Part 1: Theoretical Understanding

1. Short Answer Questions

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

Definition:
Algorithmic bias refers to systematic and repeatable errors in an AI system that create unfair outcomes, often privileging one group over others. This bias can originate from the training data, the model design, or deployment context and can lead to **discrimination**, **exclusion**, or **harm**.

Example 1: Criminal Justice (COMPAS System)
The COMPAS algorithm, used to predict criminal recidivism in the U.S., was found to assign **higher risk scores to Black defendants** than white defendants with similar profiles. This bias arose from historical arrest data and had significant implications for sentencing.

Example 2: Hiring Algorithms (Amazon Resume Screener)
Amazon's AI recruiting tool downgraded resumes that included the word "women's" (e.g., "women's chess club"), favoring male applicants. The system learned from past hiring data, which was heavily male-dominated, thus reinforcing gender bias.

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

Transparency refers to the openness of the AI system's **design, data sources, and decision-making process**. It means stakeholders can access information about how the system works — such as the algorithm used, data inputs, training process, and goals.

Explainability, on the other hand, is the ability to **understand and interpret the outputs of an AI model**, particularly for non-technical users. It focuses on making the **reasoning behind specific decisions** clear and understandable.

Why Both Are Important:

- **Transparency** builds **trust** by allowing external audits and regulatory compliance.

- **Explainability** ensures **accountability**, as stakeholders can challenge or appeal decisions (e.g., in healthcare or finance).
  Together, they uphold **ethical standards** by preventing black-box AI from making unchallengeable or harmful decisions.

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

The **General Data Protection Regulation (GDPR)** is a data privacy law that directly affects how AI systems are developed and used in the EU.

Key Impacts:

1. **Right to Explanation (Article 22):**
   Individuals have the right **not to be subject to decisions made solely by automated processing**, especially those with significant effects (e.g., loan approval, job selection). This means developers must design AI systems that are **explainable**.

2. **Data Minimization and Purpose Limitation:**
   AI systems must only use data that is **necessary, relevant, and collected for a specific, lawful purpose**, which restricts indiscriminate data harvesting.

3. **Consent and Transparency:**
   Users must give **informed consent** for their data to be used. This encourages the development of **transparent AI** with clear data usage policies.

4. **Accountability and Audits:**
   Organizations must document AI decisions and demonstrate **compliance** with privacy standards. This promotes robust **governance** and **bias monitoring**.

In summary, GDPR ensures AI is **human-centric**, **lawful**, and **fair**, though it also imposes strict constraints that may limit overly complex or opaque systems in the EU.

2. Ethical Principles Matching

Ethical 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.

Correct Matches:

- **A** → Fair distribution of AI benefits and risks
- **B** → Ensuring AI does not harm individuals or society
- **C** → Respecting users' right to control their data and decisions
- **D** → Designing AI to be environmentally friendly