Ethics for AI: From Principles to Practice

Learning goals: recognize risks • apply mitigations • understand governance

Why AI Ethics Now

- AI decisions affect hiring, lending, health, justice, education
- Mistakes scale quickly; trust is fragile
- Ethics ⇒ risk management, compliance, quality

Learning Objectives

- Identify key risks: bias, privacy, safety, accountability
- Use guardrails across the ML lifecycle
- Communicate trade-offs to non-experts

Core Principles (Toolkit)

- Fairness (equity across groups)
- Accountability (clear ownership of outcomes)
- Transparency (what, why, limits)
- Privacy (minimization, consent, security)
- Safety & Robustness (fail-safes, red teaming)

Where Bias Comes From

- Historical data & sampling bias
- Label/annotation bias
- Feature leakage & proxies (e.g., ZIP ↔ income/race)
- Deployment drift & feedback loops

Measuring Fairness

- Group metrics: demographic parity, equalized odds
- Individual/causal views: counterfactual fairness
- No single metric fits all—choose for your harm model

Mitigation Levers

- Pre-process
  - Sampling, reweighting, augmentation
- In-process
  - Constraints, regularizers, debias objectives
- Post-process
  - Thresholding, calibration per group
- Document decisions (Model Cards)

Privacy & Data Governance

- Data minimization; purpose limitation
- Consent & revocation; sensitive attributes
- Re-identification risks (linkage attacks)
- Access controls, logging, retention policies

Privacy Techniques in Practice

- Differential privacy (noise budgets)
- Federated learning & secure aggregation
- Synthetic data (benefits & pitfalls)
- PII handling & de-identification limits

Transparency & Explainability

- When do you need explanations? rights, debugging, trust
- Local vs. global explanations
- Datasheets for Datasets & Model Cards
- Communicate uncertainty & limitations

Accountability & Governance

- Roles: product, data steward, model owner, evaluator
- Approvals, audit trails, incident response
- Human-in-the-loop: escalation paths, override authority

Safety & Robustness (incl. GenAI)

- Adversarial inputs, prompt injection, data poisoning
- Hallucinations & reliability testing
- Red teaming; evaluation suites; guardrails

Responsible AI Lifecycle

- Framing → Data → Training → Evaluation → Deployment → Monitoring
- Risk reviews at gates; model/version registry
- Sunset & rollback plans

Legal & Policy Touchpoints (High Level)

- Risk-based regulation themes (high-risk use cases)
- Sectoral rules (health/finance/education) vary by jurisdiction
- Org policies: data retention, access, incident reporting

Case Exercise: Hiring Screening Tool

- Observation: Lower pass rates for Group A
- Discuss in pairs/small groups—answer:
- Questions
  - Causes of disparity?
  - Which fairness metric fits? Why?
  - First mitigations to try?
  - What will you document and to whom?

Debrief: Plausible Answers

- Causes: historical bias, proxy features, imbalanced data
- Metric: equalized odds vs demographic parity (context-driven)
- Mitigations: reweighting + threshold tuning + feature audit
- Documentation: Model Card; evaluation report; known limits

Quick Pitfalls to Avoid

- Training-serving skew; silent degradation
- Over-reliance on accuracy; no slice metrics
- Shadow deployments without monitoring
- “Ethics by assertion” (no evidence)

Practical Checklist

- Harm analysis in problem framing
- Data lineage & consent verified
- Fairness metrics on key slices; stress tests
- Explainability & documentation shipped with model
- Governance sign-offs; rollback plan; monitors in place

Resources

- Practices: Model Cards; Datasheets for Datasets; risk checklists
- Playbooks: organization RAI guidelines; red-teaming guides
- Learning: courses/readings on fairness, privacy, safety

Exit Ticket (2–3 min)

- One risk in your current/next project
- One metric or control you will adopt