AI Ethics Assignment

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

Algorithmic bias refers to systematic and repeatable errors in an AI system that lead to unfair outcomes, such as privileging one group over another. These biases often arise from biased training data, flawed model assumptions, or skewed societal representations encoded into algorithms.

Example 1: *Amazon's AI recruitment tool* penalized resumes that included the word "women" (e.g., "women's chess club"), because the model was trained on historical data dominated by male applicants, learning to favour male-coded language.

Example 2: *The COMPAS algorithm* used in the US judicial system assigned higher recidivism risk scores to Black defendants compared to white defendants, even when controlling for similar criminal histories, revealing racial bias in risk assessment tools.

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

Transparency in AI refers to the openness of the system's design, data sources, and development processes. It allows stakeholders to understand *how* the system was built, what data it uses, and what objectives it pursues.

Explainability, on the other hand, focuses on a system's ability to justify or make understandable its decisions to end-users. This involves techniques that make complex models interpretable—such as highlighting which features contributed to a prediction.

Both are essential for building **trustworthy AI**. Transparency ensures accountability and regulatory compliance, while explainability allows users to contest or understand decisions—especially in high-risk domains like healthcare, finance, and criminal justice.

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

GDPR plays a significant role in shaping ethical AI practices in the EU by enforcing **data protection**, **user rights**, and **automated decision-making constraints**.

- **Article 5** mandates data minimization, accuracy, and purpose limitation, forcing developers to limit unnecessary personal data in training sets.
- **Article 22** gives individuals the right not to be subject to decisions based solely on automated processing, including profiling, unless explicit consent is given or other legal safeguards exist.

- **Article 15** enforces a “right to explanation,” meaning users can ask why an automated decision was made.

As a result, AI systems in the EU must be designed with **privacy-by-design**, use interpretable models where possible, and incorporate human oversight. This restricts opaque black-box models in critical applications and prioritizes fairness, consent, and transparency.

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.

Case Study Analysis

Case Study 1: Biased Hiring Tool

Scenario: Amazon’s AI recruiting system was trained on historical hiring data, which resulted in penalizing resumes that included references to women or women’s organizations.

---

1. Identify the source of bias

The primary source of bias in Amazon’s hiring tool was **biased training data**. Historical hiring data was skewed toward male applicants, reflecting systemic gender inequality in the tech industry. As a result, the algorithm learned to downgrade resumes containing indicators associated with women (e.g., “women’s chess club,” “women’s university”).

Additional contributing factors:

- **Feature representation bias**: Words or phrases linked to female identity were interpreted as less favourable.
- **Lack of diversity in model validation**: The system was not rigorously tested across gender groups before deployment.

---

2. Propose three fixes to make the tool fairer

1. **Debias the training dataset**: Rebalance training samples to include equal representation of successful candidates across genders. Remove features that encode gender directly or indirectly.

2. **Introduce counterfactual fairness techniques**: Apply methods that evaluate whether the decision would remain the same if the applicant's gender were different, holding all else equal.
3. **Human-in-the-loop screening**: Integrate human reviewers into the selection process, especially in edge cases or ambiguous evaluations, to mitigate potential automated bias.

---

3. Suggest metrics to evaluate fairness post-correction

- **Disparate Impact Ratio (DIR)**: Measures the ratio of positive outcomes between protected and unprotected groups. A value close to 1 indicates fairness.
- **Equal Opportunity Difference**: Compares true positive rates between groups to ensure equal access to favourable outcomes.
- **Statistical Parity Difference**: Evaluates the difference in selection rates for different groups regardless of true outcome.

Case Study 2: Facial Recognition in Policing

Scenario: Facial recognition systems used in law enforcement have shown higher false positive rates when identifying people of color, leading to risks such as wrongful arrests.

---

1. Discuss ethical risks

- **Wrongful Arrests**: Misidentification can lead to unjust detainment, particularly affecting minority populations disproportionately.
- **Violation of Privacy**: Deploying facial recognition in public spaces without consent undermines individual privacy rights and fosters surveillance culture.
- **Erosion of Public Trust**: Lack of transparency in deployment can reduce trust in institutions, especially among already marginalized communities.
- **Reinforcement of Systemic Bias**: If used with already biased law enforcement databases, facial recognition systems can amplify discriminatory practices.

---

2. Recommend policies for responsible deployment

1. **Mandatory Bias Audits**: Regular independent evaluations of system performance across race, gender, and age groups before and during deployment.

2. **Clear Consent and Notification Protocols**: Inform individuals when and where facial recognition is used, with options to opt out in non-critical applications.
3. **Strict Use Regulation**: Limit use to specific, high-justification cases such as missing person searches, not routine public surveillance.
4. **Accountability Framework**: Create appeal processes for misidentification and assign legal liability for misuse or harm caused by the technology.
5. **Transparency Reports**: Agencies using facial recognition must publish public performance metrics, error rates, and deployment cases.