Fairness Metric Selection & Bias Mitigation Mapping

Equality AI (EAI) is a public benefit corporation dedicated to giving developers the solutions to mitigate algorithmic bias at scale. We believe that developers want to implement algorithmic fairness, but have encountered roadblocks along the way. We present our solutions: a cloud-based Responsible AI studio, open-source tools, and educational materials on how to integrate fairness into your traditional machine learning (ML) coding steps, to help our users make an algorithmically fair world.

One way that fairness can be integrated into the ML process is through creating parity (equality) on appropriate fairness metrics before model deployment, then tracking those metrics throughout deployment. Fairness should be a primary model consideration, just like measuring model utility (loss, accuracy, etc.). To learn more about fair ML, go to our GitHub.

Selecting and implementing the appropriate fairness metric is necessary to mitigate the primary sources of harm your end-users face. To make fairness metric selection easy we have provided a few essential questions you must answer to identify the appropriate fairness metric for your use case. Complete the answers to this questionnaire, then refer to the scoring guide to map your inputs to the desired metrics. Please note that while the wording of this questionnaire is phrased to assist provisioning of a binary decision/resource, we also provide the analogous probability-based risk score metrics in the scoring section.

Table of Contents:

Questionnaire

For each question, select the most appropriate response.

1. **Are your labels ground truths or are they subjective (or a proxy)?**
    - Ground truth
    - Subjective (or a proxy)

    *If you observe the label directly, it is a **ground truth**. If you do not observe the label directly, it is either **subjective** (labeled by a human) or a **proxy** (a variable that is thought to be correlated with an underlying concept that cannot be measured directly).*

2. **Are you concerned about historical bias in the labels of your dependent variable?**
    - Yes
    - No

    ***Historical bias** exists when data labels were assigned in an inequitable way, such as when they are more accurate for one group than another, or they reflect an inequitable distribution of resources, typically favoring the group with more political power.*

3. **Is the resource the model provisions rationed or not?**
    - Rationed
    - Not rationed

    *A resource is **rationed** if it is scarce, and the demand for the resource exceeds the supply of it (such as organ transplants). A resource is **not rationed** if supply of the resource exceeds demand for the resource, or it may be liberally distributed (such as over-the-counter headache medication).*

4. **Is the model recommendation or decision assistive, punitive, or neither?**
    - Assistive
    - Punitive
    - Neither

    *A recommendation is **assistive** if it is beneficial to the recipient, and they would seek to obtain it (such as an organ transplant). Note that most healthcare interventions are assistive. A recommendation is **punitive** if it is harmful to the recipient, and they would seek to avoid it (such as a criminal charge). A recommendation is **neither** punitive nor assistive if the magnitude of the potential harms of erroneously being classified as positive or negative are approximately the same, or if these harms are very small.*

Scoring Guide

Below is the questionnaire scoring guide, which maps the 4 above answers to fairness metrics. There are two sections to this questionnaire. The first section is about your confidence in your data source. The second section is about the sources of harm and incentives the end-user and institution hosting the model faces. You should complete both sections to determine the appropriate fairness metric(s).

Section 1: Confidence in Data

The first and second questions focus on the developer's confidence in the data.

Question 1: Are your labels ground truths or are they subjective (or a proxy)?

Question 2: Are you concerned about historical bias in the labels of your dependent variable?

If you answered "Subjective (or a proxy)" to Question 1 **OR** "Yes" to Question 2 then:

We recommend that you track **Statistical Parity** or **Conditional Statistical Parity** in addition to any other fairness criteria we recommend later on. Because we don't trust the labels of our dependent variables, we are going to ignore them as a ground truth when establishing fairness criteria. We are going to focus instead on the proportion of people per sensitive attribute. We will match the proportion of people the model decides are positive across groups.

> Note that **parity** is achieved when a fairness metric (such as the percent of positive predictions) has the same value across all levels of a sensitive attribute.
>
> **Sensitive attributes** are attributes such as race, gender, age, and other patient attributes that are of primary concern when it comes to fairness, and are typically protected by law.
>
> **Statistical Parity** is achieved when you observe the same percent (or number) of positive predictions across sensitive attributes. We recommend using this fairness criteria in cases where the label is unreliable, causing error metrics derived from the label to also be unreliable. When using this criterion, we distrust our labels, and focus our fairness efforts on the proportion of positive classifications per level of the sensitive attributes.

Conditional Statistical Parity is a similar concept to statistical parity, with the primary difference being that we desire the percent (or number) of positive predictions to be the same across sensitive attributes **and** legitimate predictors (reliably measured attributes that the label should be assigned based on according to experts, such as correctly assessed cancer stage for a chemotherapy recommendation). Conditional Statistical Parity should be used in cases where legitimate predictors of the label are also correlated with sensitive attributes.

We expect most studies based on real-world care to track Statistical Parity or Conditional Statistical Parity due to the presence of historical bias and lack of ground truth. You can track these metrics while also tracking other metrics as well. Continue Section 2 below.

Section 2: Sources of Harm

Here we focus on the remaining two questions about the end-user incentives of the decision being made by the model and the scarcity of the resource being provisioned based on it.

Question 3: Is the resource the model provisions rationed or not?

Question 4: Is the model recommendation or decision assistive, punitive, or neither?

The below table maps the answers to questions 3 and 4 to our recommendations for fairness criteria. Recall that the term parity means that when fairness is achieved, the value will be the same across all levels of a sensitive attribute. Note that we have provided recommendations for both binary classification and probability measures of fairness.

	Question 3	
Question 4	**Rationed (Scarce)**	**Not rationed (Common)**
Assistive	**Binary Classification:** Negative Predictive Parity	**Binary Classification:** Equal Opportunity **Probability:** Balance for Positive Class
Punitive	**Binary Classification:** Predictive Parity **Probability:** Calibration/Well Calibration	**Binary Classification:** Predictive Equality **Probability:** Balance for Negative Class
Neither	**Binary Classification:** Conditional use Accuracy Equality	**Binary Classification:** Equalized Odds **Probability:** Overall Balance

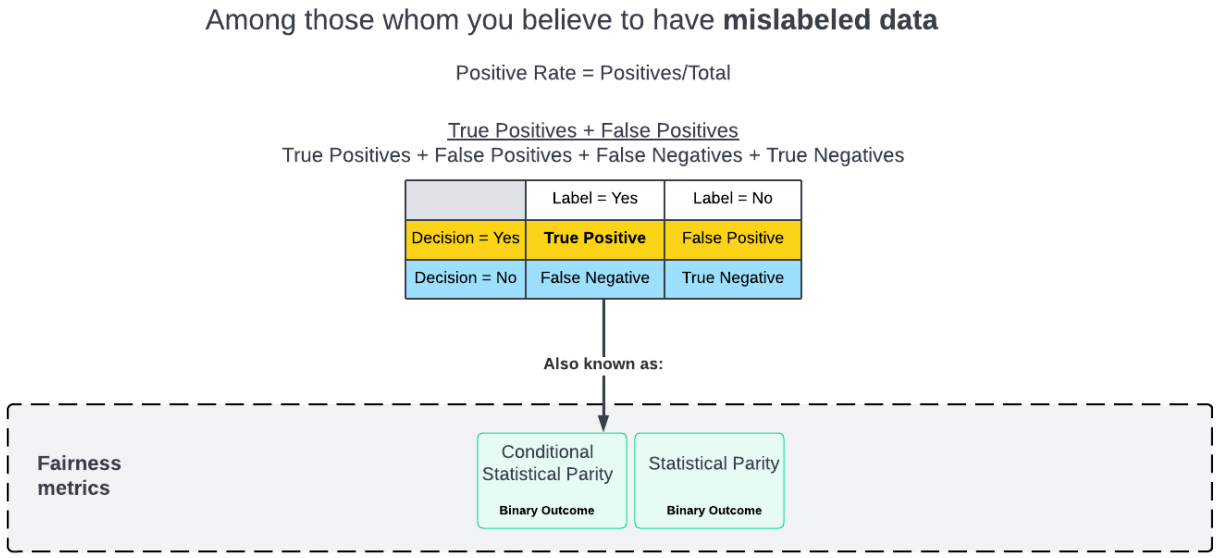
Sources of Harm

The categories of sources of harm include false negatives, false positives, both false positives and false negatives, and mislabeled data. After identifying the sources of harm of most concerned regarding your use case, you can include fairness metrics alongside performance metrics and bias mitigation methods targeting those sources of harm to model fitting, and monitoring.

If neither, false positives nor false negatives incur harm focus on minimizing your loss function or maximizing the gain function (as appropriate to your use case) instead.

Mislabeled Data

If you are concerned with mislabeled data or have little trust in your data, you want to include Statistical Parity, or Conditional Statistical Parity (see below).



When data labels are incorrect they cannot reliably be used to train a classifier.

Attempt to minimze harm by creating parity on the positive classification rate by sensitive group.

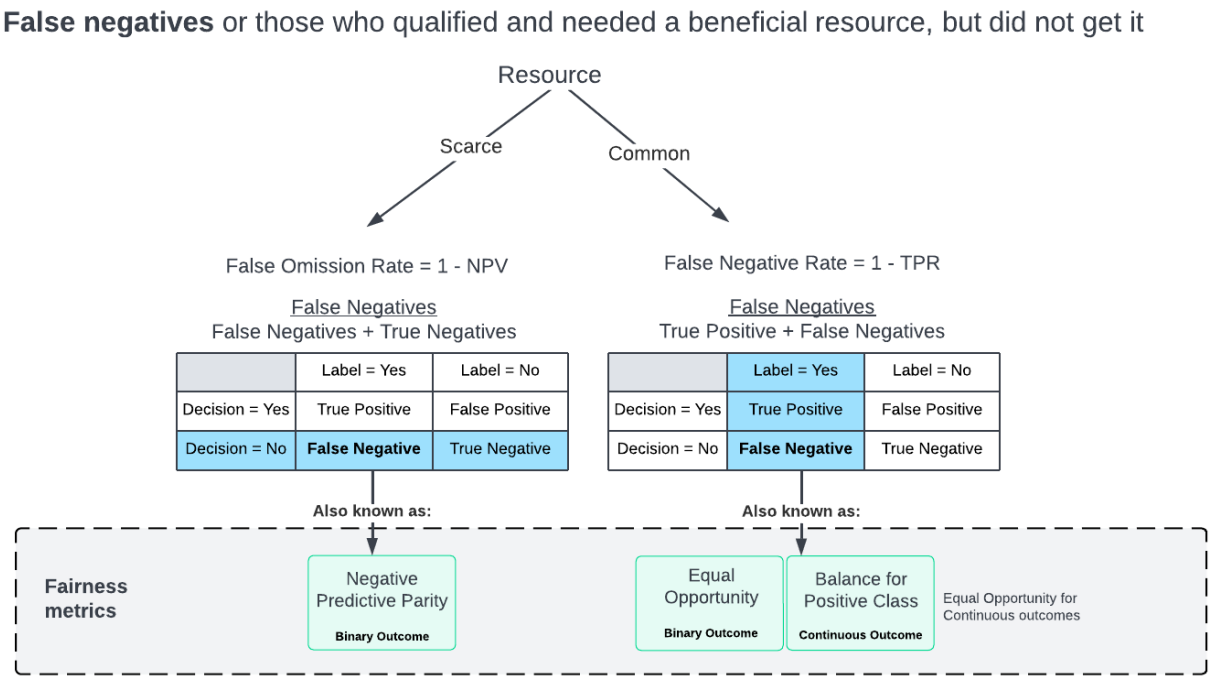
If one group's labels are assigned more accurately than another, this may result in future failure to meet expectations by a historically disadvantaged group. To mitigate this risk, support must be provided to this group to ensure this does not occur.

Accurately labeled data is very valuable, yet it is hard to create truly unbiased labels due to the complexity of the world. Acquiring correctly labeled data is another appropriate solution.

We recommend conditional statistical parity over statistical parity because statistical parity alone may ignore important risk factors and incorrectly classify individuals with important risk factors.

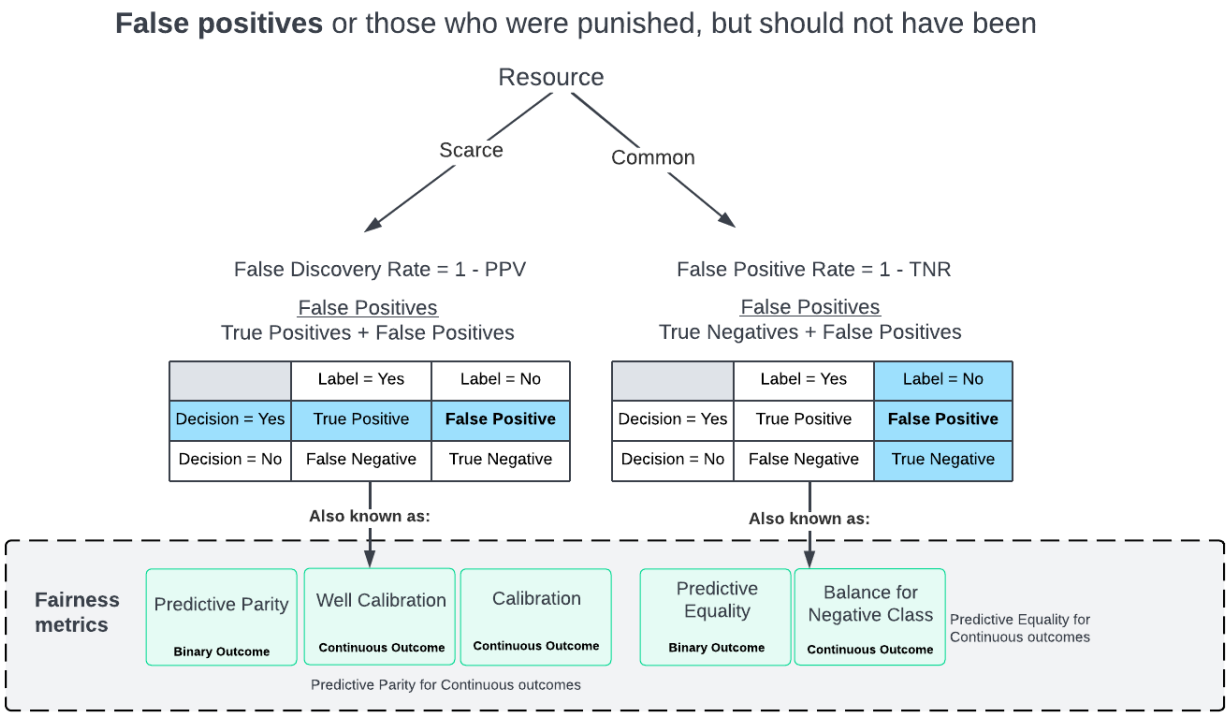
False Negatives

If you are concerned with false negatives, or people being harmed by not receiving a beneficial resource then you want to include Negative Predictive Parity, Equal Opportunity, or Balance for Positive Class (see figure below).



False Positives

If you are concerned with false positives, or people being punished when innocent, then you want to include Predictive Parity, Calibration, Well Calibration, Predictive Equality, or Balance for Negative Class (see figure below).



The resource is harmful to all who recieve it, although it is targeted to the guilty. Because of this, the greatest harm is to those punished unjustly.

The resource is scarce (fewer people will be punished than were guilty). Due to the rarity of the punishment, there is potential for greatest total harm among the punished.

To minimize unfairness, we will ensure those punished were guilty at the same rate across sensitive groups.

The resource is harmful to all who recieve it, although it is targeted to the guilty. Because of this, the greatest harm is to those punished unjustly.

The resource is common (everyone who is guilty may be punished). Due to the abundance of the punishment, there is potential for greatest total harm among the innocent.

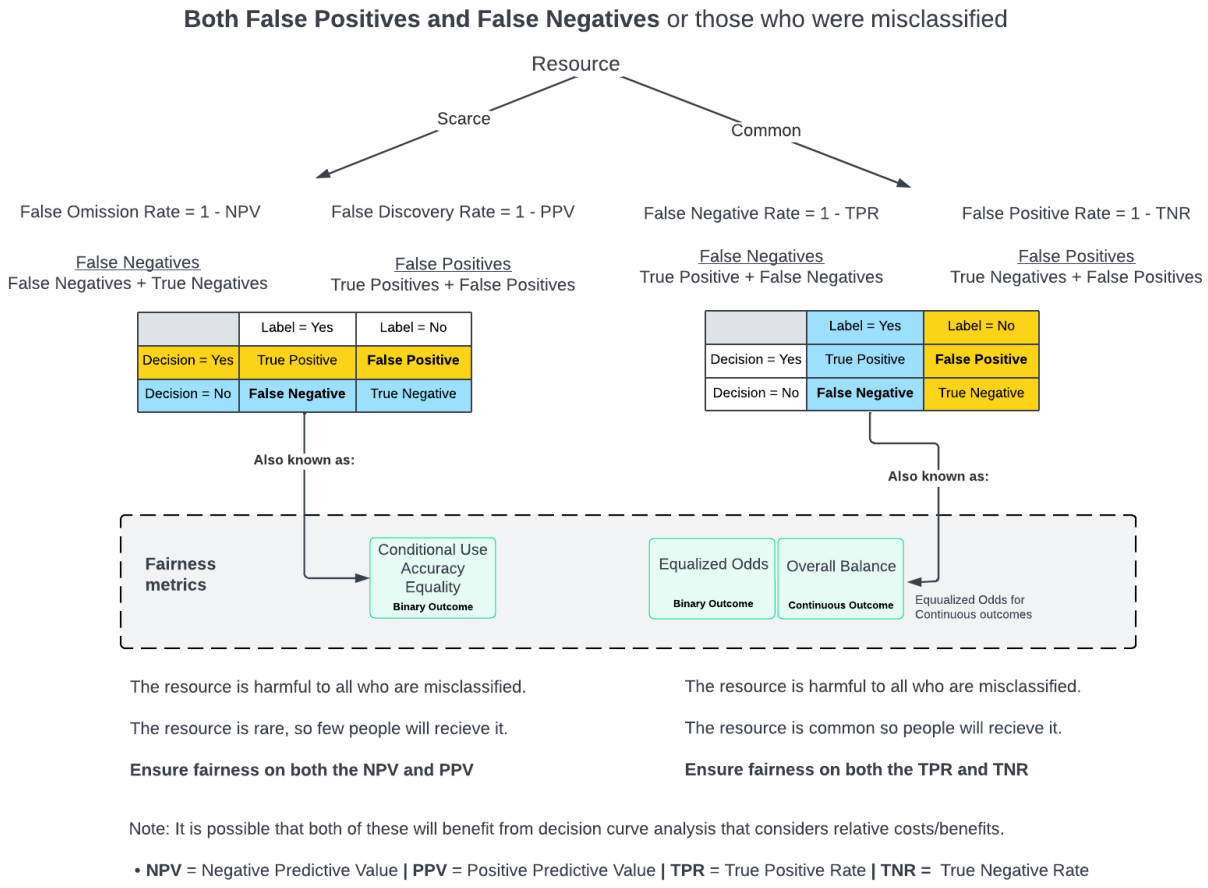
To minimize unfairness, we will ensure the innocent are protected from unjust punishment at the same rate across sensitive groups.

We recommend Well-calilibration over calibration because it requires the probability of the event to match the risk score, which is often a pragmatic choice. Calibration relaxes that constraint, but otherwise imposes the same fairness constraint across sensitive groups.

- **PPV** = Positive Predictive Value **| TNR =** True Negative Rate

Both False Positives and False Negatives

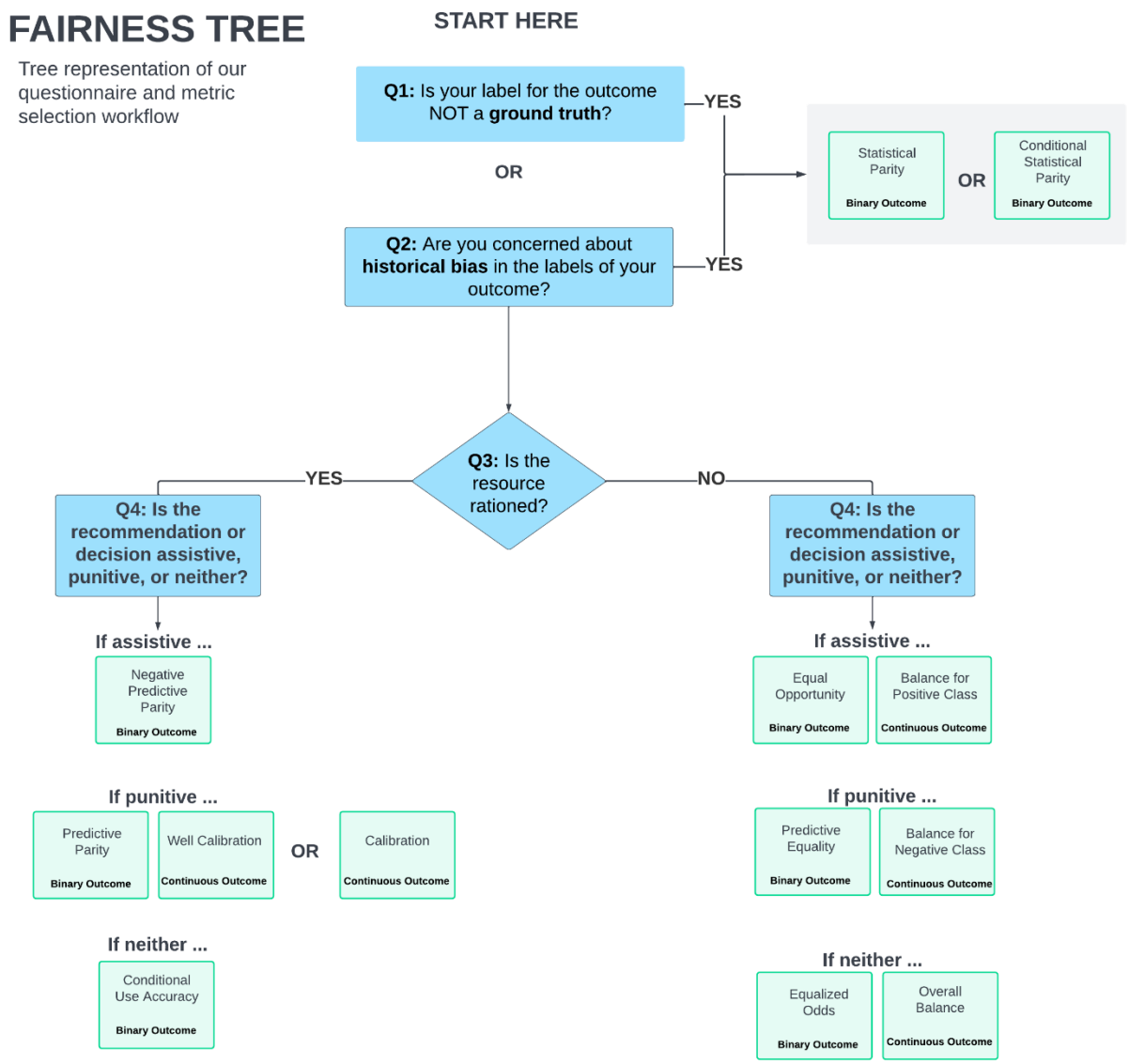
If you are concerned with false negatives, or people who have been misclassified, then you want to include Conditional Use Accuracy Equality, Equalized Odds, or Overall Balance (see figure below).



Note: It is possible that both of these will benefit from decision curve analysis that considers relative costs/benefits.

- **NPV** = Negative Predictive Value **| PPV** = Positive Predictive Value **| TPR** = True Positive Rate **| TNR =** True Negative Rate

Tree


FAIRNESS TREE
Tree representation of our questionnaire and metric selection workflow
START HERE
Q1: Is your label for the outcome NOT a ground truth?
YES
OR
Q2: Are you concerned about historical bias in the labels of your outcome?
YES
Statistical Parity
Binary Outcome
OR
Conditional Statistical Parity
Binary Outcome
Q3: Is the resource rationed?
YES
NO
Q4: Is the recommendation or decision assistive, punitive, or neither?
Q4: Is the recommendation or decision assistive, punitive, or neither?
If assistive ...
Negative Predictive Parity
Binary Outcome
If punitive ...
Predictive Parity
Binary Outcome
Well Calibration
Continuous Outcome
OR
Calibration
Continuous Outcome
If neither ...
Conditional Use Accuracy
Binary Outcome
If assistive ...
Equal Opportunity
Binary Outcome
Balance for Positive Class
Continuous Outcome
If punitive ...
Predictive Equality
Binary Outcome
Balance for Negative Class
Continuous Outcome
If neither ...
Equalized Odds
Binary Outcome
Overall Balance
Continuous Outcome

Fairness Metric to Bias Mitigation Method Mapping

After identifying the important fairness criteria, we recommend you attempt to use multiple bias mitigation strategies to try to optimize the efficiency–fairness tradeoff. Bias mitigation methods target specific fairness metrics, so use the mapping below to identify the bias mitigation methods that you should try.

Next Steps

After identifying the important fairness criteria, we recommend you attempt to use multiple bias mitigation strategies to try to optimize the efficiency-fairness tradeoff. If more than one fairness criteria is important to your learning task, you may attempt to optimize them both, however, sometimes this is not possible.

Definitions and Examples

If you observe the label directly, it is a **ground truth**. If you do not observe the label directly, it is either **subjective** (labeled by a human) or a **proxy** (a variable that is thought to be correlated with an underlying concept that cannot be measured directly).

Examples:

Ground Truth	Death Hospital readmission Lab values No show	**Subjective/Proxy**	Diagnoses in the EMR EGFR, BMI Self-reported data

Historical bias exists when data labels were assigned in an inequitable way, such as when they are more accurate for one group than another, or they reflect an inequitable distribution of resources, typically favoring the group with more political power.

Examples:

- If you have reason to believe that the dependent variable has been underreported in one or more groups, then you have historical bias in the labels of your dependent variable.
- Historically African Americans are an underserved healthcare population so they may not have complete labels for dependent variables, or perhaps smaller sample sizes. This can lead to models that will perform worse for this population, which further leads to this population continually being underserved.

A recommendation is **assistive** if it is beneficial to the recipient, and they would seek to obtain it. A recommendation is **punitive** if it is harmful to the recipient, and they would seek to avoid it. A recommendation is **neither** punitive nor assistive if a patient would not seek to obtain or avoid it.

Examples:

- Punitive includes fraud detection, overbilling, etc.

A resource is **rationed** if it is scarce, and the demand for the resource exceeds the supply of it. A resource is **not rationed** if supply of the resource exceeds demand for the resource, or it may be liberally distributed or common.

Note that **parity** is achieved when a fairness metric (such as the percent of positive predictions) has the same value across all levels of a sensitive attribute.

Sensitive attributes are attributes such as race, gender, age, and other patient attributes that are of primary concern when it comes to fairness, and are typically protected by law.