**Part 1: Theoretical Understanding (30%)**

**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 in the outcomes of AI systems, often caused by biased data or flawed model design.

Examples:

1. Facial recognition systems that perform poorly on darker-skinned individuals due to underrepresentation in the training data.
2. Hiring algorithms that favor male candidates over female ones because historical data reflects gender bias in previous hiring decisions.

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

Transparency refers to how openly the inner workings and decision-making process of an AI system are disclosed, including access to data sources, algorithms, and design choices.

Explainability focuses on how well a human can understand why an AI system made a specific decision or prediction.

Both are important because transparency builds trust by allowing oversight and auditing, while explainability helps users and stakeholders make sense of AI decisions, which is crucial for accountability and ethical use.

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

GDPR impacts AI development by requiring explicit user consent for data collection and processing. It mandates data minimization and privacy by design, grants users the right to explanation when automated decisions affect them, and imposes strict penalties for non-compliance. This influences how AI systems are trained, deployed, and audited within the EU.

**Ethical Principles Matching**

Ensuring AI does not harm individuals or society. – Non-maleficence  
Respecting users’ right to control their data and decisions. – Autonomy  
Designing AI to be environmentally friendly. – Sustainability  
Fair distribution of AI benefits and risks.

**Part 2: Case Study Analysis (40%)**

**Case 1: Biased Hiring Tool**

*Scenario:*  
Amazon’s AI recruiting tool penalized female candidates.

*Tasks:*

**Source of Bias:**  
The primary source of bias was the **training data**, which reflected historical hiring patterns that favored male candidates. As a result, the model learned to down-rank resumes that included words like “women’s” or came from all-women colleges.

**Three Fixes to Make the Tool Fairer:**

1. **Rebalance the training dataset** to include equal representation of qualified male and female candidates.
2. **Remove gender-specific indicators** (e.g., names, organizations) from resumes before training.
3. **Implement fairness-aware algorithms** that adjust model predictions to ensure demographic parity or equal opportunity.

**Fairness Evaluation Metrics Post-Correction:**

* **Disparate Impact Ratio** (selection rate for females vs. males).
* **Equal Opportunity Difference** (true positive rate across genders).
* **Statistical Parity Difference** (the difference in hiring probability across gender groups).

**Case 2: Facial Recognition in Policing**

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

*Tasks:*

**Ethical Risks:**

* **Wrongful arrests** due to false matches, leading to legal and emotional consequences.
* **Privacy violations** from mass surveillance, especially without informed consent.
* **Discrimination** against marginalized communities, reinforcing systemic biases.

**Recommended Policies for Responsible Deployment:**

* Require **independent accuracy audits** across different demographic groups.
* Mandate **human oversight** before any action based on AI predictions.
* Establish **strict data governance** to limit how and where facial data is collected and used.
* Enforce **transparency requirements**, including public disclosure of system performance and limitations.

**Part 3: Practical Audit (25%)**

**Bias Audit Report: COMPAS Recidivism Dataset**

We conducted a fairness audit on the COMPAS dataset using IBM’s AI Fairness 360 toolkit. The focus was to evaluate racial bias in the risk scores that predict recidivism, especially between African-American and Caucasian defendants.

Initial analysis revealed a **significant disparate impact**: African-American individuals were more likely to be labeled as high-risk incorrectly, resulting in a **higher false positive rate** for this group. This means the system falsely predicted that African-American defendants would reoffend at a higher rate than they actually did, which can lead to **unjust legal consequences**.

The bias largely stemmed from historical data and systemic inequalities that were encoded into the training data. To address this, we applied the **Reweighing algorithm**, a preprocessing method that adjusts the weights of training examples to ensure fairer treatment of unprivileged groups.

After applying the reweighing technique and retraining the model using logistic regression, fairness metrics showed improvement. The **disparate impact** approached 1.0 (indicating parity), and the **statistical parity difference** was reduced. The overall model accuracy remained stable, suggesting that fairness improvements did not significantly reduce predictive performance.

We also visualized the **false positive rate by race**, confirming that bias was reduced after mitigation. However, some disparity remained, emphasizing the need for continuous evaluation.

**Remediation Recommendations:**

* Regular audits using fairness metrics.
* Inclusion of diverse, representative data sources.
* Use of bias mitigation algorithms during model development.
* Human review for high-stakes decisions.

In conclusion, AI systems used in criminal justice must be carefully audited and adjusted to avoid perpetuating racial biases. Ethical AI requires both technical tools and policy safeguards.

**Part 4: Ethical Reflection (5%)**

Reflecting on a past project that involved building a loan prediction model, I realize how important it is to ensure fairness and accountability. In future projects, I will prioritize ethical AI principles by carefully examining training data for biases, applying fairness metrics during development, and ensuring transparency in model decisions. I will also engage stakeholders from diverse backgrounds to guide design choices and ensure the system respects user privacy, autonomy, and societal values.

**Bonus Task (Extra 10%)**

**Policy Proposal: Ethical AI Use in Healthcare**

This guideline outlines ethical practices for implementing AI in healthcare:

**1. Patient Consent Protocols:**

* Ensure informed, explicit consent is obtained before using patient data.
* Allow patients to opt-out and understand how their data is used.
* Communicate AI’s role in decision-making clearly.

**2. Bias Mitigation Strategies:**

* Use diverse and representative datasets to train models.
* Regularly audit for bias using fairness metrics.
* Apply mitigation techniques such as reweighting or adversarial debiasing.

**3. Transparency Requirements:**

* Maintain clear documentation of AI algorithms and training data sources.
* Explain AI predictions to healthcare providers and patients.
* Disclose limitations and error rates to build trust and accountability.

Ethical AI in healthcare should enhance care without compromising human rights, equity, or trust.