AI Ethics Coursework Answers

Part 1: Theoretical Understanding (30%)

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic and repeatable errors in an AI system that create unfair outcomes, such as privileging one group over others. It often stems from biased training data or flawed design.

Examples:
1. Hiring Tools: An AI trained on resumes of past successful candidates may favor male candidates if the historical data reflects gender bias.
2. Credit Scoring: AI models may give lower credit scores to individuals from minority communities if historical financial data reflects systemic inequality.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?

Transparency refers to openness about how an AI system is designed, trained, and operated, including access to its data sources and algorithms.
Explainability refers to the ability to interpret and understand how the AI system arrives at a particular decision or prediction.

Importance:
Transparency builds trust by allowing scrutiny of system design, while explainability ensures accountability, enabling users and developers to understand, challenge, or correct decisions.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

GDPR enforces strict rules on personal data usage, directly impacting AI development by:
- Requiring explicit user consent for data processing.
- Granting individuals the right to explanation for automated decisions.
- Limiting automated profiling that significantly affects individuals.

This encourages AI developers to prioritize privacy, data minimization, and human oversight in their systems.

Ethical Principles Matching

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.
A) Justice - Fair distribution of AI benefits and risks.

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool

Scenario: Amazon's AI recruiting tool penalized female candidates.

Source of Bias:
- Training Data: The model was trained on resumes submitted over a 10-year period, which reflected male-dominated

hiring practices.
- Model Design: The system learned to favor terms more frequently associated with male candidates and penalized resumes mentioning "women's."

Three Fixes:
1. Audit and diversify training data to include balanced representations of gender, education, and career paths.
2. Introduce fairness constraints during model training to penalize gender-based discrimination.
3. Use debiasing algorithms or pre-process data to remove gendered language and proxies.

Fairness Metrics:
1. Demographic parity - Equal selection rates across genders.
2. Equal opportunity - Similar true positive rates for different groups.
3. Disparate impact ratio - Measures the ratio of positive outcomes between groups.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

Ethical Risks:
1. Wrongful Arrests: Misidentification can lead to innocent people being detained.
2. Privacy Violations: Constant surveillance undermines individual privacy rights.
3. Discrimination: Systemic targeting of marginalized communities increases bias in policing.

Responsible Deployment Policies:
1. Bias Testing & Audits: Mandatory pre-deployment audits for racial and gender bias.
2. Human Oversight: Ensure all matches are verified by trained personnel before action is taken.
3. Use Restrictions: Ban use in high-risk contexts (e.g., crowd surveillance) until technology meets strict accuracy standards.
4. Public Transparency: Disclose system use, accuracy rates, and complaint procedures to the public.