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 unfair discrimination in AI systems, typically arising from biased training data or flawed algorithm design. It results in decisions that disproportionately disadvantage certain individuals or groups.

Examples:

1. Hiring Algorithms: Amazon's AI recruiting tool favored male candidates because it was trained on resumes submitted over a 10-year period, most of which came from men.
2. Facial Recognition: Systems have shown higher error rates particularly false positives for people with darker skin tones, as confirmed by a 2019 NIST study.

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

- Transparency in AI refers to the openness about how an AI system operates, including its design, data sources, and decision processes. It enables oversight and accountability.
- Explainability (often associated with Explainable AI or XAI) focuses on making the reasoning behind specific AI decisions understandable to users or stakeholders (e.g., showing which features influenced a loan denial).

Importance:
Both are crucial for building trust, ensuring accountability, and enabling users to challenge or understand automated decisions. While transparency provides a broad view of system behavior, explainability offers insight into individual outcomes—complementing each other in ethical AI deployment.

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

The GDPR significantly impacts AI development in the EU by:

- Requiring lawful, fair, and transparent processing of personal data used in AI systems.
- Granting individuals a "right to explanation" (Article 22), meaning they can request meaningful information about the logic behind automated decisions that affect them.
- Mandating data minimization, purpose limitation, and user consent, which limits how AI systems collect and use personal data.
- Encouraging privacy-preserving techniques (e.g., anonymization, federated learning) to comply with strict data protection standards.

These provisions ensure AI systems respect privacy and individual rights, promoting ethical and accountable AI development in 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

- **Scenario: Amazon's AI recruiting tool penalized female candidates.**
- **Tasks:**
    1. **Identify the source of bias (e.g., training data, model design).**
    2. **Propose three fixes to make the tool fairer.**
    3. **Suggest metrics to evaluate fairness post-correction.**

Case 1: Biased Hiring Tool

1. Source of Bias:
The primary source of bias was the training data. Amazon's AI recruiting tool was trained on resumes submitted to the company over a 10-year period, which were predominantly from male applicants. As a result, the model learned to associate male-dominated language and qualifications with "ideal" candidates and penalized resumes that included words like "women's" (e.g., "women's chess club captain") or came from all-women colleges.

2. Proposed Fixes to Make the Tool Fairer:

- Debias the Training Data:
  Audit and balance the historical dataset by either oversampling qualified female candidates or synthetically generating diverse, representative examples to reduce gender imbalance.

- Remove or Mask Gender-Indicative Features:
  Strip resumes of gender-proxies (e.g., names, gendered words, affiliations with women-only institutions) before feeding them into the model to prevent the algorithm from inferring gender.
- Apply Fairness Constraints During Model Training:
  Use fairness-aware machine learning techniques (e.g., demographic parity, equalized odds) to explicitly penalize gender-based disparities in scoring during model optimization.

3. Suggested Metrics to Evaluate Fairness Post-Correction:

- Demographic Parity:
  Compare the proportion of female vs. male candidates who receive high scores—should be roughly equal if the goal is equal opportunity.
- Equal Opportunity Difference:
  Measure whether the true positive rate (e.g., qualified candidates correctly ranked high) is similar across genders.
- Disparate Impact Ratio:
  Calculate the ratio of selection rates for female candidates to male candidates; a ratio close to 1.0 indicates fairness (e.g., 80% rule: ratio ≥ 0.8 is often considered acceptable).

These metrics should be tracked continuously during and after deployment to ensure sustained fairness.

Case 2: Facial Recognition in Policing

- **Scenario: A facial recognition system misidentifies minorities at higher rates.**
- **Tasks:**
  1. **Discuss ethical risks (e.g., wrongful arrests, privacy violations).**
  2. **Recommend policies for responsible deployment.**

1. Ethical Risks:

- Wrongful Arrests and Discrimination:
  As noted in the knowledge base (NIST, 2019), facial recognition systems have higher false positive rates for African American and Asian faces compared to Caucasian faces. This increases the risk of misidentification, leading to wrongful stops, investigations, or arrests disproportionately affecting minority communities and reinforcing systemic injustice.

- Privacy Violations:
  The use of facial recognition in public spaces enables mass surveillance without consent, infringing on individuals' right to privacy and potentially chilling free expression and movement.
- Lack of Transparency and Accountability:
  Many law enforcement agencies deploy these systems without clear disclosure or oversight. When decisions are made based on opaque algorithms, it becomes difficult for individuals to challenge errors or understand how they were targeted.
- Amplification of Social Bias:
  Biased AI in policing can perpetuate and exacerbate existing racial and social inequalities, eroding public trust particularly among marginalized groups.

2. Recommended Policies for Responsible Deployment:

- Ban or Suspend Use Until Bias Is Addressed:
  Follow the lead of cities like San Francisco and Boston by prohibiting law enforcement use of facial recognition until rigorous, independent testing confirms equitable performance across demographic groups.
- Mandate Algorithmic Audits and Bias Testing:
  Require third-party audits using diverse datasets (e.g., NIST-standard benchmarks) to evaluate accuracy across race, gender, and age before and during deployment.
- Require Judicial Oversight and Transparency:
  Any use of facial recognition in policing should require a warrant or judicial authorization, and agencies must publicly disclose when, where, and how the technology is used.
- Implement a "Right to Notification":
  Individuals misidentified or investigated based on facial recognition should be informed and granted access to appeal or correct the error.
- Adopt Strict Purpose Limitations:
  Restrict use to serious crimes only (e.g., terrorism, violent felonies) and prohibit use for general surveillance, protest monitoring, or minor offenses.

These policies align with ethical principles of non-maleficence (do no harm), justice (fair treatment), and transparency, as emphasized in the AI ethics framework provided.

Part 4: Ethical Reflection

- **Prompt: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?**

In a future personal project a mobile app that uses AI to recommend local educational resources for high school students in underserved communities in Kenya I will ensure it adheres to ethical AI principles in the following ways:

- Bias & Fairness: I will carefully curate training data to include diverse student profiles across gender, region, and socioeconomic background to avoid reinforcing existing educational inequalities. I'll test the model's recommendations for fairness across different demographic groups.
- Transparency: The app will clearly explain *why* a particular resource is being recommended (e.g., "Based on your interest in biology and past engagement with science content"). This supports informed user choices and builds trust.
- Privacy: I will minimize data collection only gathering what's essential and ensure all personal data is anonymized and stored securely. Where possible, I'll use privacy-preserving techniques like on-device processing to avoid sending sensitive data to external servers.
- Justice & Social Impact: The core goal is to *reduce*, not widen, educational gaps. I'll partner with local schools and community leaders to co-design the tool, ensuring it meets real needs and respects cultural context.
- Accountability: I'll include a simple feedback mechanism so users can report errors or biases, and I'll periodically audit the system's performance and impact.

By integrating these principles from the start, I aim to build an AI tool that is not only useful but also fair, respectful, and empowering especially for vulnerable or marginalized users.