Part 1: Theoretical Understanding

Q1

Algorithmic bias Define algorithmic bias as systematic and unfair discrimination embedded in AI systems, often arising from biased training data or model design. Examples: Facial recognition misidentifying minorities at higher rates. Recruitment tools penalizing female candidates unfairly.

Q2

Transparency vs Explainability Transparency is the openness about how an AI system operates (e.g., data, algorithms used). Explainability is the ability to make the AI s decision reasons understandable to humans. Both are important to build trust and enable accountability.

Q3

GDPR impact in EU. The GDPR ensures data protection and privacy, requiring AI systems to handle personal data responsibly, uphold data subject rights, and incorporate privacy-by-design principles.

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

Case 1: Biased Hiring Tool

Source of bias: Training data reflecting historical gender bias or model design ignoring fairness.

Fixes

Use balanced training data with gender diversity.

Implement fairness constraints in model training.

Incorporate bias mitigation algorithms like reweighing or adversarial debiasing.

Fairness metrics post-correction: Disparate impact ratio, equal opportunity difference, demographic parity.

Case 2: Facial Recognition in Policing

Ethical risks: Wrongful arrests, racial profiling, violation of privacy rights, reduced public trust.

Policies recommendation:

Restrict use to critical cases with human oversight.

Regular bias audits and transparency reports.

Consent and rights protections for individuals.

Establish strict accountability mechanisms.