MODULE 3

Bias & Fairness

Algorithmic Discrimination

STUDY GUIDE

AI Ethics for Higher Education

EduPolicy.ai / ScholarBar Education LLC

1 Structural Bias

AI bias isn't individual prejudice — it's historical inequity encoded in data, amplified by algorithms, and deployed at scale. A biased human loan officer denies 10 loans unfairly per year. A biased AI lending model denies thousands — and nobody at the bank knows it's happening because the system looks "objective."

2 Six Pipeline Entry Points

Bias enters at every stage: (1) **Data Collection** — who was included/excluded, (2) **Data Labeling** — whose definitions of "good" are encoded, (3) **Feature Selection** — which variables are proxies for protected characteristics, (4) **Model Training** — optimizing for accuracy using biased data produces accurately biased results, (5) **Evaluation** — aggregate accuracy masks group-level disparities, (6) **Deployment** — biased outputs become new training data through feedback loops.

3 Gender Shades Study (2018)

MIT researcher Joy Buolamwini tested commercial facial recognition from Microsoft, IBM, and Face++. Error rates: **0.8%** for light-skinned men, **7.0%** for light-skinned women, **12.0%** for dark-skinned men, **34.7%** for dark-skinned women. A 43x performance gap hidden by high overall accuracy.

4 Fairness Impossibility Theorem

Researchers proved mathematically (2016) that when base rates differ between groups, you cannot simultaneously achieve demographic parity, equal opportunity, AND calibration. Choosing which definition of fairness matters most is a **human values decision**, not a technical one.

5 Proxy Discrimination

Removing protected attributes from training data doesn't remove bias. Correlated features carry the same discriminatory signal: zip codes proxy for race, employment gaps proxy for gender, club memberships signal demographics. The bias hides in proxies.

6 Feedback Loops

Biased predictions → biased actions → biased new data → more biased predictions. Predictive policing: AI sends officers to historically over-policed neighborhoods → more arrests → data "confirms" prediction → more officers sent. The AI doesn't predict where crime happens — it predicts where police will find crime because they're already looking there.

7 Algorithmic Auditing

Systematic testing of AI systems for disparate impact. The Gender Shades study triggered IBM to improve (65% disparity reduction), then exit facial recognition entirely. Measurement drives accountability.