Historical Context Assessment Tool (HCAT)

This tool enables engineering and AI teams to assess historical discrimination patterns, data representation issues, system-level amplification risks, and ethical tradeoffs relevant to the design, development, and deployment of AI systems.

# Section 1: Domain and Application Context

**1.1 Application Domain Identification**

• What specific domain will this system operate in (e.g., lending, hiring, healthcare)?

• What function will it serve within this domain?

**1.2 Historical Discrimination Patterns**

• What documented patterns of discrimination exist in this domain (e.g., redlining, exclusion)?

### Pattern-to-Risk Mapping Table

|  |  |  |  |
| --- | --- | --- | --- |
| Historical Pattern | System Feature Affected | Encoded Social Categories | Risk |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**1.3 Technical Stratification Analysis**

• How have prior technologies contributed to or challenged social stratification?

• Are some populations treated as 'edge cases' or excluded from optimization?

**1.4 Selective Optimization**

• Are some groups disproportionately prioritized in success metrics or optimization logic?

# Section 2: Data and Representation Analysis

**2.1 Historical Data Sources**

• What dataset(s) will inform the system?

**2.2 Categorical Formation**

• How were key categories in the dataset(s) historically defined? How were they sourced, and under what context? What ethical risks do they pose, and what actions can be taken to mitigate risks.

### Risk Mapping Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Source of Bias | Affected Population | Ethical Risk | Action Taken |
| Race | Census legacy classification | Multiracial individuals | Misrepresentation | Recode with expanded categories |
| Gender | Binary-only input | Non-binary people | Exclusion | Add third-category or free text option |
|  |  |  |  |  |
|  |  |  |  |  |

**2.3 Strategic Ignorance and Missing Data**

• What groups or dimensions are missing or underrepresented in data?

### Representation Gap Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Population Share (%) | Dataset Share (%) | Gap (%) | Observation and Notes (Are these gaps deliberate or due to neglect?) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**2.4 Politics of Classification**

• Who defined the categories? Have definitions evolved over time?

**2.5 Power Asymmetries in Data Infrastructure**

• Who controls data collection and access? Do the affected have the ability to seek redress?

### Power Dynamics Mapping Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | Controller/Actor | Power Asymmetry | Ethical Risk | Ability to seek redress |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Section 3: Historical Bias Propagation, Data Representation, and Mitigation Matrix

Use this matrix to evaluate risks arising from historical patterns (Section 1) and data representation (Section 2).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Historical Bias** | **Data Representation** | **System Feature Affected** | **Likelihood** | **Severity** | **Ethical Risk Type** | **Mitigation Feasibility** | **Priority** |
| Redlining | ZIP code, home ownership | Loan thresholds | High | High | Deontological, Egalitarian | Medium | Critical |
| Gender bias | Job status, marital data | Credit score | Medium | High | Prioritarianism | Low | High |
| Informality bias | Missing income data | Feature completeness | High | Medium | Utilitarianism | Low | Medium |

# Section 4: Modern Manifestations of Historical Biases

**4.1 Historical to Technical**

How do historical patterns of bias in your context manifest today?

### Historical-to-Technical Mapping Table

|  |  |  |
| --- | --- | --- |
| Historical Pattern | Modern Mechanism | Feature or Process |
| Racial housing bias | Geo-proxies | ZIP code, Rent history |
| Gender employment disparity | Income gap | Job classification |
|  |  |  |
|  |  |  |

### 4.3 Proxy Discrimination Checklist

Identify and evaluate risks of proxy discrimination, where seemingly neutral variables act as stand-ins for protected attributes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Proxy For | Correlation Strength | Predictive Value | Action Taken |
| ZIP Code | Race | High | Medium | Replaced with distance to amenities |
| Marital Status | Gender | Moderate | Low | Removed |
| Employment Gap | Age | Low | High | Retained, monitored |

**4.4 Technological Amplification and Feedback Loop Mapping**

• How might automation amplify existing bias? Could system outputs influence future inputs (e.g., credit limit changes)?

### Feedback Loop Risk Analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Loop Type | Entry Point | Amplification Path | Risk Severity |
| Credit scores -> Loan approval -> Income recording | Biased score | System perpetuates exclusion | High |
| Geo-location -> Model trust score -> Outreach | Urban bias | Rural borrowers excluded | Medium |

# Section 5: Ethical Framework Alignment Matrix

Use the following matrix to evaluate how various ethical frameworks may interpret fairness in development of the system.

### Multi-Framework Analysis Table

|  |  |  |  |
| --- | --- | --- | --- |
| Ethical Framework | Fairness Test | Design Implications | Example in Your Context |
| Utilitarianism |  | Maximize total benefit | Optimize access without sacrificing accuracy |
| Prioritarianism |  | Prioritize worst-off groups | Adjust models for excluded populations |
| Egalitarianism |  | Promote equality of outcomes | Use re-weighting to ensure fairness |
| Deontological |  | Adhere to rights and duties | Avoid unfair exclusion even if effective |
| Virtue Ethics |  | Foster integrity and care | Embed empathy and responsibility in design |

Have the system ethical considerations been reviewed with input from affected populations?

### Stakeholder Mapping Table

|  |  |  |
| --- | --- | --- |
| Stakeholder Group | Ethical Preference | Conflict Zone |
| Borrowers | Prioritarianism | Model performance trade-offs |
| Executives | Utilitarianism | Risk tolerance vs fairness |
|  |  |  |
|  |  |  |

### Decision Table

Use this matrix to guide ethical decision-making by clearly outlining trade-offs between competing values observed in Stakeholder Mapping Table. Each option should be evaluated based on its benefit to fairness, potential ethical cost, and alignment with chosen ethical principles.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Option | Description | Ethical Benefit | Ethical Cost | Ethical Principle Prioritized | Final Choice? |
| A | Broad coverage with looser thresholds | Maximizes reach, promotes inclusion of underserved groups | May reduce precision, increasing false positives | Prioritarianism (uplift marginalized) | ✓ |
| B | Strict thresholds with high precision | Increases accuracy and minimizes error for included groups | Excludes edge cases, reinforces existing disparities | Utilitarianism (maximize efficiency) | ✗ |

**Summary of decison**:

* **Option A** was chosen because its ethical benefit—greater inclusivity for historically underserved communities—was prioritized despite a trade-off in precision.
* **Option B**, while technically appealing for its accuracy, was rejected due to its ethical cost of reinforcing systemic exclusion.

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

The selected fairness definition is…

## Mathematical Formulation

Mathematical Formulation is …

## Rationale

The choice of the above fairness definition stems from …

## Trade-off Acknowledgement

Bias Source Identification Tool

A Comprehensive Framework for Identifying and Mitigating Bias

# Introduction

The Bias Source Identification Tool (BSIT) provides a structured approach to identify, assess, and mitigate bias in various stages of system development and deployment. This document outlines the key sections of BSIT, including taxonomic classification, detection methodologies, prioritization framework, and user documentation.

# Taxonomic Classification of Bias Types

To systematically address bias, BSIT categorizes various sources of bias into distinct types, each defined by its indicators and accompanied by illustrative examples.

## Historical Bias

* **Definition:** Bias resulting from pre-existing social inequities, regardless of sampling or feature selection.
* **Indicators:**
  + Target variables reflecting historical discrimination.
  + Problematic correlations that mirror societal inequities.
  + Patterns that align with known historical discrimination.
* **Example:** A hiring algorithm trained on historical hiring decisions may perpetuate patterns of gender discrimination in technical roles.

## Sampling or Representation Bias

* **Definition:** Bias arising from how populations are sampled and measured in training data.
* **Indicators:**
  + Demographic imbalances compared to target population
  + Quality disparities across demographic groups
  + Systematic measurement differences
* **Example:** A medical diagnostic system trained primarily on data from young adult males may perform poorly for elderly female patients.

## Measurement Bias

* **Definition:** Bias arising from how features are selected, attributes are measured, proxied, or operationalized.
* **Indicators:**
  + Different measurement approaches across groups
  + Proxy variables with varying accuracy across populations
  + Inconsistent label quality across demographics
* **Example:** Using standardized test scores as a proxy for aptitude may disadvantage groups with less access to test preparation resources.

## Aggregation Bias

* **Definition:** Bias arising from combining distinct populations that may have different relationships between features and outcomes.
* **Indicators:**
  + One-size-fits-all models for heterogeneous populations
  + Features with different predictive relationships across groups
  + Unexplained performance disparities across subgroups
* **Example:** A credit scoring model might not account for different cultural approaches to credit usage, creating disparities across ethnic groups.

## Learning Bias

* **Definition:** Bias arising from model choices, and/or during model training that influence outputs that amplify disparities.
* **Indicators:**
  + Algorithms that overfit majority patterns
  + Regularization approaches that penalize minority patterns
  + Optimization objectives misaligned with fairness goals
* **Example:** A complex model might learn spurious correlations between protected attributes and outcomes that don't represent causal relationships.

## Evaluation Bias

* **Definition:** Bias arising from testing procedures that don't represent real-world performance or fairness.
* **Indicators:**
  + Test datasets with different characteristics than deployment contexts
  + Metrics that don't capture relevant fairness dimensions
  + Insufficient disaggregation of performance across groups
* **Example:** Evaluating a facial recognition system on a test set that doesn't include diverse skin tones will mask potential performance disparities in deployment.

## Deployment Bias

* **Definition:** Bias arising from how systems are implemented and used in practice.
* **Indicators:**
  + Context shifts between training and deployment
  + User interactions that reinforce biases
  + User experience that exclude minorities e.g. visually impaired
  + Feedback loops that amplify initial disparities
* **Example:** A recommendation system might create filter bubbles that limit exposure diversity based on initial demographic patterns.

# Bias Detection Methodology

Effective bias detection relies on analytical techniques tailored to each type of bias. The following outlines methods for identifying bias across the taxonomy:

## Historical Bias

* **Using Historical Context Assessment Tool:**
  + Extract documented discrimination patterns from the Historical Context Assessment results.
  + Identify specific historical mechanisms relevant to your application domain.
  + Reference the historical pattern risk classification to prioritize investigation.
* **Quantitative Techniques:**
  + Compare outcome distributions across groups identified as high-risk in the Historical Context Assessment.
  + Analyze correlations between system predictions and historical patterns documented in the assessment.
  + Test whether current data distributions match historically documented disparities.

## Sampling or Representation Bias

* **Using Historical Context Assessment Tool:**
  + Reference demographic groups identified as historically underrepresented in your domain.
  + Use historical documentation to establish appropriate population benchmarks.
  + Identify measurement approaches that have historically varied across groups.
* **Quantitative Techniques:**
  + Compare dataset demographic distribution to population benchmarks established from historical context.
    - Calculate statistical distance metrics (e.g., Kullback–Leibler divergence, Earth Mover's distance) between distributions of features across demographic groups.
    - Set acceptable thresholds based on domain-specific fairness requirements.
  + Calculate representation ratios and statistical significance of observed disparities.
    - Establish minimum representation thresholds for demographic intersections based on statistical power requirements.
    - Track improvements in representation through data augmentation or reweighting.
  + Analyze missing data patterns for correlation with protected attributes.
  + Assess data quality metrics across demographic groups identified in the historical assessment.
    - Assess feature validity across demographic groups through correlation analysis with ground truth when available.
    - Establish acceptable bounds for measurement differences between groups.
* **Qualitative Techniques:**
  + Document how samples were selected and what inclusion/exclusion criteria were applied.
  + Identify potential selection mechanisms that might create systematic under- or overrepresentation.
  + Analyze geographic, temporal, and contextual factors that influenced data collection.
  + Document distribution changes after bias mitigation interventions.

## Measurement Bias

* **Using Fairness Definition Selection Tool:**
  + Reference the selected fairness definitions to identify which measurement biases are most relevant.
  + For individual fairness definitions, focus on detecting inconsistent proxies across similar individuals.
  + For group fairness definitions, prioritize detecting systematic measurement differences across groups.
* **Quantitative Techniques:**
  + Test proxy variables for differential accuracy across groups based on your selected fairness criteria.
  + Analyze feature distributions to detect encoding schemes that create disparities.
  + Measure label consistency across annotators for different demographic groups.
* Qualitiative Techniques:
  + Document measurement improvements through alternative operationalization approaches.

## Aggregation Bias

* **Using Fairness Definition Selection Tool:**
  + Identify whether the selected fairness definitions assume uniform relationships across all demographics.
  + Determine whether performance disparities align with group-level heterogeneity assumptions.
  + Consider whether group-specific models or feature transformations might be more appropriate given population heterogeneity.
* **Quantitative Techniques:**
  + Disaggregate performance metrics (e.g., accuracy, precision, recall, F1) across demographic subgroups.
  + Use stratified analysis to compare predictive accuracy across combinations of protected attributes.
  + Identify features whose predictive relationships vary significantly across groups using interaction terms or group-wise regressions.
  + Use group fairness metrics (e.g., equal opportunity, equalized odds) for each distinct subgroup.
  + Compare global vs. group-specific model performance:train and evaluate both global models and group-specific models (per protected attribute).
  + Use statistical significance tests (e.g., likelihood ratio tests, ANOVA) to detect when group-specific modeling significantly improves performance.
* **Qualitative Techniques:**
  + Review data documentation for evidence of merged populations with divergent feature–label relationships.
  + Conduct expert reviews to determine whether population segments have distinct behavioral, cultural, or contextual dynamics.
  + Identify operational assumptions (e.g., uniform thresholds) that may not generalize across all groups.
  + Document whether modeling choices treat groups homogenously or allow for differentiated treatment.

## Learning Bias

* **Using Fairness Definition Selection Tool:**
  + Analyze model behavior specifically for violations of your selected fairness definitions.
  + For equal opportunity definitions, focus on false negative rate disparities.
  + For demographic parity definitions, examine overall prediction rate differences.
* **Quantitative Techniques:**
  + Measure model performance across demographic groups according to your selected fairness metrics.
  + Evaluate the chosen loss function.
    - Decompose performance metrics by demographic group to identify disparate optimization patterns.
    - Analyze convergence trajectories to determine whether minority group performance plateaus later than majority groups.
    - Test modified loss functions that give equal weight to examples regardless of group size.
    - Implement group-aware losses that explicitly balance performance across demographic categories.
  + Test regularization effects on minority group performance.
    - Compare feature importance across demographic groups before and after regularization.
    - Analyze how different regularization parameters affect performance disparities.
    - Implement group-specific regularization to account for different sample sizes.
  + Analyze model behavior against fairness constraints documented in your definition selection.
* **Qualitative Techniques:**
  + Document architecture-specific fairness implications to inform selection decisions.
  + Document how early stopping points affect the fairness-performance frontier.
  + Document trade-off frontiers to inform stakeholder discussions

## Evaluation Bias (mostly AI-generated)

* **Using Fairness Definition Selection Tool:**
  + Ensure that evaluation metrics align with the fairness definitions prioritized for your system.
  + Confirm whether group-specific fairness metrics (e.g., TPR, FPR by group) are included in model evaluation.
* **Quantitative Techniques:**
  + Compare evaluation dataset demographics to real-world deployment demographics.
  + Implement disaggregated evaluation that examines performance across both protected attributes and their intersections.
  + Develop statistical approaches appropriate for different group sizes.
    - Include metrics such as TPR, FPR, AUC, calibration curves, and confusion matrices across groups.
  + Create performance dashboards that highlight disparities across multiple metrics.
  + Establish minimum performance thresholds for all demographic groups rather than just in aggregate.
* **Qualitative Techniques:**
  + Document how evaluation datasets were sourced, cleaned, and sampled.
  + Compare evaluation and deployment environments (geography, technology, usage patterns).
  + Record whether evaluation metrics reflect downstream impacts (e.g., intervention thresholds).
  + Assess whether real-world feedback loops are captured in the evaluation process.
  + Document known blind spots or missing demographics in test datasets.

## Deployment Bias (mostly AI-generated)

* **Quantitative Techniques:**
  + Implement time-series tracking of fairness metrics across system iterations.
  + Calculate disparity growth rates to identify exponential amplification patterns.
  + Conduct counterfactual simulations that isolate feedback effects from other factors.
  + Measure distribution shifts in both feature spaces and outcome variables across demographic groups.
  + Implement targeted randomization to prevent self-reinforcing patterns in high-risk areas.
  + Implement A/B testing frameworks that compare system versions with different feedback intervention strategies.
  + Track long-term disparity evolution to verify intervention effectiveness.
* **Qualitative Techniques:**
  + Map all pathways through which system outputs might influence future inputs.
  + Classify identified feedback paths according to the feedback typology (direct, indirect, user-driven, system-driven).
  + Estimate potential disparity amplification risks for each path based on initial bias measurements.
  + Prioritize high-risk feedback paths for detailed monitoring and potential intervention.
  + Select appropriate intervention strategies based on feedback type and system constraints.
  + Design distribution monitoring and correction mechanisms that trigger automatically when shifts exceed thresholds.
  + Develop causal intervention approaches that modify specific feedback mechanisms without compromising overall system functionality.
  + Establish automated alerts for accelerating disparity growth rates.
  + Document observed feedback patterns to inform future system designs.

# Prioritization Framework

**Assessment Dimensions**

* **Severity:** Potential harm if the bias source remains unaddressed (1-5 scale).
* **Scope:** Proportion of decisions or individuals affected (1-5 scale).
* **Persistence:** Whether effects compound over time through feedback loops (1-5 scale).
* **Intervention Feasibility:** Relative ease of addressing the bias source (1-5 scale).
* **Historical Alignment:** Connection to historical patterns identified in Part 1 (1-5 scale).

**Priority Calculation**

Priority Score = (Severity × 0.3) + (Scope × 0.2) + (Persistence × 0.2) + (Historical Alignment × 0.2) + (Intervention Feasibility × 0.1)

**Priority Categories**

* **High Priority:** Score ≥ 4.0
* **Medium Priority:** 3.0 ≤ Score < 4.0
* **Low Priority:** Score < 3.0

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Bias Source** | **Severity (×0.3)** | **Scope (×0.2)** | **Persistence (×0.2)** | **Historical Alignment (×0.2)** | **Feasibility (×0.1)** | **Priority Score** | **Priority Category** |
| [e.g., Sampling Bias] | 4 × 0.3 = 1.2 | 5 × 0.2 = 1.0 | 3 × 0.2 = 0.6 | 5 × 0.2 = 1.0 | 2 × 0.1 = 0.2 | 1.2+1.0+0.6+1.0+0.2 = 4.0 | High |
| [Add another bias here] |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

# User Documentation

To ensure practical application of BSIT, user documentation offers step-by-step guidance and resources.

1. Review the key components of the AI system using the HCAT and FDST tools.
2. Familiarize yourself with the bias taxonomy to understand the types of biases.
3. Use the Bias Detection Methodology section to identify which biases are present in your system.
4. Apply the Prioritization Framework to assess which biases to address first.

* For each bias type identified, assign a score (1–5) for each dimension.
* Multiply each score by the weight indicated.
* Sum all weighted scores to obtain the total priority score.
* Use the Priority Categories above to classify the urgency.

1. Track mitigation interventions and outcomes using internal logs.
2. Update the system iteratively based on ongoing detection and evaluation cycles.

Fairness Metrics Tool

A Framework for Evaluating and Ensuring Fairness in AI Systems

# Introduction

The Fairness Metrics Tool provides a robust framework for selecting fairness metrics, validating them statistically, visualizing disparities, and reporting results effectively. It complements other tools, such as the Fairness Definition Selection Tool and Bias Source Identification Tool (BSIT.

# Section 1: Metric Selection Methodology

## 1.1 Determine Fairness Objective

This section begins by extracting the fairness goal using the Fairness Definition Selection Tool. The table below outlines various fairness objectives, their corresponding definitions, metric targets, and suitable applications:

|  |  |  |  |
| --- | --- | --- | --- |
| Fairness Objective | Fairness Definition | Metric Target | Suitable For |
| Equal outcomes | Demographic Parity | Statistical Parity | Outreach programs, public services |
| Equal treatment of qualified | Equal Opportunity | TPR Parity | Hiring, admissions |
| Balanced error impact | Equalized Odds | TPR + FPR Parity | Criminal justice, healthcare |
| Equal predictive reliability | Predictive Parity | PPV Parity | Medical diagnosis, insurance |
| Treat similar individuals equally | Individual Fairness | Lipschitz-based Distance | Credit scoring, housing |
| Fair across hypothetical changes | Counterfactual Fairness | Invariance under Sensitive Attribute Changes | Causal modeling, social science |

## 1.2 Base Rate Sensitivity

* If base rates are similar: Demographic Parity is the most viable metric.
* If base rates differ: Prioritize Equal Opportunity or Predictive Parity for evaluation.

## 1.3 Intersectional Considerations

* Extend metrics to account for multiple protected attributes (e.g., race × gender).
* Use subgroup disaggregation and conduct intersectional fairness audits.

## 1.4 Quantitative Metric Violations

Below are formulas for common fairness metric violations:

|  |  |  |
| --- | --- | --- |
| Fairness Definition | Metric Name | Violation Formula |
| Demographic Parity | Statistical Parity Difference | P(Ŷ=1 | A=a) - P(Ŷ=1 | A=b) |
| Equal Opportunity | Equal Opportunity Difference | P(Ŷ=1 | Y=1, A=a) - P(Ŷ=1 | Y=1, A=b) |
| Equalized Odds | TPR, FPR Difference | TPR: P(Ŷ=1 | Y=1, A=a) - P(Ŷ=1 | Y=1, A=b)<br></br>FPR: P(Ŷ=1 | Y=0, A=a) - P(Ŷ=1 | Y=0, A=b) |
| Predictive Parity | Predictive Parity Difference | P(Y=1 | Ŷ=1, A=a) - P(Y=1 | Ŷ=1, A=b) |

# Section 2: Statistical Validation Framework

## 2.1 Confidence Intervals

* Use bootstrapping, jackknife, or binomial approximations to calculate 95% confidence intervals.
* Handle small samples (n < 100) with Bayesian estimation and report credible intervals. Flag underpowered groups clearly in results.

## 2.2 Hypothesis Testing

* Null hypothesis: No disparity exists.
* Statistical tests: Chi-square test, t-test, Fisher’s exact test for small samples, and Benjamini–Hochberg for multiple comparisons.

## 2.3 Robustness Checks

* Validate metrics across data folds using cross-validation.
* Perform threshold sensitivity analysis and evaluate robustness under label noise or shifted feature distributions.

## 2.4 Threshold-Based Flagging

|  |  |  |
| --- | --- | --- |
| Metric | Suggested Threshold | Action |
| Statistical Parity Diff | ≤ 0.10 | Acceptable difference |
| Equal Opportunity Diff | ≤ 0.10 | Indicates fair TPR balance |
| TPR / FPR Difference | ≤ 0.10 | Balanced error across groups |
| Predictive Parity Diff | ≤ 0.10 | Reliable predictions across groups |

# Section 3: Visualization & Reporting Templates

## 3.1 Metric Dashboard

* Bar charts by group for TPR, FPR, PPV.
* Confidence intervals visualized as error bars.
* Trend lines to track disparities over time.

## 3.2 Disparity Heatmaps

* Create color-coded matrices of metric values across intersections.
* Use red-yellow-green coding to denote severity of violations.

## 3.3 Summary Tables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Group A | Group B | Diff | 95% CI | p-value | Threshold Met? |
| TPR | 0.85 | 0.70 | 0.15 | (0.08–0.22) | 0.002 | ❌ |

# Section 4: User Documentation

## 4.1 When to Use

* Model evaluation
* Pre-deployment audit
* Post-deployment monitoring
* Fairness or compliance reporting
* Evaluation of intervention effectiveness

## 4.2 How to Apply the Tool

1. Identify your system’s fairness goals.
2. Select corresponding fairness metrics using Section 1.
3. Calculate metric values across relevant groups.
4. Quantify violations and compare to thresholds.
5. Validate statistically using confidence intervals and hypothesis testing.
6. Evaluate metric robustness using Section 2.3.
7. Visualize results and disparities.
8. Report findings and recommend next steps.

## 4.3 Integration Points

* Model selection and evaluation phase
* Part of CI/CD pipelines
* Integrated with BSIT and other bias detection frameworks
* Used to evaluate fairness after interventions

## 4.4 Violation Interpretation

* Metrics exceeding thresholds indicate fairness risk.
* Prioritize mitigation for most violated metrics.
* Re-evaluate metrics post-intervention.
* Document trade-offs made when fairness metrics conflict.