

Introduction to Ethics in Artificial Intelligence

Introduction to Ethics in Artificial Intelligence
Study Guide – Key Concepts, Frameworks, and Practical Considerations

1. Why Ethics Matters in AI

Reason	Explanation
Impact Scope	AI systems influence decisions in healthcare, finance, law enforcement, hiring, and more.
Autonomy & Agency	Algorithms can act without direct human oversight, raising questions about responsibility.
Bias & Fairness	Data-driven models can replicate or amplify societal prejudices.
Transparency	Complex models (e.g., deep neural networks) are often “black boxes,” hindering accountability.
Trust & Adoption	Public confidence depends on perceived moral soundness of AI solutions.

2. Core Ethical Principles for AI

Principle	Definition	Typical AI-related Questions
Beneficence	Promote well-being and avoid harm.	Does the system improve health outcomes? Could it cause physical or psychological injury?
Non-maleficence	Do no harm.	Are there failure modes that could lead to unsafe actions?
Justice / Fairness	Distribute benefits and burdens equitably.	Does the model discriminate on gender, race, age, or disability?
Autonomy	Respect individuals' right to make informed choices.	Are users given meaningful control over AI-driven decisions?
Transparency / Explainability	Make processes understandable to stakeholders.	Can the system's reasoning be inspected or contested?
Accountability	Assign clear responsibility for outcomes.	Who is liable when the AI makes a mistake?

Introduction to Ethics in Artificial Intelligence

Principle	Definition	Typical AI-related Questions
Privacy	Protect personal data and limit unnecessary collection.	Is data minimised and stored securely?

3. Ethical Frameworks & Approaches

Framework	Main Idea	Use Cases
Principlism (Beauchamp & Childress)	Apply the four biomedical ethics principles (beneficence, non-maleficence, autonomy, justice) to AI.	Medical AI diagnostics, clinical decision support.
Value-Sensitive Design (VSD)	Embed human values (e.g., dignity, sustainability) throughout the design lifecycle.	Social-media recommendation engines, smart city sensors.
Fairness-Aware Machine Learning	Formalize fairness (e.g., demographic parity, equalized odds) and adjust models accordingly.	Hiring algorithms, credit scoring.
Human-Centred AI (HCAI)	Prioritise human well-being, empowerment, and oversight.	Collaborative robots (cobots), assistive technologies.
Responsible AI Governance	Combine policy, standards, and organisational processes to manage risk.	Enterprise AI deployments, government AI procurement.

4. Common Ethical Risks & Mitigation Strategies

Risk	Description	Mitigation Tactics
Bias & Discrimination	Systemic errors that favour or disadvantage groups.	• Conduct bias audits on training data. • Use fairness metrics during model validation. • Apply re-weighting or adversarial debiasing techniques.

Introduction to Ethics in Artificial Intelligence

Risk	Description	Mitigation Tactics
Opacity / Lack of Explainability	Stakeholders cannot understand decisions.	• Adopt interpretable models where feasible. • Use post-hoc explanation tools (LIME, SHAP). • Provide user-friendly decision summaries.
Safety Failures	Unexpected behavior causing physical or financial harm.	• Implement rigorous testing (simulation, adversarial testing). • Deploy fail