Ethics in Machine Learning

Bias, Fairness, Interpretability, and Responsible AI

Sarwan Ali

Department of Computer Science
Georgia State University

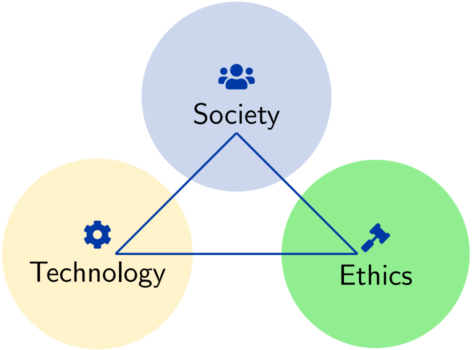
⚖ Building Ethical AI Systems ⚖

Today's Learning Journey

1. Introduction to AI Ethics
2. Understanding Bias in ML
3. Fairness in Machine Learning
4. Interpretability and Explainability
5. Responsible AI Practices
6. Practical Implementation
7. Case Studies
8. Future Directions
9. Conclusion

Why Ethics in Machine Learning?

- **Growing Impact:** ML systems affect millions of lives
- **Automated Decisions:** Systems make critical choices about people
- **Societal Trust:** Public confidence in AI technology
- **Legal Requirements:** Emerging regulations worldwide
- **Business Value:** Ethical AI reduces risks and builds reputation



Real-World Ethical Challenges

Healthcare

- Diagnostic bias in medical imaging
- Treatment recommendation fairness
- Patient privacy protection

Criminal Justice

- Risk assessment algorithms
- Predictive policing bias
- Sentencing recommendations

Employment

- Resume screening algorithms
- Performance evaluation systems
- Workplace surveillance

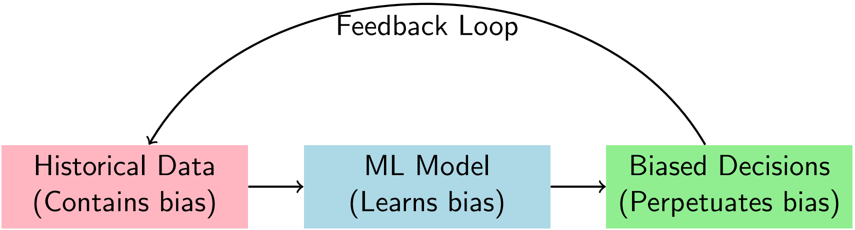
Finance

- Credit scoring fairness
- Insurance premium calculation
- Algorithmic trading impact

What is Bias in Machine Learning?

Definition

Bias in ML refers to systematic errors or unfair discrimination that occurs when algorithms consistently favor certain groups or outcomes over others.



Types of Bias in ML Systems

Data-Related Bias

- **Historical Bias:** Past discrimination in data
- **Representation Bias:** Underrepresented groups
- **Measurement Bias:** Systematic data collection errors
- **Sampling Bias:** Non-representative samples

Algorithmic Bias

- **Confirmation Bias:** Seeking confirming evidence
- **Selection Bias:** Biased feature selection
- **Evaluation Bias:** Inappropriate metrics
- **Deployment Bias:** Misuse of models

Key Insight

Bias can enter at **any stage** of the ML pipeline: data collection, preprocessing, model training, evaluation, and deployment.

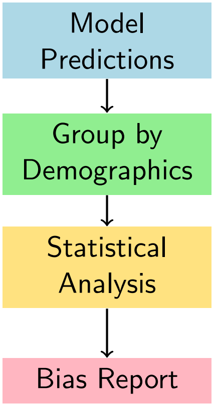
Bias Detection Techniques

Statistical Methods

- Demographic parity analysis
- Equalized odds testing
- Calibration analysis
- Disparate impact assessment

Visualization Techniques

- Confusion matrices by group
- ROC curves comparison
- Distribution plots
- Fairness dashboards



Defining Fairness

The Challenge

There is **no single definition** of fairness that works for all contexts. Different fairness criteria can be **mathematically incompatible**.

Individual Fairness

Similar individuals should receive similar treatment

Group Fairness

Statistical parity across different groups

Counterfactual Fairness

Decisions unchanged in counterfactual world

Mathematical Fairness Metrics

Demographic Parity

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

Equalized Odds

$$P(\hat{Y} = 1 | Y = y, A = 0) = P(\hat{Y} = 1 | Y = y, A = 1) \quad \forall y \in \{0, 1\}$$

Equal Opportunity

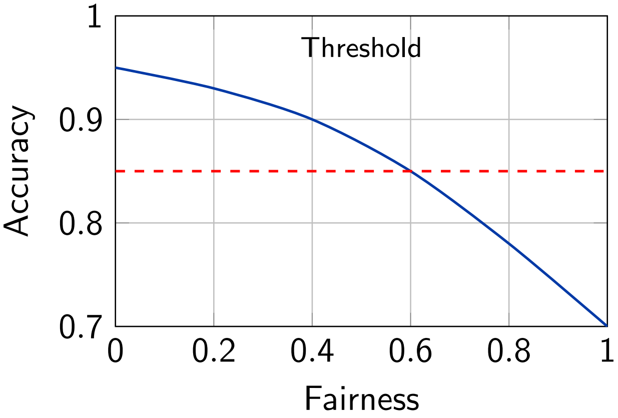
$$P(\hat{Y} = 1 | Y = 1, A = 0) = P(\hat{Y} = 1 | Y = 1, A = 1)$$

Where: $\hat{Y}$ = prediction, Y = true label, A = protected attribute

Important Note

These metrics can be **mutually exclusive** - satisfying one may violate another!

Fairness-Accuracy Trade-offs



Key Considerations:

- Perfect fairness may reduce accuracy
- Context determines acceptable trade-offs
- Stakeholder input is crucial
- Multiple models may be needed

Bias Mitigation Strategies

Pre-processing

- Data augmentation
- Re-sampling techniques
- Feature selection
- Synthetic data generation

Clean the data

In-processing

- Fairness constraints
- Adversarial training
- Multi-objective optimization
- Regularization terms

Fair training

Post-processing

- Threshold optimization
- Calibration adjustment
- Output modification
- Fairness-aware ensembles

Adjust outputs

Why Do We Need Interpretable ML?

Trust and Transparency

- Understanding model decisions
- Building user confidence
- Regulatory compliance

Debugging and Improvement

- Identifying model errors
- Feature importance analysis
- Model refinement

Accountability

- Legal requirements
- Ethical responsibility
- Risk management

Domain Knowledge

- Scientific discovery
- Medical diagnosis
- Business insights

Black box models vs **Interpretable models**

Interpretability vs Explainability

Interpretability

Intrinsic - The degree to which a human can understand the cause of a decision

Examples:

- Linear regression
- Decision trees
- Simple rule-based systems

Explainability

Post-hoc - Techniques to explain decisions made by complex models

Examples:

- LIME, SHAP
- Attention maps
- Saliency maps

Simple Model (Interpretable) $\xleftrightarrow{\text{Trade-off}}$ Complex Model + Explanation Tool

Global vs Local Explanations

Global Explanations

- Explain the **entire model**
- Overall feature importance
- Model behavior patterns
- Decision boundaries

Techniques:

- Permutation importance
- Partial dependence plots
- Feature interaction analysis

Local Explanations

- Explain **individual predictions**
- Instance-specific reasoning
- Feature contributions
- Counterfactual examples

Techniques:

- LIME (Local Interpretable Model-agnostic Explanations)
- SHAP (SHapley Additive exPlanations)
- Counterfactual explanations

Popular Explainability Techniques

SHAP (SHapley Additive exPlanations)

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f(S \cup \{i\}) - f(S)]$$

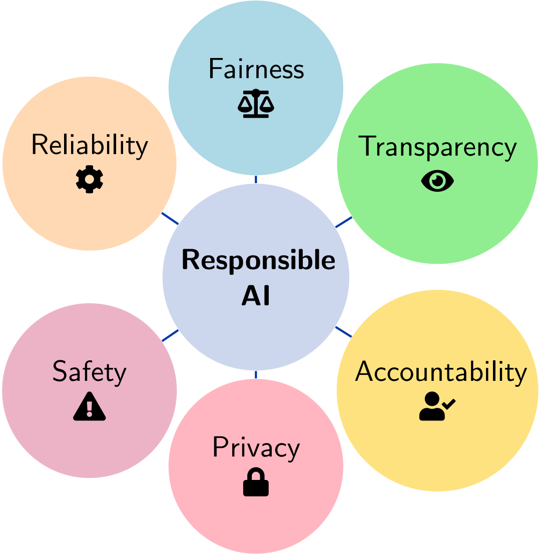
Properties:

- Efficiency: $\sum \phi_i = f(x) - E[f(X)]$
- Symmetry: Equal contribution for equal features
- Dummy: Zero contribution for irrelevant features
- Additivity: Consistent across models

LIME Approach:

1. Perturb input around instance
2. Get predictions for perturbations
3. Weight by proximity to original
4. Fit interpretable model locally

Principles of Responsible AI



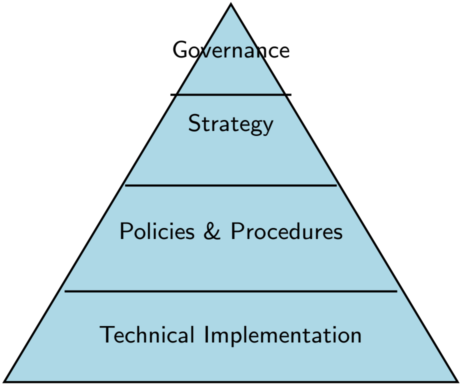
AI Governance Framework

Organizational Level

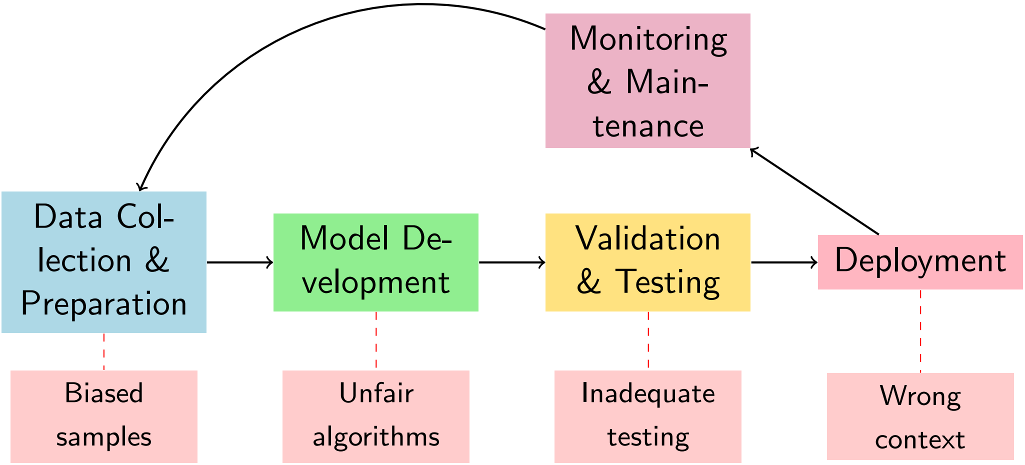
- AI ethics committee
- Clear policies and guidelines
- Regular audits and assessments
- Training and awareness programs

Technical Level

- Bias testing frameworks
- Explainability requirements
- Performance monitoring
- Continuous validation



ML Lifecycle Risk Management



Regulatory Landscape

Existing Regulations

- GDPR (EU) - Right to explanation
- CCPA (California) - Data privacy
- Fair Credit Reporting Act (US)
- Equal Employment Opportunity laws

Emerging Frameworks

- EU AI Act
- Algorithmic Accountability Act (US)
- IEEE Standards for AI
- ISO/IEC 23053 (AI Risk Management)

Key Requirements

- Risk assessment documentation
- Bias testing and mitigation
- Human oversight mechanisms
- Transparency and explainability
- Data protection and privacy
- Regular auditing and monitoring

Compliance Strategy

Stay informed about regulations in your domain and jurisdiction!

Building an Ethical AI Checklist

Pre-Development

- ☑ Define ethical requirements
- ☑ Assess potential harms
- ☑ Stakeholder consultation
- ☑ Data quality audit
- ☑ Bias risk assessment

Development

- ☑ Diverse development team
- ☑ Fairness metrics integration
- ☑ Explainability requirements
- ☑ Privacy-preserving techniques

Testing & Validation

- ☑ Bias testing across groups
- ☑ Adversarial testing
- ☑ Edge case analysis
- ☑ Performance disparities check
- ☑ Explanation quality assessment

Deployment & Monitoring

- ☑ Continuous monitoring system
- ☑ Performance degradation alerts
- ☑ Feedback mechanisms
- ☑ Regular model retraining
- ☑ Incident response procedures

⚠ **Remember:** Ethics is not a one-time check, but an ongoing process!

Tools and Frameworks for Ethical AI

Bias Detection & Mitigation

- **Fairlearn:** Microsoft's fairness toolkit
- **AIF360:** IBM's AI Fairness 360
- **What-If Tool:** Google's model analysis
- **Aequitas:** Bias audit toolkit

Explainability

- **SHAP:** Game theory-based explanations
- **LIME:** Local interpretable explanations
- **InterpretML:** Microsoft's interpretability
- **Captum:** PyTorch model interpretability

Privacy & Security

- **Differential Privacy:** TensorFlow Privacy
- **Federated Learning:** TensorFlow Federated
- **Homomorphic Encryption:** Microsoft SEAL
- **Secure Multi-party Computation**

Governance & Monitoring

- **MLflow:** ML lifecycle management
- **Weights & Biases:** Experiment tracking
- **TensorBoard:** Model monitoring
- **ModelDB:** Model versioning & governance

Code Example: Bias Detection with Fairlearn

```
from fairlearn.metrics import MetricFrame, selection_rate
from fairlearn.postprocessing import ThresholdOptimizer
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
# Train your model
model = RandomForestClassifier()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
# Analyze fairness metrics
metric_frame = MetricFrame(
    metrics={
        'accuracy': accuracy_score,
        'selection_rate': selection_rate,
        'true_positive_rate': true_positive_rate},
    y_true=y_test,
    y_pred=y_pred,
    sensitive_features=sensitive_attr
)
print("Overall metrics:")
print(metric_frame.overall)
print("\nBy group:")
print(metric_frame.by_group)
# Post-processing for fairness
postprocess_est = ThresholdOptimizer(
    estimator=model,
    constraints="equalized_odds",
    prefit=True)
postprocess_est.fit(X_train, y_train, sensitive_features=sensitive_train)
fair_predictions = postprocess_est.predict(X_test, sensitive_features=sensitive_test)
```

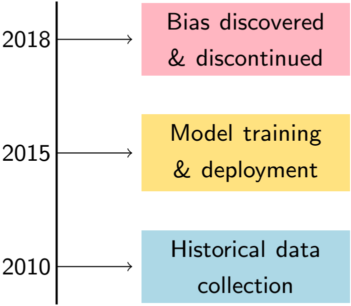
Code Example: SHAP Explanations

```python
import shap
import matplotlib.pyplot as plt
# Initialize SHAP explainer
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X_test)
# Global feature importance
shap.summary_plot(shap_values, X_test, feature_names=feature_names)
# Local explanation for single instance
instance_idx = 0
shap.waterfall_plot(
    explainer.expected_value[1],
    shap_values[1][instance_idx],
    X_test.iloc[instance_idx],
    feature_names=feature_names)
# Feature interaction analysis
shap.plots.partial_dependence(
    "feature_1", model.predict, X_train, ice=False,
    model_expected_value=True, feature_expected_value=True)
# Check for bias in SHAP explanations
shap_df = pd.DataFrame(shap_values[1], columns=feature_names)
shap_df['sensitive_attr'] = sensitive_test
# Compare average SHAP values by sensitive attribute
bias_analysis = shap_df.groupby('sensitive_attr').mean()
print("Average SHAP values by sensitive attribute:")
print(bias_analysis)
```

Case Study 1: Hiring Algorithm Bias

The Problem

- Large tech company's resume screening AI
- Trained on 10 years of historical hiring data
- Systematically downgraded resumes with "women's" keywords
- Learned from biased historical decisions



Root Causes

- Historical gender bias in tech hiring
- Insufficient diverse representation in training data
- Lack of fairness constraints during training

Lessons Learned

- Audit training data for historical biases
- Implement fairness metrics from the start
- Regular testing with diverse evaluation sets
- Human oversight in high-stakes decisions

Case Study 2: Healthcare AI Racial Bias

The Scenario

- Algorithm predicting healthcare needs
- Used healthcare spending as proxy for health needs
- Significantly underestimated Black patients' needs
- Affected millions of patients

The Bias Mechanism

- Healthcare spending $\neq$ Healthcare needs
- Structural inequalities in healthcare access
- Socioeconomic factors affecting spending

Solutions Implemented

- Changed target variable to actual health outcomes
- Included multiple health indicators
- Tested for racial disparities in predictions
- Continuous monitoring post-deployment

Impact

After correction, the percentage of Black patients identified for extra care increased from 17.7% to 46.5%

Critical: Choice of target variable can embed societal biases

Case Study 3: Criminal Justice Risk Assessment

COMPAS Algorithm Analysis

- Predicts likelihood of reoffending
- Used in sentencing and parole decisions
- ProPublica investigation revealed racial bias
- Higher false positive rates for Black defendants

Fairness Dilemma

- Algorithm satisfied **calibration**
- Failed **equalized odds**
- Mathematical impossibility to satisfy both
- Different stakeholders prefer different metrics

Confusion Matrix

	Actual	Predicted
	Low Risk	High Risk
Low	TN	FP
High	FN	TP

Higher FP rate for minorities

Key Takeaway

Context matters! Different applications may require different fairness criteria. Stakeholder input is crucial for determining appropriate trade-offs.

Emerging Trends in AI Ethics

Technical Advances

- Causal fairness approaches
- Federated learning for privacy
- Automated bias detection
- Adversarial debiasing techniques
- Uncertainty quantification

Methodological Innovations

- Participatory design approaches
- Intersectional fairness metrics
- Dynamic fairness adaptation
- Multi-stakeholder optimization

Societal Developments

- Algorithmic auditing standards
- AI ethics certification programs
- Cross-cultural fairness research
- Public participation in AI governance

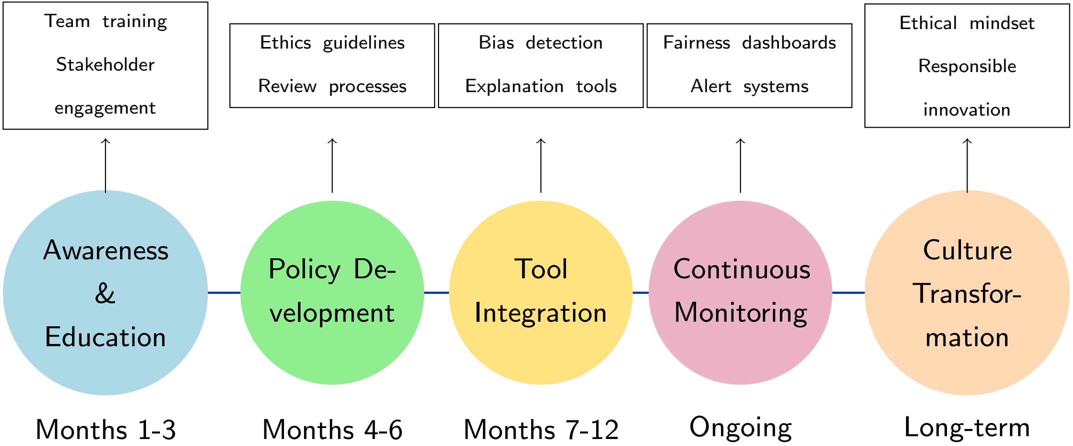
Regulatory Evolution

- Sector-specific AI regulations
- International AI governance frameworks
- Rights-based approaches to AI
- Liability and accountability laws

Challenges and Open Questions

1. **Fairness Trade-offs**
    - How to balance competing fairness criteria?
    - Who decides what constitutes "fair"?
    - Cultural and contextual variations in fairness
2. **Scalability**
    - Efficient bias detection for large-scale systems
    - Real-time fairness monitoring
    - Automated ethical decision-making
3. **Explainability vs Performance**
    - Can we have both high accuracy and interpretability?
    - Quality of explanations for non-experts
    - Cognitive biases in interpreting explanations
4. **Global Coordination**
    - Harmonizing ethical standards across cultures
    - Preventing regulatory arbitrage
    - Ensuring inclusive participation in standard-setting

Building Ethical AI: A Roadmap



Key Takeaways

Core Principles

- Ethics is not optional in AI development
- Bias can enter at any stage of ML pipeline
- Multiple fairness definitions exist and may conflict
- Explainability enhances trust and accountability
- Continuous monitoring is essential

Practical Actions

- Develop ethical AI checklists
- Use bias detection and mitigation tools
- Implement explainability from the start
- Establish governance frameworks
- Stay informed about regulations

Remember: Building ethical AI is not just a technical challenge—it's a societal responsibility that requires interdisciplinary collaboration and ongoing commitment.

Questions & Discussion

Thank you for your attention!

sali85@student.gsu.edu

Building a more ethical future with AI