Fairness Metric Selection Questionnaire & Tree

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 continuous 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 outcome?**
    - ○ 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 model recommendation or decision assistive, punitive, or neither?**
    - ○ Assistive
    - ○ Punitive
    - ○ Neither

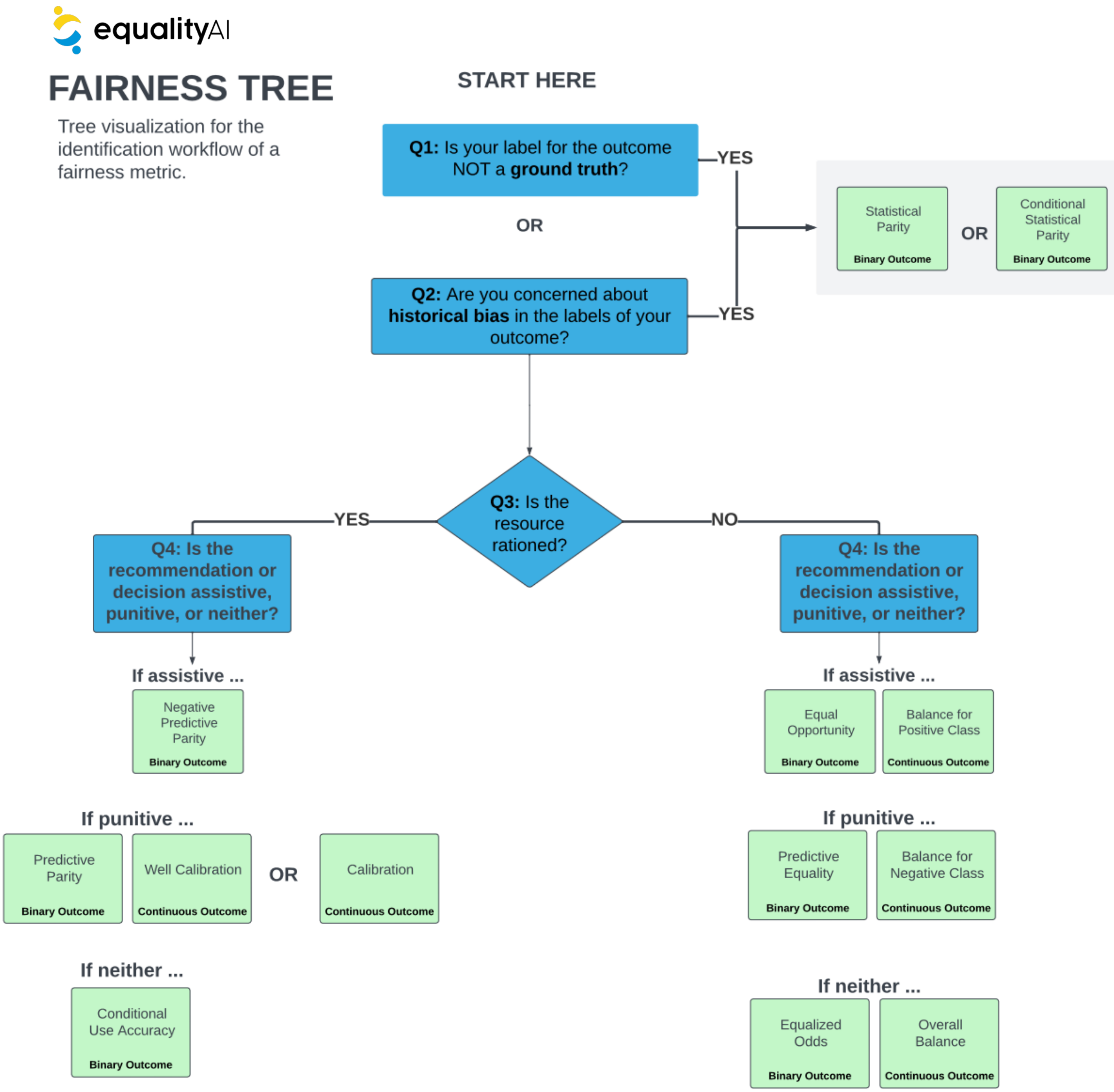
   A recommendation is ***assistive*** *if it is beneficial to the recipient, and they would seek to obtain it. A recommendation is* ***punitive*** *if it is harmful to the recipient, and they would seek to avoid it. A recommendation is* ***neither*** *punitive nor assistive if a patient would not seek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

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

Tree

Below is a visual tree representation of the questionnaire. For each question, select the most appropriate response.



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?

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

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 outcomes, 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) have 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 groups. 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 criteria, 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 groups **and** legitimate predictors.

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 and continuous measures of fairness.

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

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 outcome has been underreported in one group or more than you have historical bias in your the labels of your outcome.
- Historically African Americans are an underserved healthcare population so they may not have complete labels for outcomes or 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.

Note that **parity** is achieved when a fairness metric (such as the percent of positive predictions) have 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.