AI Ethics Assignment

Part 1: Theoretical Understanding (30%)

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

Algorithmic bias refers to systematic and repeatable errors in AI systems that create unfair outcomes, often disadvantaging particular groups. It typically arises from biased training data, flawed assumptions, or discriminatory model designs.

Examples:

1. **Hiring Systems** – An AI tool trained on historical data that favored male applicants may learn to penalize resumes with female-coded language.
2. **Credit Scoring** – Algorithms trained on financial histories may unfairly reduce creditworthiness scores for minorities due to historical inequities in loan approvals.

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

- **Transparency** refers to the openness of AI processes, such as how the system is built, what data it's trained on, and how decisions are made.
- **Explainability** refers to the ability to interpret and understand individual AI decisions in a human-comprehensible way.

Why both matter: Transparency builds **trust** and allows **external scrutiny**, while explainability empowers **users and regulators** to understand specific outcomes and **challenge unfair decisions**. Together, they support accountability and ethical compliance.

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

GDPR mandates strict **data protection and privacy** standards for AI systems operating in the EU. It impacts AI development by:

- Requiring **informed consent** for data collection and processing.
- Enforcing the **right to explanation** for automated decisions.
- Encouraging **data minimization** and **purpose limitation**.

This ensures AI respects individual rights and operates within ethical and legal bounds.

Ethical Principles Matching

Principle	Definition
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.
A) Justice	Fair distribution of AI benefits and risks.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

Bias Source:

- **Training data** – Historical data reflecting gender biases in hiring practices led to penalization of resumes with female-associated terms.

Proposed Fixes:

1. **Re-train the model** on a **balanced, de-biased dataset** that includes diverse candidates equally.
2. **Remove gender indicators** from features (e.g., names, pronouns, gendered activities).
3. Implement a **fairness-aware algorithm**, such as adversarial debiasing.

Fairness Evaluation Metrics:

- **Disparate Impact Ratio**
- **Equal Opportunity Difference**
- **Demographic Parity**

Case 2: Facial Recognition in Policing

Ethical Risks:

- **Wrongful arrests** due to higher false positive rates for minorities.
- **Privacy violations** from mass surveillance.
- **Loss of public trust** in law enforcement and technology.

Policy Recommendations:

1. Mandate **independent bias audits** before deployment.
2. Prohibit use in **high-stakes scenarios** (e.g., arrests) without human oversight.
3. Enforce **strict consent and data handling policies**.
4. Adopt **community transparency reports** and public accountability.

Part 3: Practical Audit (25%)

COMPAS Dataset Audit (Summary Report)

Toolkit Used: AI Fairness 360
Dataset: COMPAS Recidivism Dataset
Goal: Analyze racial bias in risk prediction scores.

Approach Summary:

- Measured **false positive rate** and **equal opportunity difference** across racial groups.
- Visualized **disparate impact ratios** using Matplotlib.
- Used AI Fairness 360's BinaryLabelDataset and MetricFrame.

Key Findings:

- The model had a significantly **higher false positive rate for Black defendants**.
- **Equal opportunity difference** indicated unfair advantage toward White individuals.
- **Disparate impact** was below the acceptable threshold (0.8), indicating racial bias.

Remediation Steps:

1. **Reweighing or adversarial debiasing** pre-processing techniques.
2. Use **fair classifiers** (e.g., prejudice remover).
3. Implement **post-processing adjustments** like Reject Option Classification.

Part 4: Ethical Reflection

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

Reflection:
In a future project involving predictive analytics for student performance, I will apply ethical AI principles by:

- Ensuring **data diversity** across schools and demographics.
- Incorporating **fairness checks** using tools like AI Fairness 360.
- Making predictions **explainable** to educators and parents.
- Obtaining **explicit consent** for data collection.
- Designing the system to **enhance support**, not penalize students.

This will help foster a **trustworthy and equitable** system that truly benefits learners.