Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.
Answer:
Algorithmic bias refers to unfair or discriminatory outcomes generated by AI systems due to issues in data, design, or deployment.

Examples:

- **Amazon's Hiring Tool:** This system penalized resumes containing phrases like "women's chess club," as it was trained on historical data that favored male candidates.
- **COMPAS Recidivism Algorithm:** This tool assigned higher risk scores to Black defendants than to white defendants with similar criminal records.

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

Transparency	Explainability
Reveals how the system was built, data sources, models used, and involved stakeholders.	Explains why a specific decision or prediction was made, in human-understandable terms.

Why they matter:

- **Transparency** builds public trust and supports regulatory compliance (e.g., GDPR).
- **Explainability** allows developers and users to understand, debug, and challenge decisions (e.g., understanding why a loan was denied).

Q3: How does the GDPR impact AI development in the EU?
Answer:
The General Data Protection Regulation (GDPR) influences AI in several ways:

- **Right to Explanation (Article 22):** Users can demand an explanation for automated decisions.

- **Data Minimization:** AI systems must only collect data that is strictly necessary.
- **Bias Audits:** Systems must be designed to avoid discriminatory outcomes.
- **Explicit Consent:** Users must actively opt in before their data is processed.

---

2. Ethical Principles Matching

Principle	Definition
A) Justice	Fair distribution of the benefits and risks of AI.
B) Non-maleficence	Ensuring AI does not cause harm to individuals or society.
C) Autonomy	Respecting individuals' control and consent over their data.
D) Sustainability	Designing AI systems that are environmentally and socially sustainable.

Part 2: Case Study Analysis

Case 1: Amazon's Biased Hiring Tool

Source of Bias:

- **Training Data:** The system was trained on resumes predominantly from male applicants.
- **Model Design:** It learned to downgrade terms associated with women, such as "women's college."

Proposed Solutions:

- **Rebalance Training Data:** Include more resumes from women.
- **Remove Gender Indicators:** Terms like “women’s college” should be neutral in the model.
- **Human Oversight:** Ensure HR reviews all AI-generated candidate shortlists.

Fairness Metrics:

- **Disparate Impact Ratio:** Aim for a ratio between 0.8 and 1.25.
- **Equal Opportunity Difference:** Ensure equal false negative rates across genders.

Case 2: Facial Recognition in Policing

Ethical Risks:

- **Wrongful Arrests:** Higher false positive rates for people of color.
- **Privacy Invasion:** Often deployed without the informed consent of those being surveilled.

Policy Recommendations:

- **Limit Use in Low-Stakes Scenarios:** Avoid deployment in situations like traffic stops.
- **Accuracy Thresholds:** Ensure the system is 99% accurate for *all* demographic groups.
- **Independent Audits:** Require third-party assessments to evaluate system fairness.

Part 3: Practical Audit

Fairness Analysis of the COMPAS Dataset

Tools Used:

- Python (pandas, matplotlib)
- IBM's AI Fairness 360 (AIF360)

Steps Taken:

1. Loaded COMPAS data and filtered by race (Black vs. white defendants).
2. Calculated key fairness metrics:
    - **Statistical Parity Difference:** 0.21 (bias against Black defendants)
    - **False Positive Rate Difference:** 0.17 (higher for Black defendants)

3. Visualized disparities using graphs (see notebook https://imgur.com/a/Vjuo7Si).

Remediation Strategies:

- **Re-train with Fairness Constraints:** Apply fairness-aware learning techniques like the Reductions algorithm in AIF360.

- **Simplify the Model:** Use interpretable models like logistic regression with fairness penalties.

Code and analysis available at: (Christopher1738/AI-Ethics)

Part 4: Ethical Reflection

Ethical Considerations for a Personal AI Project:

- **Data Auditing:** Examine datasets for representation gaps in gender, race, or other demographics.

- **Model Explainability:** Apply SHAP or LIME to interpret individual predictions.

- **Ongoing Monitoring:** Track fairness and performance metrics after deployment to detect potential drifts.

Bonus Task: Ethical AI in Healthcare Policy

Summary of Best Practices:

- **Informed Consent:** Patients must give explicit permission for AI-assisted diagnostics.

- **Bias Auditing:** Regularly test models for racial and gender disparities.

- **Transparency:** Healthcare providers should clearly explain AI-generated recommendations.

- **Human Oversight:** AI should support, not replace, medical professionals—final decisions must involve a human.