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

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

**Definition:**  
**Algorithmic bias** occurs when an AI system produces systematically unfair outcomes due to biases in its design, data, or assumptions. These biases often reflect existing societal prejudices or imbalances in the training data.

**Examples:**

1. **Recruitment Systems:** An AI hiring tool trained on past resumes may prefer male candidates if historical data favored them, as in Amazon’s case where female candidates were penalized.
2. **Loan Approval Models:** A credit scoring system may deny loans disproportionately to minority applicants if the training data reflects biased historical lending practices.

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

**Transparency** refers to how open and accessible the AI system’s inner workings, data sources, and design choices are to stakeholders.

**Explainability** refers to the ability of the AI system to **clearly justify** its decisions in a way humans can understand.

**Importance:**

* Transparency builds **trust** by making AI accountable and open to scrutiny.
* Explainability ensures **users and regulators** understand how and why decisions are made, which is essential for fairness, legal compliance, and identifying bias or errors.

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

**GDPR Impacts:**

* Requires **explicit user consent** for data collection and automated decision-making.
* Grants users the **“right to explanation”** for AI-based decisions.
* Enforces **data minimization** and **privacy by design**, influencing how AI models are trained and deployed.
* Non-compliance can result in **hefty fines**, making ethical design a legal requirement.

**Ethical Principles Matching**

| **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**

**Case 1: Biased Hiring Tool**

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

**1. Identify the source of bias:**

* **Training Data Bias:**  
  The tool was trained on 10 years of historical resumes submitted to Amazon, which were predominantly from male applicants. This led the model to learn that male-associated terms or experience were more desirable.
* **Feature Engineering Bias:**  
  The system penalized resumes containing words like “women’s” (e.g., “women’s chess club”) and preferred resumes with male-dominated language or formats.

**2. Three fixes to make the tool fairer:**

1. **Debias the Dataset:**
   * Remove gender-identifying features and rebalance training data to ensure equal representation of genders and qualifications.
2. **Use Fairness Constraints in Modeling:**
   * Apply algorithms that enforce fairness metrics (e.g., equal opportunity or demographic parity) during training using fairness-aware libraries like **Fairlearn** or **AI Fairness 360**.
3. **Implement Human-in-the-Loop Oversight:**
   * Include HR experts in the decision-making process to review and validate AI recommendations and flag potential bias issues.

**3. Metrics to evaluate fairness post-correction:**

* **Disparate Impact Ratio (DIR):** Measures fairness across groups (should be close to 1.0).
* **Equal Opportunity Difference:** Compares true positive rates across genders.
* **Statistical Parity Difference:** Difference in selection rates between groups.
* **Confusion Matrix per Group:** Check if error rates differ by gender.

**Case 2: Facial Recognition in Policing**

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

**1. Ethical Risks:**

* **Wrongful Arrests:**  
  Misidentification can lead to innocent individuals, especially minorities, being wrongfully detained or arrested.
* **Violation of Privacy:**  
  Surveillance without consent or legal basis may breach individuals’ privacy rights.
* **Erosion of Public Trust:**  
  Biased systems can undermine trust in law enforcement and AI technologies.
* **Discrimination and Profiling:**  
  Such systems can reinforce systemic racism and racial profiling if not addressed.

**2. Policy Recommendations for Responsible Deployment:**

1. **Mandatory Bias Audits:**  
   Require facial recognition tools to be audited regularly using diverse, representative datasets.
2. **Consent and Oversight:**  
   Deploy only in contexts where individuals are informed and give consent, except in lawful exceptions with strict oversight.
3. **Human Verification:**  
   AI decisions must be verified by trained officers or independent reviewers before any action is taken.
4. **Transparency Reports:**  
   Agencies must publish accuracy rates, usage data, and demographic impact assessments regularly.
5. **Moratorium or Ban in High-Risk Areas:**  
   Where misidentification could cause significant harm, consider banning or pausing use until bias is addressed.