AI Ethics Assignment: Designing Responsible and Fair AI Systems
PLP Group
July 12, 2025

Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.
Algorithmic bias refers to systematic and unfair discrimination in AI systems, often arising from biased training data or flawed algorithms. It occurs when AI outputs disproportionately disadvantage certain groups.
Examples:

- **Hiring Algorithms:** Amazon's AI recruiting tool was biased against women, as it was trained on male-dominated resumes, penalizing terms like "women's."
- **Facial Recognition:** NIST (2019) found facial recognition systems had higher false positive rates for African American and Asian faces compared to Caucasian faces, leading to misidentification risks.

Q2: Explain the difference between transparency and explainability in AI.
Why are both important?
Transparency refers to the openness of an AI system's processes, allowing stakeholders to understand how it functions and is governed. Explainability focuses on clarifying how specific decisions are made, often through techniques like feature importance in models.
Importance: Transparency builds trust and accountability by ensuring systems are auditable, while explainability enables users to understand and challenge decisions, reducing risks of harm and fostering ethical use.

Q3: How does GDPR impact AI development in the EU?
The EU's GDPR mandates strict data protection, requiring AI developers to ensure data minimization, obtain explicit consent, and protect user privacy. It includes a "right to explanation" for automated decisions, compelling developers to prioritize transparency and explainability. Non-compliance risks hefty fines, pushing companies to adopt privacy-preserving techniques like federated learning.

2. Ethical Principles Matching

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

Case 1: Biased Hiring Tool

Source of Bias: The bias in Amazon's AI recruiting tool stemmed from training data dominated by male resumes, causing the model to associate male-centric terms with higher suitability.
Three Fixes:

- **Diverse Training Data:** Collect balanced datasets with equal representation of genders and backgrounds.
- **Bias-Aware Algorithms**: Use fairness-aware algorithms, like adversarial training, to minimize gender correlations.
- **Regular Audits:** Implement periodic audits to detect and correct biases, using metrics like demographic parity.

Fairness Metrics: Measure disparate impact ratio (ratio of favorable outcomes across groups) and equal opportunity difference (equality in true positive rates) to ensure fairness post-correction.

Case 2: Facial Recognition in Policing

Ethical Risks: Higher misidentification rates for minorities increase risks of wrongful arrests, perpetuate systemic bias, and infringe on privacy through excessive surveillance.
Policies for Responsible Deployment:
1. Mandatory Testing: Require rigorous testing across diverse demographics to ensure equitable accuracy.
2. Transparency Protocols: Publicly disclose system limitations and error rates.
3. Human Oversight: Mandate human review of AI outputs to prevent automated decisions leading to harm.
4. Community Consent: Engage affected communities to ensure deployment aligns with public interest.