# Part 1: Theoretical Understanding (30%)

**1. Short Answer Questions**

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

**Algorithmic bias** is systematic and unfair discrimination in AI systems, often caused by biased data, skewed representation, or flawed model assumptions.

**Examples include:**

1. **Facial recognition systems** performing poorly on darker-skinned individuals due to underrepresentation in the training dataset.
2. **Job recruitment algorithms** favoring male applicants because the model was trained on historically biased hiring data.

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

- **Transparency** refers to making the internal workings, data sources, and design choices of an AI system visible and understandable to stakeholders. It answers: *"How was this system built?"*
- **Explainability** focuses on making individual AI decisions understandable. It answers: *"Why did the system make this specific decision?"*

**Importance:**

- They help build trust between users and AI systems.
- They support accountability and auditing for errors or biases.

- They allow regulators and organizations to ensure compliance with ethical and legal standards.

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

GDPR affects AI development by:

- Requiring explicit user consent before collecting or processing personal data.
- Enforcing the right to explanation, meaning individuals can ask for justification of automated decisions.
- Demanding data minimization, ensuring only necessary data is collected.
- Imposing strict penalties for misuse, which encourages responsible and ethical AI design.

**2. Ethical Principles Matching**

Match each principle to the appropriate definition:

| Principle | Definition |
|---|---|
| **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 (40%)

**Case 1: Biased Hiring Tool**

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

**1. Identify the Source of Bias**

The bias in Amazon's recruiting tool primarily arose from **historical training data**. Over a 10-year period, most resumes came from male applicants, reflecting the existing gender imbalance in the tech workforce. As a result, the AI learned to associate certain job experiences, titles, or keywords with male candidates, systematically penalizing women. Additionally, the **model design** itself failed to consider fairness constraints features like years of experience or prior job titles indirectly encoded gender information, which the algorithm reinforced rather than mitigated.

This illustrates how even well-intentioned AI can **perpetuate societal biases** if the data and design are not carefully audited for fairness.

**2. Proposed Fixes to Make the Tool Fairer**

1. **Balanced and Augmented Training Data:** The dataset should include diverse resumes representing all genders equally. Synthetic data generation or oversampling underrepresented groups can reduce imbalance.
2. **Feature Selection and Bias Filtering:** Remove features that correlate with gender but are not job-relevant. For example, removing gendered terms, schools, or certain extracurricular indicators can prevent indirect discrimination.

3. **Incorporate Fairness-Aware Algorithms:** Use AI fairness toolkits (like IBM AI Fairness 360) to implement bias mitigation strategies during model training, such as reweighting examples, adversarial debiasing, or post-processing adjustments.

By implementing these solutions, the AI can focus on **skills and qualifications**, rather than demographic characteristics, making hiring decisions more equitable.

**3. Metrics to Evaluate Fairness Post-Correction**

- **Disparate Impact Ratio (DIR):** Measures the proportion of positive outcomes between protected and non-protected groups; closer to 1 indicates fairness.
- **Equal Opportunity Difference:** Ensures candidates with similar qualifications have equal chances of selection across groups.
- **False Positive / False Negative Rates by Gender:** Confirms that misclassification is not concentrated in one gender, preventing indirect discrimination.

**Discussion:** Fixing bias is not just a technical task it requires ongoing **auditing, monitoring, and stakeholder engagement**. AI can inadvertently reinforce societal inequalities, so fairness interventions must be continuous, not one-off.

**Case 2: Facial Recognition in Policing**

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

**1. Ethical Risks**

The ethical risks of biased facial recognition are profound:

- **Wrongful Arrests & Legal Consequences:** High misidentification rates for minorities can result in innocent people being detained, creating a direct threat to life and liberty.
- **Reinforcement of Social Inequalities:** Minority communities already disproportionately impacted by policing could face additional harm, perpetuating systemic bias.
- **Privacy Violations:** Continuous surveillance of individuals raises concerns about consent and civil liberties.
- **Erosion of Public Trust:** Communities lose faith in both law enforcement and AI technologies when they experience discriminatory outcomes.

**Discussion:** This scenario shows that AI doesn't operate in isolation; it interacts with **societal structures**. Misuse or unmitigated bias in law enforcement has serious human and social consequences.

**2. Recommended Policies for Responsible Deployment**

1. **Bias Auditing:** Regularly evaluate the system for racial and demographic bias using diverse datasets. Metrics like false positive rates, false negative rates, and predictive parity should be monitored.
2. **Human Oversight:** Ensure AI decisions are **verified by trained personnel**. No arrests should be made solely based on AI output.
3. **Transparency & Accountability:** Organizations must publish accuracy statistics, error rates, and audit reports to allow public scrutiny.
4. **Privacy Safeguards:** Collect minimal personally identifiable information, encrypt data, and anonymize records where possible.

5. **Regulatory Compliance:** Follow frameworks like GDPR, EU Ethics Guidelines for Trustworthy AI, and local laws to uphold ethical standards.

**Discussion:** Deploying AI in policing requires a **human-centric approach**. Technology can assist law enforcement, but ethical safeguards, clear accountability, and community trust are non-negotiable. Without these, AI could exacerbate injustice instead of reducing it.