Fairness Definition Selection Tool

This document outlines a Fairness Definition Selection Tool, including a catalog of fairness definitions, a decision tree for selection, trade-off analysis, and a user guide to facilitate informed and strategic choices. The selection of an appropriate fairness definition is crucial to ensuring equitable and just treatment across diverse populations.

User Guide

Deploying Predictive Equality Effectively

This section provides a step-by-step approach to implementing Predictive Equality as the guiding fairness definition for your tool.

# Step 1: Preparation

* Define the scope of fairness within your system.
* Collect and analyze data on sensitive attributes and demographic groups.
* Understand the ethical and regulatory requirements specific to your domain.

# Step 2: Explore the Fairness Definition Catalog

* Review fairness definitions to identify the best fit for your needs.
* Evaluate strengths, weaknesses, and practical examples for each definition.

# Step 3: Apply the Decision Tree

* Navigate the decision tree systematically.
* Document your selections and the rationale behind them for transparency.

# Step 4: Trade-Off Analysis

* Perform a detailed trade-off analysis to understand the implications of your choice.
* Factor in accuracy, complexity, and ethical considerations.

# Step 5: Monitor and Adapt

* Continuously monitor system performance and fairness metrics.
* Iterate your approach as needed to address evolving circumstances or entrenched biases.

# Fairness Definition Catalog

The following are the fairness definitions with their mathematical formulations of fairness. They are divided into 3 groups namely; Group, Individual, and Counterfactual.

## Group Fairness (group-level parity)

## Demographic Parity or Statistical Parity

**Description**: Ensures that different demographic groups have equal probabilities of receiving positive outcomes (predictions), regardless of individual characteristics e.g. *In hiring decisions, demographic parity would mean that candidates from different gender groups have equal chances of being selected.*

**Mathematical Formulation**:

*P(Ŷ=1|A=a) = P(Ŷ=1|A=b)*

Where

* Ŷ is the predicted outcome,
* A is the protected attribute e.g. gender
* Ŷ=1 is the positive outcome e.g. hired
  1. Equal Opportunity

**Description**: Ensures that different demographic groups have equal positive rates i.e. individuals who are qualified for a positive outcome (based on relevant criteria) are equally likely to receive that outcome, irrespective of their demographic group e.g. *Admissions to a university, equal opportunity ensures that applicants from different socioeconomic backgrounds who meet the eligibility criteria have the same probability of acceptance.*

**Mathematical Formulation**:

*P(Ŷ=1|Y=1, A=a) = P(Ŷ=1|Y=1, A=b)*

Where

* Ŷ is the predicted outcome,
* Y is the true outcome,
* A is the protected attribute e.g. ZIP code
* Ŷ=1 is the positive outcome e.g. admitted

Use when you are concerned about fairness in giving opportunities to those who deserve them.

* 1. Equalized Odds

**Description**: requires both true positive (TP) rates and false positives (FP) rates to be equal across groups e.g. *in hiring processes, it is vital to ensure that candidates from different demographic groups are assessed equitably, not only in terms of their suitability for a role but also in the likelihood of being misclassified.*

**Mathematical Formulation:**

*P(Ŷ=1|Y=y, A=a) = P(Ŷ=1|Y=y, A=b)*

*for y in {0,1}*

Where

* Ŷ is the predicted outcome,
* Y is the true outcome,
* A is the protected attribute e.g. ZIP code
* y is either of the outcomes i.e negative outcome or positive outcome
* Ŷ=1 is the positive outcome e.g. admitted

Use when you want fairness in both benefits and harms i.e. errors do not unfairly benefit ot harm certain groups.

* 1. Predictive Parity

**Description**: requires equal positive predictive values across groups e.g. In healthcare, all positively classified should be equally likely to be truly positive.

**Mathematical Formulation:**

*P(Y=1 | Ŷ=1, A=a) = P(Y=1| Ŷ=1, A=b)*

Where

* Ŷ is the predicted outcome,
* Y is the true outcome,
* A is the protected attribute e.g. ZIP code
* Ŷ=1 is the positive outcome e.g. admitted

Use when you want trust in the predictions e.g. in healthcare diagnosis.

## Individual Fairness

Treats similar individuals similarly, regardless of their demographic group.

* 1. Similarity-Based Fairness

Description: requires that similar individuals receive similar predctions, regardless of protected attributes.

**Mathematical Formulation:**

For individuals xi and xj:

*dy(Ŷ(*xi *), Ŷ(*xj *)) ≤ L\* dx(*xi *,* xj*’)*

Where

* *d* is the similarity metric in
  + *dx* input space,
  + *dy* output space
* Ŷ is the predicted outcome,
* L is a Lipschitz constant

Use when you care about cas-by-case fairness not just group-level.

* 1. Fairness Through Awareness

**Description**: Incorporates awareness of sensitive attributes to ensure equitable treatment e.g. adjusting financial predictions for individuals with systemic barriers.

* 1. Counterfactual Fairness

Description: asks whether predictions would change if an individual’s protected attribute were different, everything else being equal.

**Mathematical Formulation:**

*P(*ŶA←a​(U)=y | X=x, A=a) = *P(*ŶA←a​(U)=y | X=x, A=a)

Where

* ŶA←a​ is the prediction in a world where there is a change in the protected attrubute
* U represents background factors,
* A is the protected attribute
* X represents the observed variables

Useful when you want a causal definition of fairness that deals with historical patterns of bias or inequality.

# Definition Selection Decision Tree

The decision tree simplifies the process of selecting a fairness definition by providing a step-by-step guide based on the specific needs and priorities of the system.

## Step 1: Define the Fairness Goal

Determine the primary goal of fairness in your system.

* If equitable outcomes across groups are essential, consider Demographic Parity.
* If error distribution matters, consider Equalized Odds or Predictive Equality.
* If individual equity is critical, select Individual Fairness.

|  |  |
| --- | --- |
| **Goal** | **Use This Fairness Definition** |
| Ensure equal outcomes across demographic groups (regardless of qualification) | **Demographic Parity (Statistical Parity)** then proceed to step 2. |
| Ensure fair treatment of individuals with similar qualifications | **Individual Fairness** or **Fairness Through Awareness** then proceed to step 2. |
| Ensure error rates are fairly distributed across groups | Proceed to Step 2 |

## Step 2: Error-Impact Analysis

## Determine which type of errors are most harmful in your application context:

## Question: Which error type has greater negative impact in this application?

## If false negatives (FN) are more harmful: Make equal opportunity a mandatory fairness definition.

## If false positives (FP) are more harmful: Make predictive equality a mandatory fairness definition.

## If both error types are equally critical: Make equalized odds a mandatory fairness definition.

## After addressing the relevant error impacts, proceed to Step 3.

## Step 3: Outcome Calibration Assessment

## Question: Will the system expose probabilistic scores to users or analysts (e.g., credit-risk scores, insurance pricing, ranking algorithms)?

## If yes: Add sufficiency (group-calibrated scores) to your fairness metric set.

## If no: Use the definitions selected in the previous steps.

Step 4 (Optional): Consider Counterfactual Fairness

Do you want to ensure fairness across “what-if” scenarios (e.g., what if a person’s gender or race were different)?

* Yes → Consider Counterfactual Fairness, which ensures that decisions remain the same under a change in sensitive attribute, all else being equal.
* No → Proceed with the prior fairness metrics.

## Step 5: Evaluate Trade-Offs

Analyze the compromises between fairness definitions and system performance. Use the trade-off analysis in the next section to guide your decision.

# Trade-Off Analysis

No fairness definition is universally applicable. Each comes with trade-offs between fairness objectives and practical implications. The trade-off analysis should provide insight into balancing these competing priorities.

Analytical Matrix for Fairness Definitions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Fairness Definition | Impact on Accuracy | Implementation Complexity | Ethical Considerations | Domain Constraints |
| Demographic Parity | Prioritizes group equality | Can reduce predictive accuracy | Relatively straightforward to implement | May oversimplify fairness | Applicable to general use cases |
| Equalized Odds | Balances error rates across groups | Requires detailed error analysis | Demanding advanced tools | Encourages equitable outcomes | Useful in high-stakes decision-making |
| Individual Fairness | Ensures similar individuals are treated similarly | May reduce efficiency due to complex models | Requires ongoing monitoring | Addresses systemic inequities | Best suited for personalized systems |
| Fairness Through Awareness | Incorporates sensitive attributes | Improves equity but risks privacy concerns | Ethically challenging | Can address deep-rooted disparities | Applicable in domains with historical bias |
| Predictive Parity | Focuses on consistent predictive performance | Balances accuracy without major compromise | Moderate complexity | Minimizes ethical dilemmas | Ideal for criminal justice, risk assessments |

# Selected Fairness Definition

This section outlines the chosen fairness definition, its mathematical formulation, the rationale behind its selection, and the trade-offs associated with its implementation.

## Fairness Definition

Predictive Equality has been selected as the guiding fairness definition for this tool.

## Mathematical Formulation

Predictive Equality ensures that the predictive performance (such as false positive rates or false negative rates) is consistent across sensitive demographic groups. Mathematically, this can be expressed as:

* False Positive Rate (Group A) ≈ False Positive Rate (Group B)
* False Negative Rate (Group A) ≈ False Negative Rate (Group B)

## Rationale

The choice of Predictive Equality stems from its ability to balance ethical concerns and practical applicability. This definition minimizes ethical dilemmas while maintaining accuracy, making it ideal for domains such as criminal justice and risk assessments. Furthermore, Predictive Equality is moderate in complexity and can address systemic biases without major compromises to model performance.

## Trade-off Acknowledgement

While Predictive Equality is effective in ensuring balanced predictive outcomes, it does not fully eliminate disparities in outcomes or access. In contexts where historical bias is more entrenched, additional measures may be required to complement this approach. The trade-off lies in achieving fairness in predictions while accepting potential sacrifices in broader equity outcomes.