Ethical AI Assessment

Part 1: Theoretical Understanding *(30%)*

1. Short Answer Questions

◆ **Q1: What is Algorithmic Bias? Provide two examples.**

Definition: Algorithmic bias occurs when an AI system produces unfair or discriminatory outcomes due to biases embedded in its training data, model design, or deployment context.

Examples:

- **Hiring algorithms** trained on past recruitment data may favor male applicants, replicating historical gender bias.
- **Facial recognition tools** misidentify people of color more frequently than white individuals, leading to wrongful arrests or misclassification.

◆ **Q2: Explain the difference between Transparency and Explainability in AI. Why are both important?**

Transparency: The extent to which the inner workings of an AI system—its data sources, structure, and training methods—are open to inspection.

Explainability: The ability of an AI system to communicate its decisions in a way humans can understand.

Importance: Together, they enable **trust**, **accountability**, and **user empowerment**. Without transparency, it's difficult to audit a system. Without explainability, it’s difficult to contest or interpret a decision.

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

Impact: GDPR enforces strict rules around the use of personal data, which directly affects how AI systems are built and deployed in the EU.

Key points:

- Requires **user consent** before processing data
- Ensures a **right to explanation** for automated decisions
- Restricts **profiling** without justification
- Mandates **data minimization** and protection through encryption and access controls
- **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 *(40%)*

🧠 Case 1: Biased Hiring Tool

Scenario: Amazon’s internal AI recruiting tool penalized female candidates, even when their qualifications were similar to male applicants.

🔹 **Source of Bias**

- **Training Data:** Historically male-dominated hiring data led the model to rank male applicants more favorably.

- **Model Design:** The AI learned from resumes that contained male-centric patterns or keywords.
- **Feature Selection:** Gender-correlated features (like certain colleges or activities) influenced scores indirectly.

🔧 Fixes to Improve Fairness

1. **Data Rebalancing:** Reconstruct the training dataset to include equal representation of genders across roles and industries.
2. **Algorithmic Debiasing:** Use fairness-aware techniques such as adversarial debiasing, reweighting, or reject option classification.
3. **Feature Monitoring and Filtering:** Audit features for gender proxies (e.g. "captain of men's soccer") and exclude sensitive or correlated inputs.

📊 Fairness Metrics to Evaluate Corrections

- **Statistical Parity:** Measures whether different groups receive outcomes at similar rates.
- **Disparate Impact Ratio:** Compares selection rates between groups (e.g. female vs. male candidates).
- **Equal Opportunity Difference:** Assesses true positive rates across groups to ensure fairness in favorable outcomes.

🕵️‍♀️ Case 2: Facial Recognition in Policing

Scenario: A police department deploys facial recognition software that misidentifies people of color at disproportionately high rates, leading to wrongful detainment.

⚠️ Ethical Risks

- **Wrongful Arrests:** Misidentification can lead to unjust legal consequences and trauma.

- **Discriminatory Surveillance:** Minority communities may be over-policed, infringing civil liberties.
- **Privacy Violations:** Widespread, unconsented facial tracking erodes individual privacy.

🛡️ Responsible Deployment Policies

1. **Bias Auditing & Public Reporting:** Mandate third-party testing before deployment and publish impact assessments.
2. **Usage Restrictions:** Limit usage to pre-approved cases such as missing persons or active threats—not mass surveillance.
3. **Informed Consent & Oversight:** Include opt-in participation for public datasets and require civic advisory boards to govern AI use.
4. **Ethical AI Assessment**
5. **Part 3: Practical Audit — COMPAS Dataset *(25%)***
6. **💻 Dataset Audit using AI Fairness 360**

📊 Summary Report (300 Words)

The audit of the COMPAS Recidivism Dataset using IBM's AI Fairness 360 revealed significant racial disparities in predictive outcomes. Specifically, African-American defendants showed a notably higher false positive rate than Caucasian defendants, suggesting they were wrongly flagged as high-risk more often.

Disparate impact and equal opportunity metrics highlighted imbalances in how the model treated racial groups, with fairness scores falling below acceptable thresholds. These results underscore how historical biases embedded in training data can perpetuate harm when used in sensitive domains like criminal justice.

To mitigate the bias, we applied Equalized Odds Postprocessing, which adjusted predictions to align fairness across groups while maintaining reasonable accuracy. Post-correction metrics showed meaningful improvement in equal opportunity and reduction in disparity scores.

Remediation efforts also included transparency in feature engineering, ethical evaluation of proxy variables (e.g., age, charge degree), and periodic bias assessments using updated data.

Moving forward, fairness auditing must be a core part of AI development pipelines in criminal risk scoring. Public agencies should demand model explainability, support oversight frameworks, and establish ethical safeguards to prevent wrongful predictions that could affect people's liberty and livelihoods.