Ethical AI: Addressing Bias, Privacy, and Security

Welcome to Week 3 of our AI Ethics course. We'll explore critical ethical challenges that shape modern AI development and deployment.

by S MM

Understanding Algorithmic Bias

Source Data Issues

Biases emerge from historical inequities reflected in training datasets.

Algorithm Design

Feature selection and model architecture choices can amplify existing biases.

Deployment Context

Real-world applications reveal unexpected biases through feedback loops.

Real-World Bias Manifestations

Healthcare

- Diagnosis algorithms less accurate for minorities
- Treatment recommendations favoring certain demographics

Criminal Justice

- Recidivism prediction disparities across racial groups
- Facial recognition false positives for non-white faces

Employment

- Resume screening systems penalizing women candidates
- Interview analysis algorithms favoring specific speech patterns

Financial Services

- Credit scoring algorithms discriminating against minorities
- Loan approval rates varying by neighborhood demographics

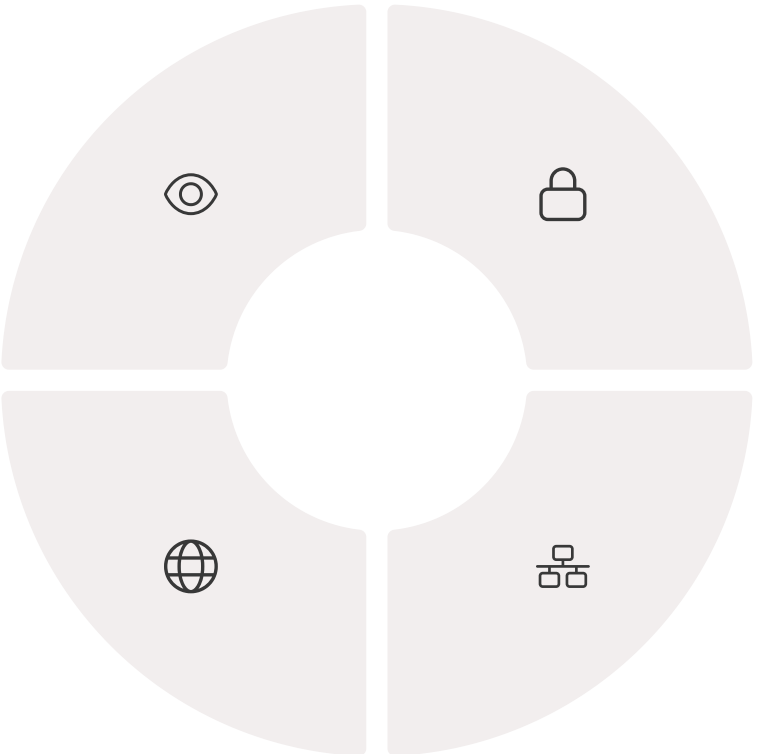
Privacy Concerns in AI Systems

Surveillance Capitalism

Companies monetize personal data through prediction products. Users become both resource and target.

Cross-Border Data Flows

Data transfers between jurisdictions create regulatory gaps. Protections vary widely by location.



Consent Issues

Complex terms of service obscure how data is collected. Meaningful consent becomes impossible.

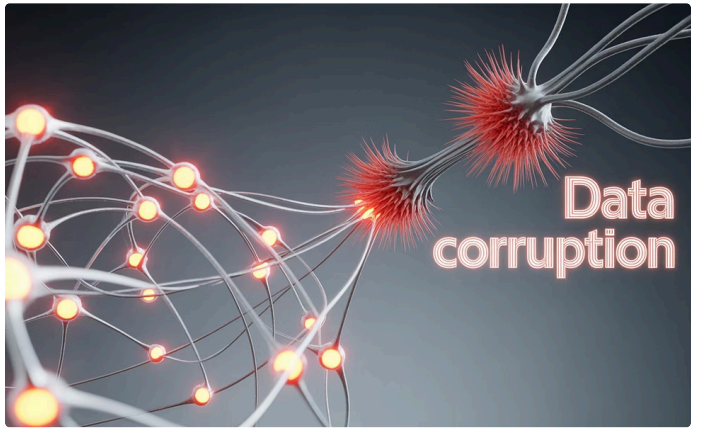
Data Aggregation

Seemingly innocuous data points combine to reveal sensitive information. De-anonymization becomes trivial.

Security Vulnerabilities in AI

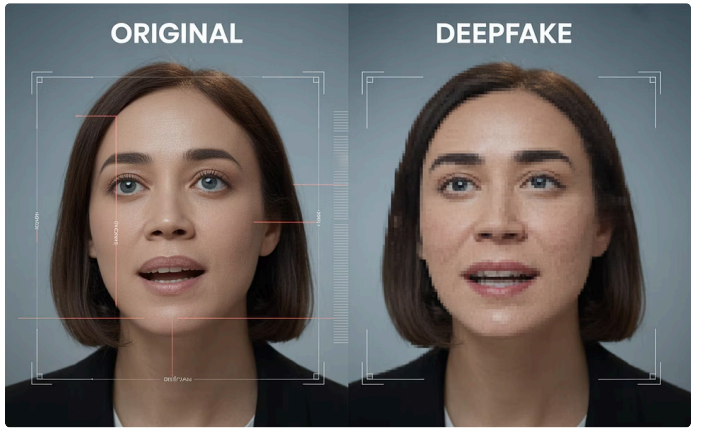
Adversarial Attacks

Subtle input manipulations cause classification errors. Imperceptible changes lead to catastrophic failures.



Model Poisoning

Manipulated training data compromises model integrity. Backdoors create hidden vulnerabilities.



Deepfakes

Synthetic media undermines trust in visual evidence. Detection becomes increasingly difficult.

Fairness-Aware Machine Learning

1 Pre-Processing Techniques

Modify training data to remove or balance bias before model training.

- Reweighting samples from underrepresented groups
- Transforming features to ensure fairness constraints

2 In-Processing Methods

Incorporate fairness constraints directly into learning algorithm.

- Adversarial debiasing during training
- Fairness regularization in objective functions

3 Post-Processing Approaches

Adjust model outputs to ensure fair predictions across groups.

- Threshold adjustment for different demographics
- Calibration techniques to equalize error rates

Differential Privacy & Secure ML

Differential Privacy

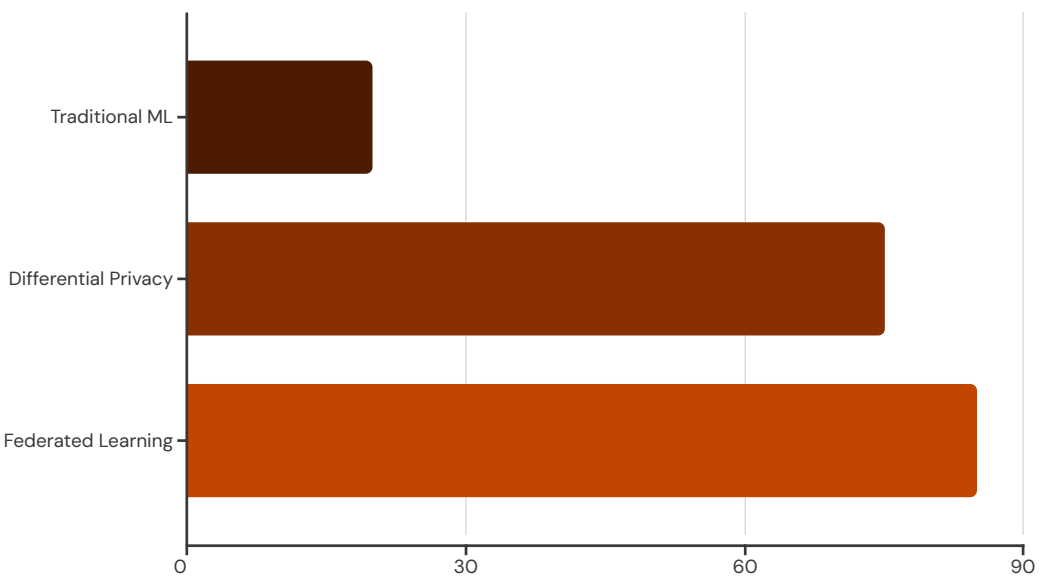
Mathematical framework for sharing information without exposing individuals.

- Adds calibrated noise to statistical queries
- Provides formal privacy guarantees
- Offers privacy-utility tradeoff control

Federated Learning

Trains models across devices while keeping data local.

- Only model updates leave user devices
- Raw data never centrally collected
- Can combine with encryption techniques



Ethical AI Development Framework

Audit & Assessment

Evaluate systems for bias, privacy risks, and security vulnerabilities through rigorous testing.

Mitigation Strategies

Apply appropriate fairness, privacy, and security techniques based on assessment findings.

Continuous Monitoring

Implement ongoing surveillance for emergent biases and vulnerabilities in deployed systems.

Stakeholder Inclusion

Involve diverse perspectives in development to identify blind spots and potential harms.

The ethical AI development cycle requires commitment to ongoing evaluation and improvement.