AI Ethics Assignment: Designing Responsible and Fair AI Systems

Part 1: Theoretical Understanding (30%)

Q1: Algorithmic bias is the systematic and unfair discrimination produced by an AI system.
Examples:
1. Hiring algorithms favoring male candidates due to biased historical data.
2. Facial recognition systems misidentifying darker■skinned individuals at higher rates.

Q2: Transparency means users can see how the AI works (data sources, rules, processes), while explainability means the AI can justify individual predictions. Both are critical for trust, accountability, and debugging bias.

Q3: GDPR impacts AI by enforcing data protection, requiring consent, enabling the right to explanation, and limiting automated decision-making without human oversight.

Ethical Principles Matching:
A) Justice → Fair distribution of AI benefits and risks.
B) Non■maleficence → Ensuring AI does not harm individuals or society.
C) Autonomy → Respecting users' right to control their data and decisions.
D) Sustainability → Designing AI to be environmentally friendly.

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool
Source of Bias: Historical male■dominated hiring data; model learned gender proxies; skewed feature weighting.
Fixes:
1. Remove gender■related features and proxies.
2. Rebalance or augment training data.
3. Use fairness■constrained models.
Fairness Metrics:
- Disparate Impact Ratio
- Equal Opportunity Difference
- Statistical Parity

Case 2: Facial Recognition in Policing
Ethical Risks: Wrongful arrests, privacy invasion, racial discrimination, surveillance overreach.
Policies:
- Mandatory human review.
- Independent bias audits.
- Strict usage limitations.
- Transparency and public reporting.

Part 3: Practical Audit Summary (300 words) (25%)

Using the COMPAS recidivism dataset, a fairness audit reveals significant racial bias in false positive rates.

African■American defendants are far more likely to be misclassified as "high■risk" despite not reoffending.
AIF360 metrics such as disparate impact, equal opportunity difference, and average odds difference confirm these disparities.
Visualizations show skewed score distributions, higher FPR for Black defendants, and inconsistent calibration across groups.

Remediation:
- Reweighing to equalize dataset balance.
- Adversarial debiasing during model training.
- Post■processing with equalized odds adjustments.

Part 4: Ethical Reflection (5%)

In future AI projects, I will prioritize fairness by conducting bias audits, ensuring transparency, applying privacy■first data practices, and validating models with diverse stakeholders.