Part 1: Theoretical Understanding

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 repeatable errors in an AI system that create unfair outcomes, such as privileging one group over another. These biases often arise from skewed training data, flawed assumptions in design, or lack of representativeness in the dataset.

Examples:

1. **Hiring Algorithms**: An AI tool trained on past resumes might favor male applicants for technical jobs if historical data shows gender bias in hiring.
2. **Facial Recognition**: Some facial recognition systems have been found to be significantly less accurate for individuals with darker skin tones due to underrepresentation in training datasets.

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

- **Transparency** refers to how open and accessible information is about an AI system's design, data sources, and decision-making processes. It's about *knowing what's inside the black box*.
- **Explainability**, on the other hand, is the ability to understand and interpret why an AI system made a specific decision or prediction, especially in non-technical terms.

Why both matter:

- **Transparency** builds **trust** and accountability—stakeholders know what data and algorithms are used.
- **Explainability** supports **interpretation** and **responsibility**—particularly crucial in sectors like healthcare, finance, and law, where AI decisions affect lives.

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

The **GDPR**, enforced in the European Union, places strict rules on how personal data is collected, processed, and used. Its impacts on AI include:

- **Right to Explanation**: Individuals have the right to understand automated decisions made about them (e.g., loan denials).
- **Consent & Data Minimization**: Developers must obtain explicit consent and use only the minimum data necessary.
- **Bias & Fairness**: Developers are obligated to ensure systems do not lead to discriminatory outcomes.

GDPR encourages **ethically aligned AI** by mandating data protection, fairness, and accountability, shaping how AI is developed and deployed across the EU.

2. Ethical Principles Matching

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

1. Source of Bias

The **bias** in Amazon's AI recruiting tool stemmed primarily from the **training data** used. The model was trained on **10 years of historical hiring data**, which reflected existing male-dominated hiring practices in the tech industry. As a result, the AI system **learned to replicate gender biases**, for example:

- Penalizing resumes that included the word "women's" (e.g., "women's chess club captain")
- Downgrading graduates from all-women's colleges

Additionally, **model design** played a role: the system was not explicitly designed to detect or correct for gender bias, and **gender-specific signals were not neutralized**.

2. Proposed Fixes

To make the hiring tool **fairer**, the following actions are recommended:

A. Debias the training data

- Curate a balanced dataset that includes **equal representation of genders**, ethnic backgrounds, and educational paths.

- Remove or mask gender proxies (e.g., names, pronouns, clubs).

B. Incorporate fairness constraints during training

- Use algorithms that **enforce demographic parity** or **equal opportunity** during model optimization.

C. Implement human-AI collaboration

- Use the AI tool as a **support system**, not a decision-maker. Human recruiters should have final judgment and be trained to interpret AI outputs critically.

3. Fairness Evaluation Metrics

After making corrections, it's essential to **measure fairness** using quantitative metrics such as:

- **Disparate Impact Ratio**
  Measures whether the selection rate for different groups (e.g., male vs. female) is balanced.
- **Equal Opportunity Difference**
  Evaluates whether true positive rates are similar across groups.
- **Demographic Parity**
  Ensures outcomes are not significantly skewed in favor of one group.
- **False Positive/Negative Rate Parity**
  Checks if errors are distributed fairly between demographics.

Case 2: Facial Recognition in Policing

1. Ethical Risks

Facial recognition technology, especially when deployed by law enforcement, poses serious **ethical risks** particularly when it disproportionately misidentifies people from minority groups. Key risks include:

A. Wrongful Arrests

- Facial recognition systems have shown significantly **higher error rates for people of color**, especially Black and Asian individuals.
- These misidentifications can lead to **false accusations, arrests, or surveillance**, violating the principle of **justice and non-maleficence**.

B. Privacy Violations

- Individuals are often scanned without their **informed consent**, violating **autonomy and data rights**.
- Widespread use in public spaces contributes to **mass surveillance**, eroding personal privacy and creating a chilling effect on freedom of movement and expression.

C. Reinforcement of Systemic Bias

- If biased data (e.g., mugshots, crime statistics skewed by historic over-policing of certain communities) is used to train or validate the system, the tool may **amplify existing inequalities**.

2. Policy Recommendations for Responsible Deployment

To reduce harm and promote ethical use, the following **policies** should guide deployment:

A. Mandatory Bias Audits

- Require **independent fairness testing** before deployment. Use tools like **AI Fairness 360** to check for demographic parity and error rates by group.

B. Transparent Governance & Oversight

- Establish **civilian oversight boards** or independent AI ethics committees to review and approve use cases.
- **Public disclosure** of system performance and accountability reports should be mandatory.

C. Consent and Notification Policies

- Implement clear **opt-in systems** for facial recognition use.
- Notify individuals when and where such technologies are in use, except in strictly defined emergency cases.

D. Limited and Targeted Use

- Restrict deployment to **high-stakes scenarios** with proper judicial oversight e.g., verifying identity for violent crime investigations only.
- **Ban real-time facial recognition** in public surveillance unless authorized by court order.

E. Right to Appeal & Human Oversight

- Ensure that AI-generated matches are always **verified by a human officer**.
- Create a **clear appeals process** for individuals to contest AI-generated accusations.