
2. The decision maker has fundamentally good intentions, and the fairness mechanism helps them to avoid unintentional discrimation.

Do fairness definitions and the subsequent algorithms that implement them need to be robust against bad actors? Is this even possible?

### *Question 6: Why don't we decide on the fairness of a metric through a democratic process?*

The question is the idea behind **process fairness**, or giving each classifying feature a fairness value by asking people, likely other practitioners of FAT/ML to rate these features through a survey. This would perhaps allow process fairness to be an optimization problem. Of course, this faces its own problems. Different classifiers may or may not be biased in the context of different datasets. Further, some aspects of machine learning are notoriously inscrutable, perhaps limiting human intuition as a guide to process fairness across broad contexts. Searching democratically for a "one true fairness definition" is likely to be another dead end.

**Exercise 3 (15 minutes)**

This section asks various questions about fairness in machine learning. Choose one of the questions above and write a short response to it. Your response could incorporate something surprising you learned, a thought-provoking question, your personal experience, an additional resource that builds upon or shifts the discussion.

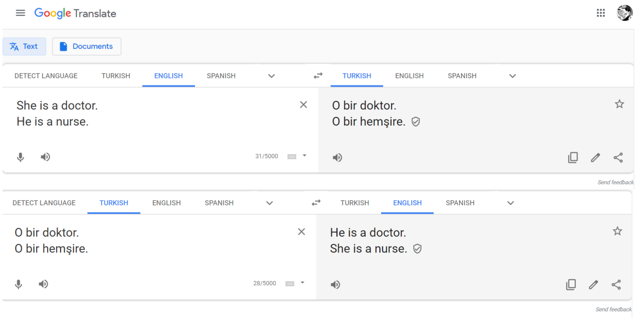
## *4 Other Sources of Bias*

As if the above considerations of fairness weren't enough, there have been other considerations of bias in machine learning systems.

### *Allocative vs. Representational Harms*

An **allocative harm** is one where a system withholds opportunities based on a protected attribute, taking the form of discrete transactions with immediate effects. A **representational harm** is one that has more diffuse, long-term effects on a society, reinforcing subordination of certain demographic groups.

Representational harms may take the form of beauty filters that claim to make people more attractive by lightening their skin in a photo, as one by FaceApp did, or biases that appear in translations by Google Translate.



She is a doctor.
He is a nurse.

O bir doktor.
O bir hemşire.

O bir doktor.
O bir hemşire.

He is a doctor.
She is a nurse.

Figure 5: An experiment on Google Translate performed December 12, 2019 by me.

### *Stereotype Mirroring and Exaggeration*

One disagreement about representational harms can be framed as a discussion about stereotype mirroring and exaggeration. **Stereotype mirroring** is the view that algorithms based on real-world data merely show the strength of a stereotype. In fact, many technologists might argue that stereotype mirroring by an algorithm is a good thing, showing that it conforms to standards set by stereotypes in real life, and therefore is "unbiased", exhibits accuracy, and is therefore correct. **Exaggeration** in this case refers to the idea that the strength of a stereotype can be amplified in an algorithmic representation, when compared to real-world data.

It's important to have conversations about whether or not mirroring is desired in an algorithm. Should machine learning reflect pre-existing stereotypes?

### *Dataset Bias and Responsibility*

There are many commonly-used datasets used to teach machine learning. Theoretically, these datasets are unbiased representations of the visual world. However, a study done by Antonio Torralba and Alexei A. Efros shows that many of these datasets can be told apart by most people. This raises the notion of **cross-dataset generalization**: how well does a model built on one dataset generalize to another? This may be an appropriate test to evaluate the demographic representation of datasets.

Further, should dataset curators be required to do a bias assessment of their datasets? This might include an analysis of demographic representations and biases inherent to their dataset, as well as the intended and unintended contexts of use for these datasets. Similarly, should researchers who release pre-trained models have obligations to do the same, or even to "de-bias" their systems?

### Exercise 4 (15 minutes)

You may have noticed by now that this assignment asks for a lot of shared short reflections! Because there are so many varied perspectives on fairness in ML, one of the biggest priorities of this assignment is to foster conversation about fairness in machine learning within the community (that's you!). So without further ado, here are a couple more items to write and reflect on:

(a) List one recent example of representational harm that you've seen. This could be an example outside of machine learning.

(b) Do you believe machine learning should reflect pre-existing stereotypes? Or do you believe that algorithms and their designers have a responsibility to correct, or at the very least, not reinforce societal stereotypes?

(c) Do you believe practitioners of machine learning have an obligation to perform bias assessment on their datasets and pre-trained models? Do you believe models found to have bias in them should be "de-biased" by the creator?

## 5 *Takeaways and Using Fairness Algorithms*

We've seen that there are many definitions for fairness, and even more considerations to have in mind when considering fairness in an algorithm. Again, though it may be unsatisfying to not have one unifying definition of fairness, different definitions and considerations are useful to have for different stakeholders, contexts, and applications. Viewing the impossibility theorems as trolley problems might motivate us to avoid "multilane drifting", or over-constraining our system with different definitions of fairness. Ultimately, we should ask ourselves what is most important in the algorithms we produce: mathematical correctness, or their ability to support human values. This assignment is only a starting point; if you are interested in these topics, there are many great resources out there, including (but certainly not limited to):

- Papers and tutorials from the ACM FAT/ML conference
- Just(1) a(2) few(3) of(4) the(5) many(6) academic(7) papers(8) discussing(9) various(10) fairness(11) metrics and algorithms to check for it.
- A really interesting paper on computational empircism that I didn't get to include a discussion on in this assignment (thought it might be a little overkill at this point)

Though we might decide that mathematical models of fairness are not enough, they are certainly a starting point. Many excellent fairness algorithms have been developed, among which is the IBM AI Fairness 360 Toolkit package.

## External Resource(s) (40 minutes)

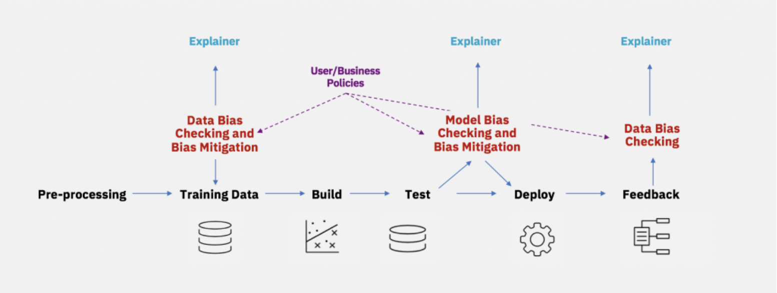
(a) Take a look at the IBM AI Fairness 360 Toolkit. This toolkit uses 9 different algorithms, developed by the algorithmic fairness community, to mitigate unwanted bias. According to its own intro, the toolkit "focuses on bias mitigation, as opposed to simply on metrics, its focus on industrial usability, and its software engineering". (10 minutes)

(b) Read through the section Guidance on Using the Toolkit. (15 minutes)

(c) Explore and work through the toolkit demo. (15 minutes)

(d) (optional) If you would like, you can use the toolkit in your own machine learning algorithms. The toolkit comes with two tutorials that may help you implment these fairness checks and mitigation methods.

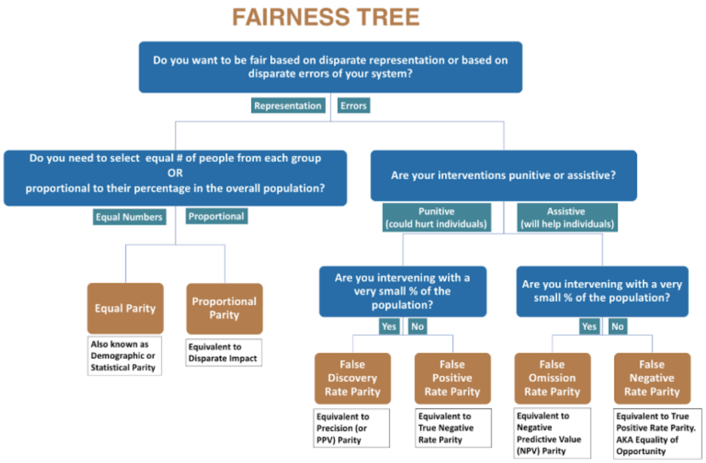
## Exercise 5 (30 minutes)



(a) The above diagram from the blog introduction shows three places in the machine learning workflow where fairness checking and mitigation can occur. For each of the three places (Training Data, Test/Deploy cycle, Feedback), brainstorm the types of fairness might be relevant at each stage.

(b) As you read through the Guidance on Using the Toolkit, write down your responses to the following prompts:

- When reading through the "vs" section (Individual vs. Group, Data vs. Model, WAE vs. WYSIWYG), do you find yourself leaning towards one or the other?
- Can you think of datasets or situations that would work better under different measures of fairness?


FAIRNESS TREE
Do you want to be fair based on disparate representation or based on disparate errors of your system?
Representation
Errors
Do you need to select equal # of people from each group OR proportional to their percentage in the overall population?
Equal Numbers
Proportional
Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Equal Parity
Also known as Demographic or Statistical Parity
Proportional Parity
Equivalent to Disparate Impact
Are you intervening with a very small % of the population?
Yes
No
Are you intervening with a very small % of the population?
Yes
No
False Discovery Rate Parity
Equivalent to Precision (or PPV) Parity
False Positive Rate Parity
Equivalent to True Negative Rate Parity
False Omission Rate Parity
Equivalent to Negative Predictive Value (NPV) Parity
False Negative Rate Parity
Equivalent to True Positive Rate Parity. AKA Equality of Opportunity

(c) The decision tree (pictured above) attaches certain decisions to different measures of bias. Does it make sense why each consideration is linked to its corresponding measure of bias? (You do not need to write a response to this question; just check your understanding here).

(d) As you work through the demo, reflect on your findings by answering the following:

- Which demo dataset did you look at?
- Which measure (or measures) of statistical bias would you aim to mitigate for this particular dataset? Explain why you would choose to focus on mitigating these measures of bias.
- Try the different methods of mitigation. Which one was most effective for mitigating the bias you identified for your chosen dataset? Why do you think this was the case? What might this say about the data?
- Repeat the process for the other two datasets. Was there any one method that worked best for all the datasets? (Hint: did your results remind you of the No Free Lunch Theorem?)