Part 1 — Theoretical Understanding

Q1: Define algorithmic bias and provide two examples.
Algorithmic bias is systematic unfairness in an AI system that causes certain groups to be disadvantaged due to skewed data, flawed model design, or biased assumptions.

Examples:
A hiring algorithm trained mostly on male applicants ranks women lower for technical jobs.
A facial recognition model misidentifies darker-skinned individuals at higher rates because it was trained predominantly on lighter-skinned faces.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?
Transparency refers to openness about how an AI system is built: data sources, model type, training process, and decision pipeline.
Explainability refers to the ability to understand why the model produced a specific output or decision.

Importance:
Transparency builds trust and accountability, while explainability ensures that users, regulators, and auditors can interpret decisions, detect biases, and challenge harmful outcomes.

Q3: How does GDPR impact AI development in the EU?
GDPR imposes strict requirements on data collection, processing, and automated decision-making. Key impacts include:
Users have the right to access their data and request correction or deletion.
Automated decisions with legal or significant effects require human oversight.
Developers must ensure lawful basis for processing, implement privacy-by-design, and minimize data use.
Clear explanations must be provided for algorithmic decisions affecting individuals.
These rules force AI developers to prioritize fairness, transparency, accountability, and privacy when designing systems.

Ethical Principles Matching

Definition	Principle
Ensuring AI does not harm individuals or society.	B) Non-maleficence
Respecting user 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

Case 1: Biased Hiring Tool (Amazon Recruiting System)

Source of Bias

The hiring model became biased because it learned from historical recruitment data dominated by male applicants. The training examples reflected existing gender imbalance in the tech industry, so the algorithm inferred that male candidates were “more suitable”. Additional contributors include:

- Biased feature selection (e.g., penalizing women-only schools or activities).

- Lack of fairness constraints during model training.
- Absence of gender-balanced evaluation datasets.

Proposed Fixes

1. **Rebalance and clean the training dataset**
   Remove gender-correlated features, ensure equal representation of male and female applicants, and eliminate historical patterns that encode discrimination.

2. **Apply fairness-aware machine learning techniques**
   Use methods such as reweighting, adversarial debiasing, or fairness constraints to prevent the model from linking gender to hiring outcomes.

3. **Implement continuous fairness monitoring**
   Track model decisions in real time to detect emerging gender disparities and enforce human oversight whenever the model flags borderline cases.

Fairness Metrics to Evaluate Post-Correction

Use a combination of quantitative fairness metrics:

Disparate Impact
Ratio Checks whether the selection rate for women is at least 80% of the rate for men.

Equal Opportunity
Difference Measures the difference in true positive rates across gender groups.

Demographic Parity
Difference Ensures both genders receive similar positive predictions overall.

Case 2: Facial Recognition in Policing

I. Ethical Risks

Wrongful arrests and misidentification
Higher error rates for minorities lead to unjust detentions, criminalization of innocent people, and long-term personal harm.

Violation of privacy and surveillance
overreach Mass collection of biometric data can be used without consent, enabling constant monitoring, profiling, or tracking.

Discrimination and systemic inequality
Biased recognition disproportionately targets racial minorities, reinforcing existing policing disparities and public distrust.

Lack of transparency and accountability
Police may rely on inaccurate outputs without understanding the model's limitations, reducing oversight and due process.

II. **Policies for Responsible Deployment**

Mandatory bias and accuracy audits
Before use, systems must undergo independent fairness evaluations across demographic groups, with transparent results.

Strict human oversight
AI predictions should never be the sole basis for arrest or investigation. Officers must validate matches manually and require multiple evidence sources.

Limited and lawful use
Deploy only for specific, high-risk cases where alternatives are not available. Avoid real-time mass surveillance.

Clear consent and data governance rules
Biometric data must be collected, stored, and deleted according to privacy laws. Individuals should know when and how data is used.

Accountability and public reporting
Police departments must publish accuracy statistics, incident reports, and audit outcomes to maintain public

Part 3 — Ethical Reflection

1) Ensuring Responsible AI in a Personal Project

In a recent personal project, I developed a predictive system to forecast product demand for a small e-commerce business. While the system primarily focuses on inventory optimization, it involves AI-driven predictions that directly affect business decisions and, indirectly, customers and employees.

To ensure the project adheres to ethical AI principles, I plan to implement the following measures:

Fairness and Bias Mitigation
I will audit the input data to identify potential biases, such as seasonal trends or regional disparities, that could disproportionately affect certain products or customer segments. If bias is detected, I will apply preprocessing techniques like reweighing or post-processing adjustments to reduce its impact on predictions.

Transparency and Explainability
The system will include interpretable outputs and visualizations, allowing business users to understand why certain demand predictions are made. This

ensures decisions are explainable and not blindly accepted, supporting trust and accountability.

Data Privacy and Security
Customer data will be anonymized and stored securely. Only aggregated or de-identified information will be used for AI training, in compliance with GDPR and other relevant regulations.

Continuous Monitoring and Feedback
I will establish a monitoring framework to track prediction errors and performance across different product categories, ensuring the system adapts to changing conditions while maintaining fairness and reliability.

Human Oversight
Critical decisions, such as restocking high-value items or discounting, will always involve human review. The AI system serves as a decision support tool rather than a fully autonomous agent.