AI Ethics

Part 1: Theoretical Understanding (30%)

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

Definition:
Algorithmic bias occurs when an AI system reflects the prejudices present in its training data or design, leading to systematically unfair outcomes for certain groups (e.g., based on gender, race, or socioeconomic status).

Examples:

1. Gender bias in hiring tools: Amazon's AI recruiting tool downgraded resumes containing words like "women's" or all-female colleges, reflecting historical underrepresentation of women in tech roles.
2. Racial bias in facial recognition: Some facial recognition systems have higher error rates for darker-skinned individuals due to underrepresentation in training data, leading to misidentification in law enforcement.

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

Transparency refers to how clear and open the system is about its processes, data sources, and decision-making logic. It includes information like who developed the system, what data was used, and how the model was validated.

Explainability refers to the ability to explain, in understandable terms, why a model made a particular decision. It focuses on making the reasoning behind individual predictions interpretable.

Why Both Are Important:

- Transparency builds trust and enables accountability, especially in regulated domains like healthcare or finance.
- Explainability ensures users can understand and challenge decisions, which is crucial for fairness and ethical compliance.

Q3: How does GDPR impact AI development in the EU?

The General Data Protection Regulation (GDPR) impacts AI development in the EU by:

- Requiring explicit consent for collecting and processing personal data.

- Granting individuals, the right to explanation for automated decisions affecting them (e.g., loan approvals, hiring).
- Enforcing data minimization and purpose limitation, which affects how much data AI systems can collect and use.
- Mandating data protection impact assessments (DPIAs) for high-risk AI systems.
- Imposing heavy fines for non-compliance, incentivizing ethical and privacy-conscious AI development.

---

3. Ethical Principles Matching

Principle	Definition
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 (Amazon AI Recruiting Tool)

Identify the source of bias:

- Training data bias: The model was trained on historical hiring data that reflected existing gender imbalances in tech.
- Model design: The system learned to associate male-dominated language with better candidates.

Three fixes to make the tool fairer:

1. Use balanced and representative training data including diverse candidate profiles.
2. Apply fairness-aware machine learning techniques (e.g., adversarial debiasing, reweighting).
3. Audit and monitor model outputs regularly for demographic disparities.

Metrics to evaluate fairness post-correction:

- Statistical Parity Difference (SPD)
- Disparate Impact Ratio

- Equal Opportunity Difference
- False Positive Rate Disparity

Case 2: Facial Recognition in Policing

Ethical risks:

- Wrongful arrests due to misidentification of minorities.
- Privacy violations from mass surveillance and lack of consent.
- Discriminatory policing reinforcing systemic racism.
- Lack of transparency and accountability in how decisions are made.

Policies for responsible deployment:

1. Ban or restrict use in high-stakes contexts without rigorous fairness testing.
2. Implement strict oversight and audit requirements for law enforcement use.
3. Ensure informed consent where possible, and provide avenues for redress.
4. Mandate racial and demographic impact assessments before deployment.