AI Ethics & Fairness

Student Self-Training Guide · Grade 6+ · Module 6 of 10

1. What is AI Ethics?

Ethics means deciding what is right and wrong. AI Ethics asks: "Is the AI being fair? Is it honest? Is it safe? Could it harm someone?" As AI becomes more powerful, these questions become more important.

2. The 5 Principles of Ethical AI

Principle	Meaning	Example
Fairness	Treat everyone equally	Hiring AI should not prefer one gender
Transparency	Explain how decisions are made	AI should say why it denied a loan
Privacy	Protect personal data	AI should not share your health data
Safety	Do no harm	Self-driving AI must prioritise human life
Accountability	Humans stay responsible	A doctor reviews all AI diagnoses

3. Real-World Ethics Problems

- Facial recognition AI identified innocent people as criminals due to biased training data.
- AI hiring tools rejected women because past data had mostly male employees.
- Social media AI pushed harmful content to teenagers to keep them engaged.
- Deepfake AI created fake videos of real people saying things they never said.

4. Bias in AI

KEY IDEA

AI learns bias from humans. If the training data reflects unfair human decisions, the AI will make unfair decisions too. Garbage in = Garbage out.

5. Your Role as a Future AI Builder

- Always ask: "Who could be harmed by this AI?"
- Test your AI on diverse groups of people.
- Be honest about what your AI can and cannot do.
- Design AI with a human expert always in the loop for critical decisions.

6. Debate Topic

CLASS DEBATE

"AI should be allowed to decide who gets bail in criminal cases." — Argue FOR or AGAINST in 3 minutes. Use examples and the 5 ethical principles.