AI Ethics Assignment – Designing Responsible and Fair AI Systems

Theoretical Understanding

Q1: Algorithmic Bias

Definition:
Algorithmic bias occurs when an AI system systematically produces unfair or discriminatory outcomes against certain groups due to issues in training data, model design, or human assumptions embedded in the system.

Examples:

1. **Hiring Algorithms:** A recruitment AI penalizes female candidates because historical data reflects gender bias in past hires.
2. **Facial Recognition Systems:** Misidentifying minority ethnic groups at higher rates due to underrepresentation in training datasets.

Q2: Transparency vs Explainability

- **Transparency:** Refers to the openness of the AI system's processes, design, and decision-making logic. It allows stakeholders to see what data and algorithms are used.
- **Explainability:** Refers to the ability to provide human-understandable reasons for individual AI predictions or decisions.

Importance:

- Ensures accountability and trust in AI.
- Allows detection and correction of biases.
- Required for regulatory compliance (e.g., GDPR).

Q3: GDPR Impact on AI in the EU

- Requires **explicit user consent** for collecting and processing personal data.
- Grants' users the **right to explanation** for automated decisions.
- Enforces **data protection by design**, limiting use of sensitive data in AI systems.
- Encourages **auditability**, accountability, and data minimization in AI models.

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

Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

Scenario: AI recruiting tool penalized female candidates.

Source of Bias:

- Historical hiring data biased toward men.
- Model optimizes for past hires without fairness constraints.

Proposed Fixes:

1. Balance the training dataset to include equal representation of genders.
2. Implement fairness-aware algorithms (e.g., reweighing, adversarial debiasing).
3. Introduce human oversight in the final hiring decisions.

Fairness Metrics:

- **Demographic parity:** Equal hiring rates across genders.
- **Equalized odds:** Similar false positive and negative rates across groups.
- **Predictive parity:** Same predictive accuracy across groups.

Case 2: Facial Recognition in Policing

Scenario: System misidentifies minorities at higher rates.

Ethical Risks:

- Wrongful arrests and discrimination.
- Violation of privacy and consent.
- Erosion of public trust in law enforcement.

Responsible Deployment Policies:

1. Limit use to critical law enforcement scenarios.
2. Require human verification of AI predictions.
3. Publish system accuracy and bias audits publicly.
4. Ensure diverse training datasets and perform regular audits.

Summary Report for the Dataset Audit:

Analysis of the COMPAS dataset revealed a significant racial disparity in recidivism risk scores. African American defendants had higher false positive rates compared to Caucasian defendants, implying a risk of unjust higher sentencing. Visualizations showed disproportionate risk score distributions. To mitigate bias, the Reweighing algorithm was applied, adjusting sample weights to equalize the importance of each racial group. Post-mitigation metrics showed reduced disparate impact, demonstrating a fairer prediction model. Continuous monitoring is recommended for responsible deployment. This process highlights the importance of fairness audits and bias mitigation in AI systems affecting human decisions.

Ethical Reflection:

In my predictive credit risk project, I will ensure ethical AI by auditing datasets for gender and racial bias, documenting decision logic for transparency, and obtaining user consent for sensitive data. I will implement fairness-aware models and evaluate predictions across different groups to prevent discrimination. Regular monitoring and human oversight will ensure alignment with ethical principles, fostering trust and responsible AI use.

Ethical AI in Healthcare

To ensure the responsible use of AI in healthcare, patient consent is of paramount importance. All patients must provide explicit opt-in consent for data collection and the use of AI-assisted treatments. Additionally, patients should receive clear and understandable explanations of how AI systems inform care decisions, including the role of predictive algorithms in diagnosis, treatment recommendations, or risk assessments. This transparency empowers patients to make informed decisions regarding their care.

Bias mitigation is another critical pillar of ethical AI in healthcare. Regular audits must be conducted to detect demographic disparities in AI predictions, ensuring that no patient group is unfairly disadvantaged. Efforts should be made to balance training datasets and, where necessary, employ synthetic oversampling techniques to improve representation for underrepresented groups. Such measures help maintain fairness and equity across diverse patient populations.

Transparency requirements must be strictly upheld. AI outputs should be explainable not only to clinicians but also to patients, allowing both parties to understand the rationale behind automated recommendations. Furthermore, decision logs must be maintained and readily accessible, providing a clear record of AI-assisted decisions to support accountability and oversight.

Finally, review and accountability mechanisms must be established. A dedicated oversight committee should monitor AI deployment, review system performance, and investigate any errors or adverse outcomes. Continuous monitoring is essential to ensure compliance with ethical standards, regulatory requirements, and legal frameworks, thereby fostering trust in AI systems while safeguarding patient welfare.