**Part 1: Theoretical Understanding (30%)**

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

Answer:

Algorithmic bias refers to systematic and repeatable errors in an AI system that create unfair outcomes, often discriminating against certain groups. It arises due to biased data, flawed assumptions, or opaque model design.

Examples:

1. Hiring Algorithms: Amazon’s recruiting tool favored male resumes due to biased historical data reflecting a male-dominated tech industry.

2. Credit Scoring Systems: AI models have been found to offer less favorable loan terms to minority groups, even when income and creditworthiness were similar.

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

Answer:

Transparency refers to the openness about how an AI system operates—its data sources, algorithms, and decision-making process.

Explainability is the ability to interpret and understand why a system made a specific decision.

Importance:

Transparency builds trust by revealing how the system functions. Explainability ensures stakeholders can audit, debug, or contest outcomes—vital for fairness, especially in high-stakes domains like healthcare or criminal justice.

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

Answer:

GDPR enforces strict data privacy and protection laws. For AI developers, it mandates:

Data minimization and informed consent.

The right to explanation for automated decisions.

Restrictions on profiling and sensitive data use.

This compels developers to adopt privacy-by-design, accountable algorithms, and human oversight.

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.

D) Sustainability Designing AI to be environmentally friendly.

**Part 2: Case Study Analysis (40%)**

Case 1: Biased Hiring Tool (Amazon)

Source of Bias:

Training data: The model was trained on resumes from the past decade, which were predominantly from male candidates.

Model design: It penalized resumes with words like “women’s” (e.g., “women’s chess club”).

Proposed Fixes:

1. Data Rebalancing: Curate a diverse and inclusive training set, ensuring equal representation across genders.

2. Feature Engineering: Remove or mask gender-related attributes and proxies (e.g., clubs, names).

3. Bias-aware Models: Use algorithms that support fairness constraints (e.g., adversarial debiasing).

Fairness Metrics to Evaluate:

Disparate Impact Ratio

Equal Opportunity Difference

Demographic Participation

Case 2: Facial Recognition in Policing

Ethical Risks:

Wrongful Arrests: Misidentifications have led to unjust incarcerations.

Privacy Violations: Mass surveillance infringes on civil liberties.

Disproportionate Harm: Minorities face higher error rates, reinforcing systemic bias.

Recommended Policies:

1. Human-in-the-loop validation before any action is taken based on recognition.

2. Mandatory bias audits before deployment.

3. Ban real-time facial recognition in public spaces without court oversight