AI Ethics Assignment – Final Report

Part 1: Theoretical Understanding

Short Answer Questions

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

Answer:
Algorithmic bias occurs when an AI system produces systematically unfair outcomes due to skewed data, flawed model design, or implicit societal biases.

Examples:

1. **Hiring Algorithms**: A recruiting AI may favour male candidates if trained on historical data dominated by male hires.
2. **Facial Recognition**: Systems often misidentify people of colour due to underrepresentation in training datasets.

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

Answer:

- **Transparency** refers to how openly an AI system's design, training data, and decision-making processes are disclosed.
- **Explainability** is the degree to which a human can understand why an AI made a specific decision.

Importance:
Transparency builds trust and enables accountability, while explainability ensures users can interpret and challenge decisions—critical in high-stakes domains like healthcare or criminal justice.

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

Answer:
The **General Data Protection Regulation (GDPR)** enforces:

- User consent for data use.

- Right to explanation for automated decisions.
- Data minimization and anonymization.

This compels AI developers to prioritize data privacy, transparency, and accountability, often requiring algorithmic adjustments or human oversight.

Ethical Principles Matching

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

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon AI)
Scenario:
Amazon's AI recruiting tool penalized female candidates, favoring male-dominated resumes based on historical hiring data.

1. Identify the Source of Bias

- **Training Data Bias:** Historical resumes predominantly from male applicants led the model to associate male terms (e.g., "executed," "captured") with success.
- **Feature Selection:** The model learned to downgrade resumes containing terms linked to women's experiences (e.g., "women's chess club").
- **Feedback Loop:** Reinforced gender stereotypes due to biased hiring outcomes being used to retrain the system.

2. Three Fixes to Make the Tool Fairer

1. **Bias-Aware Data Curation**
   - Remove gender-indicative language or attributes from training data.
   - Balance the dataset by including equal numbers of successful candidates of different genders.
2. **Fairness-Aware Modeling Techniques**
   - Apply algorithms like reweighing or adversarial debiasing (e.g., using AI Fairness 360).
   - Include fairness constraints during model training.

3. **Human-in-the-Loop Evaluation**
    - Add diverse human reviewers to audit AI decisions before they are implemented.
    - Use an ensemble model combining AI scoring with human judgment.
4. **Fairness Evaluation Metrics**
    - Disparate Impact Ratio: Compares selection rates for different gender groups.
    - Equal Opportunity Difference: Compares true positive rates across genders.
    - Demographic Parity: Checks if both genders have equal likelihood of being shortlisted.
    - Calibration by Group: Ensures predicted probabilities are equally accurate across genders.

Case 2: Facial Recognition in Policing

Scenario:

Facial recognition systems disproportionately misidentify individuals from minority communities, leading to privacy violations and wrongful arrests.

1. **Ethical Risks**
    - **Wrongful Arrests:** Misidentification can lead to unjust detainment and legal consequences.
    - **Privacy Invasion:** Continuous surveillance without consent undermines civil liberties.
    - **Racial Profiling:** Reinforces systemic biases against marginalized groups.
    - **Lack of Accountability:** Black-box systems make it difficult to trace errors and assign responsibility.
2. **Policy Recommendations for Responsible Deployment**
    1. **Ban in High-Risk Areas Until Proven Fair**
        - ➢ Restrict use in law enforcement unless the system passes independent fairness audits.
    2. **Mandatory Bias Audits & Public Reporting**
        - ➢ Evaluate error rates across demographic groups and publish findings.
    3. **Consent-Based Deployment**
        - Inform communities and gain consent before system implementation in public spaces.
    4. **Human Oversight & Appeal Mechanism**
        - Ensure AI decisions are reviewed by humans and appeal processes are in place.

5. **Transparency & Explainability**

   - Disclose model training data sources, design, and decision-making logic.

Part 3: Bias Audit Report – COMPAS Dataset

The COMPAS Recidivism Dataset was analyzed to investigate racial bias in risk assessments for re-offending. Using IBM's AI Fairness 360 toolkit, we audited the dataset and found significant disparities in how individuals were scored based on race.

Initially, metrics like Disparate Impact and Mean Difference revealed systemic bias favoring Caucasian individuals over African-American defendants. The False Positive Rate (FPR) was notably higher for African-Americans, suggesting they were more often incorrectly predicted to re-offend.

To mitigate this, we applied a reweighing algorithm to the training dataset. This technique adjusts the data distribution to neutralize the effect of sensitive attributes (in this case, race). Post-training, we retrained a logistic regression classifier and re-evaluated it.

The results showed moderate improvements. While the false positive rate difference remained non-zero, it was reduced, and the disparate impact approached parity, indicating fairer predictions.

In conclusion, while COMPAS demonstrates predictive power, it also inherits and perpetuates racial bias. For real-world applications, bias mitigation techniques like reweighing or adversarial debiasing should be standard. In addition, transparency, continuous auditing, and human oversight are essential to ensure that AI models do not reinforce systemic inequalities.

Part 4: Ethical Reflection

In a recent project, I designed a predictive model for hospital readmission risk, aiming to help healthcare providers optimize follow-up care. As I continue to develop this model, I am committed to ensuring it adheres to ethical AI principles.

First, I will prioritize fairness and inclusion by auditing the dataset for demographic imbalances—especially across age, race, and gender. Tools like AI Fairness 360 will help me identify and mitigate biases before model deployment.

Second, I will uphold transparency and explainability by choosing interpretable models or integrating model explanation tools such as LIME or SHAP. These will help clinicians understand why the AI flagged a patient as high-risk, facilitating trust and better decision-making.

Third, I’ll integrate data protection and user autonomy by anonymizing sensitive patient data and complying with data privacy laws such as the POPIA and GDPR. Patients' consent will be required before their data is used in training or testing.

Lastly, I’ll seek interdisciplinary feedback from healthcare professionals and ethics advisors throughout the project lifecycle to ensure ethical blind spots are addressed early.

By embedding these ethical principles into the project workflow, I aim to create a socially responsible and trustworthy AI solution for the healthcare sector.

Bonus Task: Ethical AI Guideline for Healthcare

1. Patient Consent Protocols

- ➢ Informed Consent: Patients must be clearly informed when AI is used in diagnosis, decision support, or treatment recommendations.
- ➢ Data Usage Clarity: Patients should know how their data will be collected, stored, and analyzed.
- ➢ Right to Opt-Out: Patients must have the right to refuse AI-based processing without compromising access to standard care.

2. Bias Mitigation Strategies

- ➢ Diverse and Representative Datasets: Ensure datasets include various demographic groups (race, gender, age, disability status).
- ➢ Regular Bias Audits: Use tools (e.g., AI Fairness 360, Fairlearn) to detect and correct disparities.
- ➢ Algorithmic Rebalancing: Apply reweighing, adversarial debiasing, or oversampling to reduce training bias.
- ➢ Inclusive Design: Engage healthcare workers, ethicists, and patient groups in design/testing.

3. Transparency Requirements

- ➢ Explainable AI (XAI): Use interpretable models or tools (e.g., SHAP, LIME) to clarify predictions.
- ➢ Audit Trails: Maintain decision and data access logs for accountability.
- ➢ Open Communication: Clearly communicate limitations and uncertainty.
- ➢ Governance Oversight: Form an ethics board to assess AI tools regularly.

Conclusion:
This guideline ensures AI in healthcare supports equitable treatment, safeguards patient rights, and aligns with public trust and ethical governance. Responsible development and

deployment of AI tools will contribute to a healthcare system that is both innovative and just.