1. Short Answer Questions

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

Algorithmic bias: refers to recurring, systematic mistakes made by an AI system that lead to unfair results, like giving preference to one group over another. These biases frequently result from skewed training data, incorrect model design assumptions, or data that reflect societal injustices.

Examples:

Biased Hiring Algorithms: If the training data is derived from past hiring trends that were skewed toward men, AI systems used for recruitment may give preference to men over women. For example, Amazon eliminated an AI hiring tool that penalized resumes that used the word "women's," like "women’s chess club captain."

Facial Recognition Errors: People with darker skin tones, particularly women, have been found to have substantially higher error rates in facial recognition systems. Because minority groups were underrepresented in the training datasets, one study by MIT Media Lab found that error rates for dark-skinned women could reach 34.7%, while error rates for light-skinned men were less than 1%.

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

Transparency: refers to how transparent and intelligible an AI system's internal operations are. It focuses on the model's construction, data sources, and decision-making process. A transparent model, for instance, would make its architecture, sources of training data, and design decisions visible.

Explainability: refers to the ease with which a human can comprehend the logic underlying a particular AI output or choice. Despite the complexity of the model (such as a deep neural network), it focuses on the reasons behind the AI's specific predictions or actions.

Why Both Are Important

Transparency: is crucial for accountability, auditing, and trust. It enables system analysis and the identification of possible risks or biases by developers, regulators, and users.

Explainability: is essential for ethical compliance, decision support, and user comprehension. It makes it possible for users, particularly non-experts, to comprehend and challenge AI outputs, which is crucial in industries like healthcare, law, and finance.

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

The General Data Protection Regulation (GDPR), which enforces stringent guidelines on the collection, processing, and use of personal data, has a major impact on AI development in the EU. These rules are designed to safeguard people's privacy and guarantee that data is used ethically in automated systems, such as artificial intelligence.

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

Case 1: Biased Hiring Tool

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

1. Identify the Source of Bias

The training data was the primary cause of bias in Amazon's AI hiring tool. Ten years' worth of hiring data, which primarily represented the company's male-dominated tech workforce, were used to train the model. The AI consequently learned to penalize resumes that contained terms or experiences associated with women (e.g., “women’s chess club”) and to associate successful candidates with patterns related to men (e.g., resumes from male-dominated schools or activities or resumes that contained words like “executed”).

Secondary sources:

Issues with the model's design, like its failure to take gendered language into consideration.

Absence of audits for bias during development.

2. Three Fixes to Make the Tool Fairer

a) Debias the Training Data

- Equal representation of successful resumes from both genders will help to balance the dataset.
- Eliminate any gendered indicators that might cause biased associations, such as names and organizations that are specific to a particular gender.

b) Use Fairness-Aware Algorithms

- During model training, apply algorithmic fairness constraints (such as equalized odds or demographic parity).
- To reduce the model's capacity to infer or act on gender, employ strategies such as adversarial debiasing.

c) Human-in-the-Loop Oversight

- Add audit checkpoints or manual reviews to the AI pipeline.
- Permit human recruiters to confirm or override automated judgments, particularly in cases that are unclear or indecisive.

3. Metrics to Evaluate Fairness Post-Correction

Metric	Purpose
Disparate Impact Ratio	Measures selection rates across gender; ideally close to 1:1 between groups.
Equal Opportunity Difference	Compares true positive rates (e.g., qualified candidates accepted) between genders.
Demographic Parity	Ensures outcomes are not disproportionately skewed toward a particular gender.
Fairness Confusion Matrix	Breaks down TP, FP, FN, and TN per demographic to detect imbalances.
Calibration by Group	Checks if prediction scores are equally reliable across gender groups.

Case 2: Facial Recognition in Policing

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

1. Ethical Risks

a) Wrongful Arrests and Legal Injustice

- False accusations, erroneous arrests, and incarceration can result from misidentification; this is particularly risky when law enforcement heavily relies on AI with minimal human oversight.
- Black people have been arrested in real-world instances due to incorrect facial recognition matches.

b) Discrimination and Civil Rights Violations

- Systemic inequalities and racial bias are reinforced by disproportionate mistakes.
- Impacts marginalized communities' faith in institutions.

c) Privacy Violations

- Facial recognition technology used for mass surveillance can track people without their consent, violating their privacy and sense of autonomy.
- Free speech and public protest are stifled as a result, particularly for marginalized groups.

d) Lack of Accountability

- People's rights are violated when AI decisions are opaque because it is hard to assign blame or challenge inaccurate results.

2. Recommended Policies for Responsible Deployment

a) Mandatory Bias Audits

- Demand that facial recognition systems undergo routine third-party audits to evaluate accuracy across age, gender, and race groups.
- Make audit findings openly accessible and transparent.

b) Use Restrictions in High-Stakes Scenarios

- Ban or strictly restrict the use of facial recognition technology in border control, immigration, and law enforcement until it is shown to be impartial and trustworthy.
- Prior to deployment in delicate situations, obtain judicial approval or public consultation.

c) Consent and Transparency Policies

- Notify people when facial recognition is being used.
- Give them easily accessible information about the use and storage of their data, along with clear opt-out options.

d) Human Oversight and Accountability

- Before acting on AI outputs, require human verification (e.g., no arrests without human investigation).

- Provide explicit accountability procedures for abuse or damage, along with legal redress for misidentification victims.

e) Inclusive Development Standards

- During development, require representative and varied datasets.
- Engage community members in the design and assessment process, particularly those from impacted minority groups.

Part 4: Ethical Reflection

Reflection on My AI-Based Health Follow-Up Reminder System

In order to assist clinics in lowering patient no-show rates and enhancing continuity of care, I am presently creating a Healthtech Follow-Up Reminder System that automatically sends appointment reminders via SMS or WhatsApp.

I'm implementing the following safeguards to make sure this project complies with ethical AI principles:

- Fairness and Non-Discrimination
- Transparency and Explainability
- Privacy and Data Protection
- Sustainability and Social Benefit
- Accountability and Human Oversight