Algorithmic Bias

Problems	Solutions
Sources of bias	Mitigation strategies
Health disparities	Continuous monitoring
Fairness metrics	Regular auditing

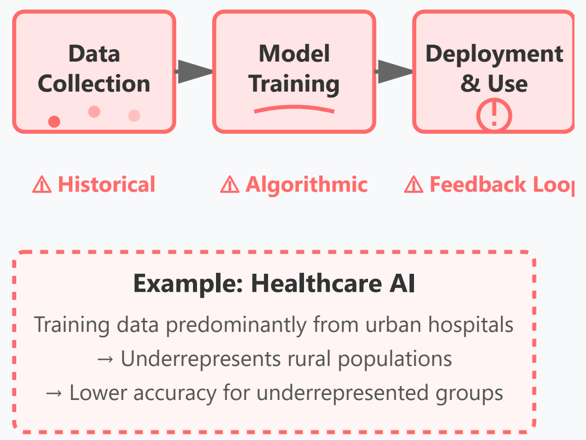
Problems: Understanding Algorithmic Bias

Sources of Bias

Algorithmic bias originates from multiple sources throughout the machine learning pipeline. These biases can be introduced during data collection, model training, or deployment phases, often reflecting and amplifying existing societal inequalities.

Common Sources:

- → **Historical Bias:** Training data reflecting past discriminatory practices
- → **Representation Bias:** Underrepresentation of certain demographic groups in datasets
- → **Measurement Bias:** Inconsistent or biased data collection methods across populations
- → **Aggregation Bias:** Inappropriate grouping that obscures important subgroup differences



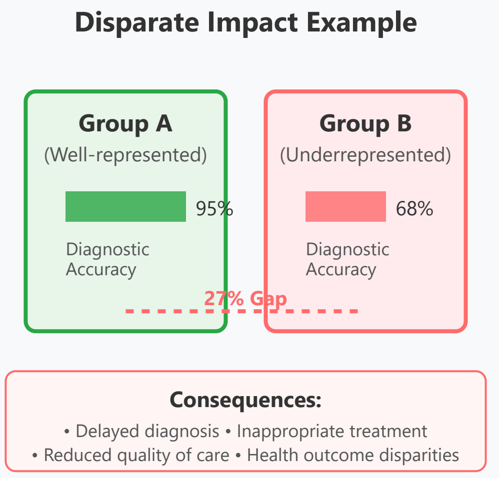
Bias can enter at any stage of the ML pipeline

Health Disparities

Algorithmic bias in healthcare can exacerbate existing health disparities, leading to unequal access to care, misdiagnosis, or inappropriate treatment recommendations for marginalized populations. These systems may perpetuate systemic inequalities if not carefully designed and monitored.

Real-World Impact:

- → **Risk Prediction:** Algorithms using healthcare costs as a proxy may underestimate illness severity in communities with less access to care
- → **Diagnostic Tools:** Image recognition systems trained primarily on certain skin tones may perform poorly on others
- → **Resource Allocation:** Biased predictions can lead to inequitable distribution of medical resources and interventions
- → **Clinical Trials:** Underrepresentation in research data affects treatment efficacy predictions



Performance gaps can lead to worse health outcomes

Fairness Metrics

Measuring fairness in algorithms is complex and multifaceted. Different fairness metrics may conflict with each other, and choosing the appropriate metric depends on the specific context and stakeholder values. No single metric can capture all aspects of fairness.

Key Fairness Metrics:

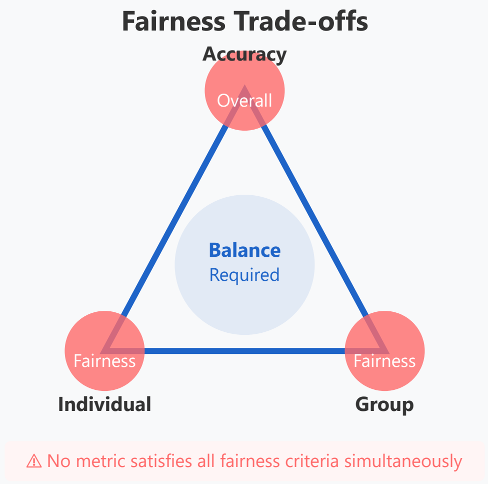
→ **Demographic Parity:** Equal positive prediction rates across groups

→ **Equal Opportunity:** Equal true positive rates for all groups

→ **Equalized Odds:** Equal true positive and false positive rates across groups

→ **Calibration:** Predictions are equally accurate across all groups

→ **Individual Fairness:** Similar individuals receive similar predictions

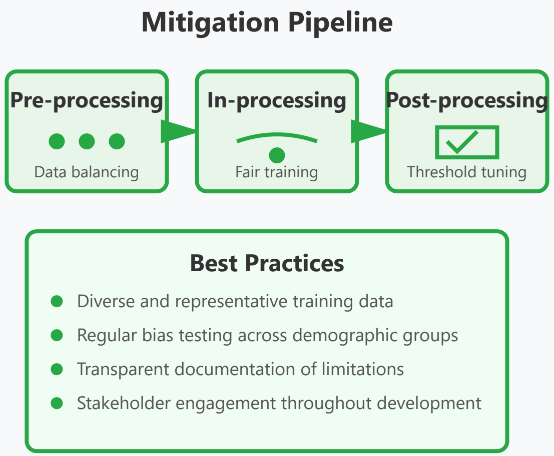


Different fairness metrics often involve trade-offs

Solutions: Addressing Algorithmic Bias

Mitigation Strategies

Effective bias mitigation requires a comprehensive approach that addresses issues at every stage of the ML pipeline. These strategies should be implemented proactively during design and development, rather than as afterthoughts following deployment.



Multi-stage approach to bias mitigation

Mitigation Approaches:

→ **Pre-processing:** Resampling, reweighting, or augmenting training data to ensure balanced representation

→ **In-processing:** Incorporating fairness constraints directly into model training objectives

→ **Post-processing:** Adjusting model predictions to meet fairness criteria

→ **Diverse Teams:** Including stakeholders from affected communities in design and review processes

Continuous Monitoring

Bias can emerge or evolve over time due to changing data distributions, shifting societal contexts, or feedback loops. Continuous monitoring ensures that deployed systems maintain fairness and allows for timely intervention when issues arise.

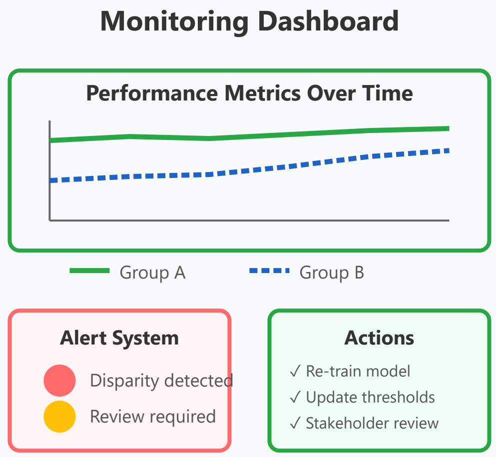
Monitoring Components:

→ **Performance Metrics:** Track accuracy, precision, and recall across demographic groups over time

→ **Disparity Detection:** Automated alerts when fairness metrics exceed acceptable thresholds

→ **User Feedback:** Systematic collection and analysis of complaints and concerns

→ **Data Drift:** Monitor changes in input data distributions that may affect fairness



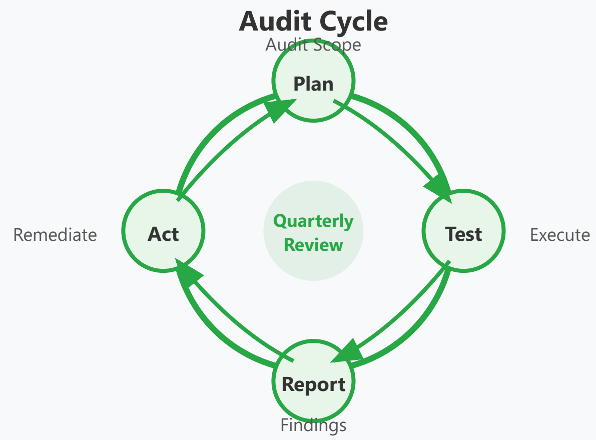
Real-time monitoring enables rapid response to emerging issues

Regular Auditing

Regular audits provide comprehensive evaluations of algorithmic systems, examining technical performance, fairness outcomes, and compliance with ethical standards. Both internal and external audits help ensure accountability and build public trust.

Audit Components:

- → **Technical Audit:** Comprehensive testing of model performance across all relevant subgroups
- → **Impact Assessment:** Evaluation of real-world effects on affected populations
- → **Compliance Review:** Verification of adherence to regulations and ethical guidelines
- → **Documentation Audit:** Review of decision-making processes and transparency materials



Continuous improvement through regular audit cycles