Week 7 Assignment

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 is systematic error in AI that creates unfair outcomes, often mirroring societal biases present in training data. It manifests as gender/racial bias in hiring systems that penalize certain demographics, and lower accuracy in facial recognition for certain groups due to non-diverse training data.

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

Transparency is the openness about an AI system's design, data, and processes (the what and how). Explainability is providing clear reasons for a specific, individual decision (the why). Both are vital for building trust, ensuring accountability, and debugging issues like bias.

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

The GDPR significantly impacts EU AI development by enforcing strict rules on personal data processing. Key impacts include mandatory data minimization, the need for fairness and transparency in models, and the "right not to be subject to solely automated decisions, which is the primary regulatory driver compelling developers to implement Explainable AI (XAI).

2. Ethical Principles Matching

Match the following principles to their definitions:

A) and 4: *Justice*- Fair distribution of AI benefits and risks.

B) and 1: *Non-maleficence*- Ensuring AI does not harm individuals or society.

C) and 2: *Autonomy*- Respecting users' right to control their data and decisions.

D) and 3: *Sustainability*- Designing AI to be environmentally friendly.

Part 2: *Case Study Analysis*

Case 1: *Biased Hiring Tool*

1. ***Identify the source of bias (e.g., training data, model design)***

★ **Historical Data Reflection:** The model was trained on **10 years of historical résumés** submitted to the company, which were predominantly from **male applicants** in technical and leadership roles.

★ **Discrimination Learning:** The AI learned to associate words and phrases common on male-dominated résumés (e.g., "executed," "captured," or attendance at all-male colleges) with successful outcomes, while effectively **penalizing résumés** that contained words traditionally associated with female candidates (e.g., "woman's," or mention of women's chess clubs). The system essentially learned to mimic and amplify the **historical gender disparity** already present in the tech industry's hiring records.

2. ***Propose three fixes to make the tool fairer***

★ Ablate or neutralize gender-specific terms in the training data.

★ Implement a Fairness Regularization or Adversarial Debiasing technique during the model's training process.

★ Apply a Post-Processing Calibration technique to the model's output scores.

3. ***Suggest metrics to evaluate fairness post-correction***

The system must avoid Disparate Impact and uphold Equal Opportunity across demographic groups, such as gender.

The key metric for Disparate Impact is the Adverse Impact Ratio (AIR). AIR is the ratio of the selection rate for the disadvantaged group versus the advantaged group. The "4/5ths Rule" requires this ratio to be 0.8 or greater to indicate fairness.

Equal Opportunity uses the True Positive Rate (TPR): the proportion of qualified candidates (who would have been successful) that the model correctly selects. Fairness requires an Equal Opportunity Difference (EOD) near zero, ensuring all highly qualified individuals are identified equally.

Finally, Predictive Parity assesses the Positive Predictive Value (PPV)—the probability that a candidate predicted as qualified actually proves successful. The difference in PPV between groups must be near zero to confirm prediction reliability is consistent across all candidates, regardless of their gender.

Case 2: *Facial Recognition in Policing*

1. ***Discuss ethical risks (e.g., wrongful arrests, privacy violations)***

The primary ethical risks associated with facial recognition systems that misidentify minorities stem from the potential for wrongful arrests and systemic discrimination. When algorithms show significantly higher error rates for certain demographics—often due to unrepresentative training data—these errors are amplified in real-world use by law enforcement. A false positive, which incorrectly identifies an innocent person as a suspect, occurs disproportionately to minority individuals, leading directly to unjust detentions, stops, and arrests. This automated bias reinforces and scales existing societal inequalities,

imposing severe consequences on communities that are already over-policed. Furthermore, the constant, pervasive deployment of this flawed technology erodes privacy and civil liberties. The fear of being incorrectly identified or permanently tracked by a biased system creates a chilling effect on fundamental rights like freedom of assembly and speech, as individuals may avoid public protest or certain gatherings to evade surveillance and potential misidentification, thereby undermining democratic participation and autonomy.

2. ***Recommend policies for responsible deployment.***

★ Mandate Rigorous Bias Testing and Auditing.
★ Implement Strong Regulatory and Legal Safeguards.
★ Ensure Transparency and Public Oversight.

Part 3: *Practical Audit*

Part 4: *Ethical Reflection*

Bonus Task

Patient Consent Protocols

AI systems must respect patient autonomy through clear, comprehensive, and dynamic consent mechanisms.

1. Informed and Granular Consent

- **Clear Disclosure:** Patients must be explicitly informed when an AI system is involved in their diagnosis, treatment planning, or care management. This disclosure must use plain language, avoiding technical jargon.
- **Purpose Specificity:** Consent must specify the exact clinical purpose for which their data will be used by the AI (e.g., "for cancer risk prediction," or "for personalised medication dosing").
- **Dynamic Withdrawal:** Patients retain the right to withdraw their consent for the future use of their data in AI training or system updates at any time, without prejudice to their ongoing care.

2. Consent for Data Usage

- **Anonymisation Strategy:** Document and disclose the methods used to de-identify or anonymise patient data for AI development.
- **Secondary Use:** If patient data, aggregated or anonymised, is intended for secondary uses (e.g., commercial licensing, external research), a separate, explicit consent process is required.

Bias Mitigation Strategies

AI systems must be designed, trained, and validated to ensure equitable performance across all patient populations, including those defined by demographics, socioeconomic status, and cultural background.

1. Data Equity and Diversity

- **Representative Data Audits:** Conduct mandatory audits of training datasets to identify under-representation of specific demographic groups (age, gender, ethnicity, geography).
- **Bias Remediation:** Actively balance or augment training data to ensure adequate representation, and implement algorithmic techniques (e.g., adversarial debiasing) to correct identified data biases before deployment.

2. Performance Equity Monitoring

- **Disparity Analysis:** Validate and monitor AI performance (accuracy, false positive/negative rates) segmented by protected attributes *during development and post-deployment*.
- **Threshold Calibration:** Where performance disparities exist, recalibrate AI decision thresholds to optimise outcomes for the most vulnerable or historically underserved populations, rather than optimising solely for overall average accuracy.
- **Human Review:** Mandate structured human review (e.g., by clinical staff) of all high-risk AI outputs that affect populations where known performance gaps persist.

Transparency and Accountability Requirements

Documentation must be rigorous and accessible to allow for effective auditing, explainability, and clear allocation of clinical responsibility.

1. Documentation and Auditability

- **Model Card Creation:** For every deployed AI model, maintain a "Model Card" detailing its intended use, training data provenance, bias testing results, performance metrics (including segmented metrics), and known limitations.
- **Decision Logging:** AI systems must maintain an immutable log of all inputs, outputs, confidence scores, and time stamps associated with every clinical decision or recommendation. This log must be accessible for regulatory and clinical audits.

2. Explainability (XAI)

- **Clinician-Facing Explanations:** Provide clinically relevant explanations (e.g., saliency maps, feature importance scores) that clearly illustrate *why* the AI arrived at a specific conclusion. Explanations must be integrated into the clinical workflow.
- **Patient Explanation Framework:** Develop standardised communication frameworks to explain high-stakes AI recommendations to patients in an understandable and empathetic manner.

3. Accountability

- **Clear Ownership:** The ultimate clinical responsibility for patient care, including decisions informed by AI, resides with the licensed healthcare provider. The AI system acts as a sophisticated tool, not an autonomous decision-maker.

- **Reporting Mechanism:** Establish a formal process for reporting, investigating, and remediating AI errors or unintended consequences in patient care, mirroring existing medical incident protocols.