🌍 AI Ethics Assignment — Designing Responsible and Fair AI Systems

Part 1: Theoretical Understanding

Q1: Algorithmic Bias

Algorithmic bias occurs when AI systems produce systematically unfair outcomes for certain groups.
Examples:

1. A hiring algorithm penalizing women due to biased historical hiring data.
2. A credit scoring model assigning lower scores to minority groups due to biased training sets.

Q2: Transparency vs Explainability

- **Transparency** means open access to system design, data sources, and model structure.
- **Explainability** means being able to understand *why* a model made a specific decision.
  Both are crucial for building trust and accountability in AI systems.

Q3: GDPR and AI

GDPR enforces data privacy, fairness, and accountability. It impacts AI by requiring:

- Explicit consent for data use.
- The right to explanation for automated decisions.
- Strict data minimization and protection standards.

Ethical Principles Matching

Principle	Definition
Justice	Fair distribution of AI benefits and risks
Non-maleficence	Ensuring AI does not harm individuals or society
Autonomy	Respecting users’ right to control their data and decisions
Sustainability	Designing AI to be environmentally friendly

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

- **Source of Bias:** Historical data reflecting male dominance in tech hiring.
- **Fixes:**
  1. Balance training data across genders.
  2. Use bias detection metrics during model training.
  3. Include fairness constraints in optimization.
- **Metrics for Fairness:**

    - Disparate Impact Ratio
    - Equal Opportunity Difference
    - Statistical Parity Difference

Case 2: Facial Recognition in Policing

- **Ethical Risks:**
    - Higher false positives for minorities.
    - Privacy violations through mass surveillance.
- **Policy Recommendations:**
    - Mandate independent audits before deployment.
    - Enforce human oversight in decisions.
    - Require transparency and bias documentation.

Part 3: Practical Audit (COMPAS Dataset)

Bias Audit Results

Model Performance:

```
Precision (0.0): 0.68 | Recall: 0.74
Precision (1.0): 0.65 | Recall: 0.58
Accuracy: 0.67
```

Fairness Metrics:

```
Before Mitigation:
 - Mean difference: -0.097
 - Disparate impact: 0.840

After Mitigation (Reweighing):
 - Mean difference: 0.000
 - Disparate impact: 1.000
```

Summary (300 words)

The COMPAS dataset displayed measurable racial bias before mitigation. African-American defendants were predicted as “high-risk” more often than Caucasian defendants, even with similar backgrounds. Using **AI Fairness 360**, bias was quantified via disparate impact and mean difference. After applying **Reweighing**, the disparate impact improved to 1.0, indicating parity in treatment.
Mitigation successfully balanced risk predictions while maintaining acceptable accuracy (0.67). However, fairness improvement came with a minor tradeoff in precision. This underscores that fairness interventions may reduce performance but increase social equity and accountability. The final audit recommends ongoing monitoring, dataset diversification, and human oversight to sustain fairness in high-stakes applications like justice and healthcare.

Part 4: Ethical Reflection

In future projects, I will ensure ethical AI by:

- Conducting fairness audits at every stage.
- Documenting data provenance.
- Maintaining transparency through explainable models.
- Involving interdisciplinary review teams to minimize bias.

Bonus Task: Ethical AI in Healthcare

Policy Proposal Summary:

- **Patient Consent:** Obtain explicit consent for data collection and AI-based decisions.
- **Bias Mitigation:** Train on diverse populations; apply reweighing or adversarial debiasing.
- **Transparency:** Provide interpretable model explanations to clinicians and patients.
- **Accountability:** Regularly audit model outputs and retrain using up-to-date data.

Prepared by: Bikila Keneni
PLP Academy — AI for Software Engineering(2025)