AI Ethics Assignment

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 occurs when an AI system produces systematically prejudiced results due to erroneous assumptions in the machine learning process, imbalanced training data, or flawed model design. It often leads to **unfair outcomes** that disadvantage certain groups of people based on characteristics such as race, gender, age, or socioeconomic status.

Case Study Example 1: Facial Recognition and Racial Bias

System: Commercial facial recognition systems (e.g., Amazon Recognition, IBM, Microsoft)

Issue: A 2018 MIT Media Lab study revealed that facial recognition systems had error rates of:

- **0.8% for light-skinned men**
- **Up to 34.7% for dark-skinned women**

Reflection: This disparity was primarily due to **underrepresentation of darker-skinned individuals** in the training datasets. As a result, these systems performed poorly on marginalized populations, raising ethical and legal concerns in areas like surveillance and law enforcement. The issue pushed major tech companies to pause or halt facial recognition deployments.

Case Study Example 2: Hiring Algorithms and Gender Bias

System: An AI recruitment tool used by Amazon

Issue: Amazon developed an internal AI system to screen resumes. It was discovered that the tool **penalized resumes that included the word "women's"** (e.g., "women's chess club captain") and favoured male-dominated language.

Reflection: The model was trained on historical hiring data, which reflected **existing gender imbalances** in the tech industry. The AI learned to prefer male applicants, unintentionally reinforcing gender discrimination. Amazon eventually scrapped the tool, highlighting the importance of **bias auditing** in hiring algorithms.

Key Reflection:

Algorithmic bias is not just a technical flaw—it is an ethical and societal risk. If left unchecked, it can **amplify discrimination** in critical areas like healthcare, policing, finance, and employment. Responsible AI development must include:

- Diverse datasets
- Bias testing tools (like IBM AIF360)
- Human oversight

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

- **Transparency:** Refers to how openly the AI system's design, data, and processes are shared. It means users can see how the system works.
- **Explainability:** Refers to how easily a human can understand the reasoning behind specific AI decisions or outputs**.**

Key Difference:

- **Transparency** is about access to information.
- **Explainability** is about understanding the AI's behavior.

Case Study 1: Transparency — Cambridge Analytica Scandal

- **What Happened:** Facebook user data was harvested without consent and used to influence elections.
- **Issue:** Lack of transparency in how user data was collected and used by AI algorithms for political targeting.
- **Analysis:** If Facebook's data practices and AI systems had been more transparent, the misuse could have been prevented or detected earlier. Transparency ensures ethical standards and legal compliance.

Case Study 2: Explainability — AI in Healthcare (IBM Watson for Oncology)

- **What Happened:** IBM Watson recommended unsafe cancer treatments in some cases.
- **Issue:** Doctors couldn't understand how Watson arrived at certain decisions.
- **Analysis:** The lack of explainability made it difficult for doctors to trust or verify AI outputs. Explainability is crucial in high-stakes fields where lives are involved.

Why Both Matter

- **Transparency** builds public trust and allows oversight.
- **Explainability** ensures users can interpret and challenge AI decisions, especially in critical areas like healthcare, finance, or criminal justice.

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

GDPR affects AI development by enforcing **data protection, transparency, and user rights**, which shapes how AI systems are designed and deployed in the EU.

Key Impacts on AI:

1. **Consent:** AI systems must get clear user permission before collecting personal data.
2. **Right to Explanation:** Users can ask for **reasons behind AI decisions** (Article 22).
3. **Data Minimization:** Only necessary data should be used for AI training.
4. **Accountability:** Developers must ensure ethical use of data and document decision processes.
5. **Fines for Violations:** Up to €20 million or 4% of global revenue.

Case Study: Google GDPR Violation (France, 2019)

- **What happened:** Google was fined €50 million for unclear data processing in AIbased ad targeting.
- **Impact:** Highlighted the need for transparency and proper consent in AI systems.

Reflection:

GDPR pushes AI developers to build responsible, privacy-focused systems, promoting trust and fairness in AI use across the EU.

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.