

# Module 12: Fairness and Responsible AI

## Overview

Machine learning systems are increasingly used for high-stakes decisions: hiring, lending, criminal justice, healthcare. These systems can perpetuate or even amplify existing societal biases. This module explores sources of algorithmic bias, formal definitions of fairness, measurement techniques, and mitigation strategies. You will grapple with the inherent tensions between different fairness criteria and consider the broader ethical implications of deploying ML systems.

---

## 1. Why Fairness Matters

### High-Stakes Applications

- **Hiring:** Resume screening, interview scheduling
- **Lending:** Credit scoring, loan approval
- **Criminal Justice:** Recidivism prediction, sentencing
- **Healthcare:** Diagnosis, treatment recommendations
- **Advertising:** Job ads, housing ads targeting

### Real-World Examples of Bias

- COMPAS recidivism algorithm: Higher false positive rates for Black defendants
  - Amazon hiring tool: Penalized resumes with “women’s” activities
  - Facial recognition: Higher error rates for darker-skinned faces
  - Healthcare algorithm: Underestimated Black patients’ needs
- 

## 2. Sources of Bias

### Data-Level Bias

- **Historical bias:** Data reflects past discrimination
- **Representation bias:** Groups under/overrepresented
- **Measurement bias:** Features measured differently across groups
- **Sampling bias:** Non-representative data collection

### Algorithm-Level Bias

- **Aggregation bias:** Single model for different populations
- **Objective function:** Optimizing for accuracy may not mean fairness
- **Feature selection:** Proxies for protected attributes

### Human-in-the-Loop Bias

- Annotation bias in labels
- Confirmation bias in interpretation

- Deployment context differs from training
- 

### 3. Protected Attributes

#### Common Protected Classes

- Race/ethnicity
- Gender/sex
- Age
- Religion
- National origin
- Disability status

#### Legal Context

Protected classes vary by jurisdiction and application: - US: Civil Rights Act, Equal Credit Opportunity Act - EU: GDPR has some fairness provisions - Sector-specific: Healthcare, employment, housing

#### The Proxy Problem

Even without protected attributes as features, models can learn to predict them from proxies: - ZIP code  $\square$  race/income - Name  $\square$  gender/ethnicity - Purchasing patterns  $\square$  various attributes

---

### 4. Fairness Definitions

#### Demographic Parity (Statistical Parity)

Equal positive prediction rates across groups:

$$P(\hat{Y}=1 \mid A=0) = P(\hat{Y}=1 \mid A=1)$$

**Pros:** Simple, directly addresses disparate impact **Cons:** May require different thresholds, ignores base rate differences

#### Equalized Odds

Equal true positive rates AND false positive rates across groups:

$$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1) \quad (\text{TPR})$$

$$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1) \quad (\text{FPR})$$

**Pros:** Conditions on true outcome, reduces harm from false positives **Cons:** May be impossible to achieve perfectly

## Equal Opportunity

Relaxed version of equalized odds—only requires equal TPR:

$$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1)$$

**Pros:** Qualified individuals in each group have equal chance **Cons:** Doesn't address false positives

## Calibration

Equal accuracy of probability estimates across groups:

$$P(Y=1 \mid \hat{Y}=p, A=0) = P(Y=1 \mid \hat{Y}=p, A=1) = p$$

For the same predicted probability, actual positive rate should be the same.

---

## 5. Impossibility Results

### The Fairness Tradeoff

**Except in special cases, you cannot simultaneously satisfy:** - Calibration - Equal false positive rates - Equal false negative rates

When base rates differ between groups ( $P(Y=1|A=0) \neq P(Y=1|A=1)$ ), these criteria conflict.

### Implications

- Must choose which fairness definition matters most for your context
  - No single “correct” definition—depends on values and application
  - Tradeoffs are fundamental, not just technical limitations
- 

## 6. Measuring Fairness

### Group-Level Metrics

For each group  $A \in \{0, 1\}$ : - Positive rate:  $P(\hat{Y}=1 \mid A)$  - True positive rate:  $P(\hat{Y}=1 \mid Y=1, A)$  - False positive rate:  $P(\hat{Y}=1 \mid Y=0, A)$  - Precision:  $P(Y=1 \mid \hat{Y}=1, A)$

### Disparity Ratios

Compare metrics between groups:

$$\text{Demographic Parity Ratio} = P(\hat{Y}=1|A=1) / P(\hat{Y}=1|A=0)$$

The 80% rule: Ratios below 0.8 indicate potential disparate impact.

---

## 7. Bias Mitigation

### Pre-processing

Modify training data before learning: - Resampling: Balance representation - Reweighting: Give underrepresented groups more weight - Representation learning: Learn fair embeddings

### In-processing

Modify the learning algorithm: - Constrained optimization: Add fairness constraints - Adversarial learning: Prevent prediction of protected attribute - Regularization: Penalize unfair predictions

### Post-processing

Modify predictions after learning: - Threshold adjustment: Different thresholds per group - Calibration: Adjust probabilities to be group-calibrated - Reject option: Abstain in uncertain cases

---

## 8. Beyond Binary Fairness

### Intersectionality

Fairness at group level may hide problems at intersections: - Fair for women overall, but not for Black women specifically - Need to consider multiple attributes simultaneously

### Individual Fairness

Similar individuals should receive similar predictions:

If  $d(x_1, x_2)$  is small, then  $d(\hat{y}_1, \hat{y}_2)$  should be small

Requires defining appropriate similarity metric.

### Counterfactual Fairness

Would the prediction change if the protected attribute were different?

Requires causal modeling of how attributes relate.

---

## 9. Practical Considerations

### When Perfect Fairness Is Impossible

- Document tradeoffs explicitly
- Choose criteria based on context and values
- Be transparent about limitations

## Continuous Monitoring

- Fairness can degrade over time (distribution shift)
- Regularly audit deployed models
- Collect data on outcomes, not just predictions

## Stakeholder Engagement

- Affected communities should have input
  - Technical definitions alone are insufficient
  - Consider broader systemic impacts
- 

## Key Takeaways

1. **Bias** can enter ML systems through data, algorithms, and deployment
  2. **Protected attributes** may be learned from proxies
  3. **Multiple fairness definitions** exist with different implications
  4. **Fairness criteria conflict** when base rates differ
  5. **Mitigation** can occur at pre-processing, in-processing, or post-processing stages
  6. **Context matters**: No universal fairness definition
  7. **Monitoring and engagement** are ongoing requirements
- 

## Connections to Other Modules

- **Module 5**: Metrics beyond accuracy
  - **Module 11**: Evaluation methodologies
- 

## Further Reading

- [Links to be added]
- Fairness and Machine Learning ([fairmlbook.org](http://fairmlbook.org))
- AI Fairness 360 documentation