Part 1: Theoretical Understanding

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

Algorithmic bias occurs when an AI system produces systematically prejudiced results due to erroneous assumptions in the machine learning process.

Examples:

1. **Hiring Tools** – Amazon's recruitment AI preferred male resumes because it was trained on historical male-dominated data.

2. **Facial Recognition** – Systems like Clearview AI misidentify people of color at higher rates, leading to false arrests.

Q2: Difference Between Transparency and Explainability in AI

- **Transparency**: Refers to how open and accessible the internal workings of the AI system are (e.g., data sources, model structure).

- **Explainability**: Refers to how easily a human can understand the reasons behind a decision or prediction made by the AI.

Importance:

- Transparency builds trust and accountability.

- Explainability ensures users and stakeholders can understand and challenge AI decisions.

Q3: GDPR Impact on AI Development in the EU

- **GDPR** mandates data minimization, explicit user consent, and the right to explanation (Article 22).

- AI developers must ensure:
    - **Data anonymization**
    - **Explainability of decisions**
    - **Right to opt-out of automated processing**

Ethical Principles Matching

Definition	Ethical Principle
Ensuring AI does not harm individuals or society	**B) Non-maleficence**
Respecting users' right to control their data and decisions	**C) Autonomy**
Designing AI to be environmentally friendly	**D) Sustainability**
Fair distribution of AI benefits and risks	**A) Justice**

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool (Amazon)

- **Source of Bias**:
    - Training data was biased toward resumes from male candidates (reflecting past hiring practices).
    - Model rewarded male-associated language/phrases.
- **Three Fixes**:
    - **Balanced Training Data**: Ensure gender-neutral or balanced data is used.
    - **Bias Detection Pipelines**: Incorporate bias-checking frameworks like AIF360 during model evaluation.

    - **Remove Sensitive Features**: Avoid using gendered features or proxies like school names or hobbies.

- **Fairness Metrics**:
    - Disparate Impact Ratio
    - Equal Opportunity Difference
    - Demographic Parity

Case 2: Facial Recognition in Policing

- **Ethical Risks**:
    1. **Wrongful arrests** due to misidentification.
    2. **Privacy violations** from unconsented surveillance.
    3. **Racial profiling** amplifying existing social inequalities.
- **Recommended Policies**:
    1. **Mandatory third-party audits** of facial recognition tools.
    2. **Ban usage** in high-stakes decisions until proven fair.
    3. **Transparency** about data sources and algorithms.
    4. **Opt-in policies** for public usage.