AI Ethics and Bias Mitigation

Building Responsible and Fair Artificial Intelligence Systems

Introduction

AI Ethics ensures technology aligns with human values such as fairness, accountability, and transparency. Bias in AI can lead to discrimination, misinformation, or reputational harm. Ethical AI aims to be transparent, inclusive, and safe.

Why AI Ethics Matters

AI systems influence hiring, healthcare, finance, and law enforcement. Without ethical considerations, models may amplify societal inequalities and reduce public trust in AI.

Pillars of AI Ethics

Fairness: Avoid discrimination
Transparency: Explainable and auditable decisions
Accountability: Defined ownership of outcomes
Privacy: Data protection and consent
Safety: Prevention of harm and misuse

Sources of Bias in AI

- Data bias: Skewed or non-representative data
- Labeling bias: Subjective or erroneous annotations
- Algorithmic bias: Optimization favoring certain groups
- Societal bias: Inherited cultural or systemic inequalities
- Measurement bias: Incorrect proxies for fairness

Real-World Examples

• COMPAS criminal justice tool – racial bias
• Facial recognition misidentifying people of color
• Amazon’s AI recruiting favoring male candidates
• GPT models producing gendered stereotypes

Bias Detection Techniques

Data audits, fairness metrics, and explainability tools help identify bias.
Example:
```
from fairlearn.metrics import demographic_parity_difference
metric = demographic_parity_difference(y_true, y_pred, sensitive_features=gender)
```

Bias Mitigation Approaches

Pre-processing: Re-sampling, data balancing
In-processing: Fair loss functions, adversarial debiasing

Post-processing: Re-ranking and output calibration

Fairness-Aware Algorithms

Adversarial debiasing minimizes correlation between bias attributes and predictions.
Reweighing assigns balanced sample weights.
Example:

```
from aif360.algorithms.preprocessing import Reweighing
rw = Reweighing(unprivileged_groups=[{'gender':0}], privileged_groups=[{'gender':1}])
dataset_transf = rw.fit_transform(dataset)
```

Explainable AI (XAI)

Interpretability improves trust and accountability.
Popular techniques:
- SHAP (Shapley values)
- LIME (Local Interpretable Model Explanations)
- Counterfactual explanations

Ethical AI Frameworks

• OECD AI Principles (2019)
• EU AI Act (2024)
• NIST AI RMF
• UNESCO AI Ethics Recommendations
• NITI Aayog: Responsible AI for All (India)

Organizational Practices

Create AI Ethics Boards
Maintain Model Cards and Datasheets for Datasets
Conduct Ethical Impact Assessments
Train employees on Responsible AI principles

Tools for Bias & Ethics Management

- AIF360 (IBM)
- Fairlearn (Microsoft)
- What-If Tool (Google)
- Explainable AI Dashboards
- Model Cards Toolkit

Best Practices

✔ Use diverse and representative datasets
✔ Include human oversight
✔ Make AI models explainable
✔ Continuously monitor models post-deployment

✔ Establish escalation channels for ethical risks

Case Study

COMPAS Recidivism Prediction System:
Displayed racial bias by predicting higher risk for black defendants.
Outcome: Sparked introduction of fairness-aware metrics and auditing practices.

Future Directions

AI-by-Design ethical frameworks
Federated and privacy-preserving learning
Transparent LLM pipelines with audit trails
Integration of ethics into AI education and governance

Summary

Ethical AI ensures fairness, transparency, and accountability throughout the lifecycle. Bias mitigation requires interventions in data, model, and governance levels.

Q&A; / Discussion

Prompt: How can we balance accuracy and fairness without compromising ethical principles?