Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

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

Algorithmic bias refers to systematic and unfair discrimination created or amplified by an AI system due to biased data, flawed design choices, or unintended model behaviors.

Examples:

1. **Biased facial recognition systems** misidentify people with darker skin tones more often because the training data is skewed toward lighter-skinned individuals.

2. **Job recruitment algorithms** that downgrade CVs from women because historical hiring data favored male candidates

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

- **Transparency** is about openly revealing how an AI system is built—its data sources, algorithms, architecture, and decision-making processes.

- **Explainability** is about making an AI system's outputs understandable to humans—explaining *why* the model made a particular prediction or decision.

Why both are important:

- Transparency builds **trust**, ensures **accountability**, and makes auditing possible.

- Explainability helps users and regulators **understand**, **question**, and **challenge** AI decisions, especially in high-stakes areas like healthcare or finance.

Together, they promote ethical, safe, and responsible AI deployment.

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

GDPR affects AI development in several ways:

- **Data protection requirements:** AI developers must follow strict rules on collecting, storing, and processing personal data.

- **Right to explanation:** Users can request an explanation of automated decisions that affect them, pushing developers to create explainable AI systems.

- **Consent and lawful basis:** AI systems must have a lawful basis for using personal data, and individuals must give informed consent.

- **Data minimization:** Developers must use only the data necessary for the AI's purpose.

- **Accountability:** Organizations must demonstrate compliance and conduct Data Protection Impact Assessments (DPIAs) for high-risk AI systems.

2. Ethical Principles Matching

Match the following principles to their definitions:

- **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

Scenario: Amazon's AI recruiting tool penalized female candidates.

1. Identify the source of bias

- Training data: Historical hiring data reflected gender imbalance (more male hires), causing the AI to favor male candidates.

- Model design: The algorithm prioritized patterns correlated with gender indirectly (e.g., word choices in CVs).

- Feature selection bias: Certain features reinforced existing stereotypes.

2. Propose three fixes to make the tool fairer

1. Use balanced datasets: Include equal representation of genders and diverse backgrounds.

2. Remove biased features: Exclude gender-related indicators (e.g., names, pronouns, schools with gender bias).

3. Bias mitigation algorithms: Apply techniques like reweighting or adversarial debiasing to reduce gender bias in predictions.

3. Metrics to evaluate fairness post-correction

- Demographic parity: Check if the hiring rate is similar across genders.

- Equal opportunity: Ensure qualified candidates of all genders have equal chances of being recommended.

- False positive/negative rates by group: Compare misclassification rates between male and female applicants.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

1. Ethical risks

- Wrongful arrests or convictions: Misidentification can lead to innocent people being accused.

- Privacy violations: Constant surveillance of marginalized groups undermines privacy rights.

- Reinforcement of systemic bias: Disproportionate targeting of minority communities can exacerbate social inequality.

- Erosion of public trust: Communities may lose trust in law enforcement and AI systems**.**

2. Policies for responsible deployment

1. Bias auditing: Regularly test models for accuracy across demographic groups.

2. Human oversight: Ensure final decisions involve human review, especially in law enforcement contexts.

3. Limited use cases: Restrict deployment to serious crimes or critical scenarios, not routine policing.

4. Transparency & accountability: Maintain logs, explain AI decisions, and allow independent audits.

5. Data minimization: Avoid collecting unnecessary personal data, and ensure secure storage.

Part 4: Ethical Reflection

Project Example: Voice-Based Medication Reminder System for Elderly Patients

Ethical Reflection:
In developing my medication reminder system, I will prioritize ethical AI principles to ensure it benefits users safely and fairly.

1. Privacy and Data Protection: Patient information, including medication schedules and health data, will be encrypted and stored securely. Access will be restricted to authorized caregivers only, complying with data protection standards.

2. Transparency: Users and caregivers will be clearly informed about what data is collected, how it is used, and how reminders are generated. This ensures trust and informed consent.

3. Accessibility and Inclusivity: The system will cater to elderly and visually impaired users, using clear voice prompts and intuitive interfaces, ensuring no group is disadvantaged.

4. Fairness: All patients will receive reminders equitably, avoiding biases in scheduling or prioritization that could favor certain users over others.

5. Accountability and Safety: Any system errors, such as missed reminders, will be logged and addressed promptly. Human oversight ensures critical decisions are never fully automated.

Conclusion:
By embedding privacy, transparency, fairness, accessibility, and accountability into the system's design, this project will use AI responsibly, support vulnerable populations, and maintain trust while improving medication adherence and patient health outcomes.