MODULE 2

The Ethics Problem

Frameworks for AI Ethics

STUDY GUIDE

AI Ethics for Higher Education

EduPolicy.ai / ScholarBar Education LLC

1 Three Core Tensions

Every AI ethics problem stems from at least one: **power without accountability** (decisions affecting millions with no individual oversight), **speed without deliberation** (millisecond decisions that ethics requires reflection for), and **scale without individual consideration** (applying the same rule to everyone when fairness often requires differentiation).

2 Utilitarianism

Judges actions by outcomes — the greatest good for the greatest number. An AI that reduces crime by 30% through facial recognition is justified under this framework, even if privacy is compromised. **Weakness:** can justify harming minorities if the majority benefits.

3 Deontology

Rules-based ethics — some actions are wrong regardless of outcome. Mass surveillance violates individual rights even if it prevents crime. **Weakness:** rules can conflict with each other.

4 Virtue Ethics

Character-based — asks "what kind of society are we becoming?" rather than calculating outcomes or applying rules. "A just society doesn't treat its citizens as suspects." **Weakness:** "good person" is subjective and culturally dependent.

5 The AI Trolley Problem

Autonomous vehicle decisions differ from human ones because they are **pre-programmed**. Someone wrote the code that weighted the options months before the collision. This creates moral responsibility for decisions made faster than human thought — a fundamentally new ethical category.

6 Proxy Discrimination

Removing protected attributes (gender, race) from training data doesn't remove bias. Correlated features — zip codes, "women's chess club" on a resume, employment gaps from parental leave — carry the same discriminatory signal. The bias hides in proxies.

7 The Accountability Gap

When AI causes harm, responsibility distributes across developers, deployers, users, and regulators. Each party can point to the others. Nobody accepts full responsibility. This gap is structural, not accidental.

8 Case Study: COMPAS

AI system used in U.S. courts for sentencing. Predicted recidivism but flagged Black defendants as "high risk" at nearly twice the rate of white defendants. Company refused to reveal its algorithm. Courts used it despite known flaws. Demonstrates bias, opacity, and accountability failure in one case.