

Introduction to AI Ethics

Introduction to AI Ethics – Structured Study Notes

1. Why AI Ethics Matters

Reason	Explanation
Impact on Society	AI systems influence employment, health care, finance, law enforcement, and everyday decisions.
Scale & Speed	Algorithms can affect millions of people instantly, amplifying both benefits and harms.
Opacity	Complex models (e.g., deep neural networks) are often “black boxes,” making accountability difficult.
Power Asymmetry	Large tech firms and governments can wield AI for surveillance, manipulation, or market dominance.
Legal & Regulatory Pressure	Emerging laws (EU AI Act, US AI Bill of Rights, China’s AI Governance Guidelines) require ethical compliance.

2. Core Ethical Principles for AI

Principle	Key Questions	Typical Mitigations
Beneficence	Does the system promote well-being?	Human-centered design, impact assessments.
Non-maleficence	Could it cause harm?	Risk analysis, safety testing, fail-safes.
Autonomy	Does it respect user agency?	Transparent consent flows, opt-out options.
Justice & Fairness	Are outcomes equitable across groups?	Bias audits, diverse training data, fairness-aware algorithms.
Transparency & Explainability	Can stakeholders understand decisions?	Model interpretability tools, documentation (model cards, datasheets).
Accountability	Who is responsible for outcomes?	Governance frameworks, audit trails, clear liability clauses.

Introduction to AI Ethics

Principle	Key Questions	Typical Mitigations
Privacy	Are personal data protected?	Data minimization, differential privacy, secure storage.
Sustainability	What is the environmental footprint?	Energy-efficient models, carbon accounting.

3. Common Ethical Risks in AI

Risk	Description	Example
Bias & Discrimination	Systematic prejudice encoded in data or model.	Facial recognition misidentifying darker-skinned faces.
Opacity (Black Box)	Inability to interpret model reasoning.	Credit scoring algorithm denying loans without explanation.
Surveillance & Privacy Erosion	Excessive data collection and tracking.	Smart city cameras linked to facial recognition databases.
Automation-Induced Job Displacement	Large-scale workforce disruption.	AI chatbots replacing call center agents.
Weaponization	Use of AI for lethal or coercive purposes.	Autonomous drones targeting civilians.
Manipulation & Misinformation	AI-generated content influencing opinions.	Deepfake videos spreading false political statements.
Security Vulnerabilities	Models susceptible to adversarial attacks.	Slight pixel changes causing misclassification in self-driving cars.

4. Ethical AI Development Lifecycle

1. Problem Definition

- Clarify objectives, stakeholders, and societal impact.
- Conduct a **pre-impact assessment** (benefits vs. potential harms).

2. Data Collection & Management

Introduction to AI Ethics

- Verify data provenance, consent, and representativeness.
- Apply **deidentification** and **bias detection** tools.

3. Model Design & Training

- Choose architectures with interpretability in mind (e.g., rule-based hybrids).
- Incorporate **fairness constraints** (e.g., demographic parity, equalized odds).
- Use **privacy-preserving techniques** (differential privacy, federated learning).

4. Evaluation & Testing

- Test on