AI Theoretical and Case Study Assignment

Part 1: Theoretical Understanding (30%)

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples.

Definition: Algorithmic bias occurs when an AI system produces systematically unfair outcomes for certain groups due to biased data, model design, or societal prejudices.

Examples: 1. **Hiring Algorithms:** Penalizing female candidates due to male-dominated historical data. 2. **Facial Recognition:** Higher error rates for darker-skinned individuals due to underrepresentation in training datasets.

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

- **Transparency:** Openness about AI system design, data, and decision-making processes.
- **Explainability:** The AI system’s ability to provide understandable reasoning for its outputs.

Importance: - Transparency ensures accountability and enables auditing. - Explainability builds trust and helps stakeholders make informed decisions.

Q3: How does GDPR impact AI development in the EU?

- Enforces lawful, transparent, and secure handling of personal data.
- Grants user rights: access, correction, deletion of personal data.
- Encourages data minimization, consent, and explainable AI.

---

2. Ethical Principles Matching

Principle	Definition
Justice	Fair distribution of AI benefits and risks.
Non-maleficence	Ensuring AI does not harm individuals or society.
Autonomy	Respecting users’ right to control their data and decisions.
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.

Source of Bias: - Training data bias: historical male-dominated resumes. - Model design: overemphasis on features correlated with male candidates.

Proposed Fixes: 1. Rebalance training data to equally represent genders. 2. Remove gender-related features from the model. 3. Apply fairness-aware algorithms to ensure equal opportunity.

Fairness Metrics: - Demographic parity - Equal opportunity - False negative/positive rates by gender

---

Case 2: Facial Recognition in Policing

Scenario: Misidentification rates higher for minorities.

Ethical Risks: - Wrongful arrests - Discrimination against marginalized communities - Privacy violations

Recommended Policies: 1. Human oversight for critical decisions. 2. Regular audits and accuracy benchmarking across demographics. 3. Strict consent, data retention, and usage policies. 4. Transparency reports detailing errors and biases.

Reflections

Ethical AI requires both technical and human considerations. Addressing bias, ensuring explainability, and respecting privacy are essential to build trustworthy AI systems. Real-world case studies, such as biased hiring tools or facial recognition errors, highlight the consequences of neglecting these principles.