AI Ethics Assignment
Designing Responsible and Fair AI Systems

Part 1: Theoretical Questions

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

Algorithmic bias refers to systematic errors in an AI system that lead to unfair, discriminatory, or unequal outcomes for certain groups. These biases often arise from skewed training data, flawed model design, or historical inequalities reflected in datasets.

Examples:

1. Hiring Algorithms Discriminating Against Women
   A model trained on historical resumes (mostly from men) learns to downgrade resumes containing words like "women's chess club," penalizing female applicants.
2. Facial Recognition Misidentifying Darker-Skinned People
   Some systems show higher error rates for Black individuals because the training data overrepresents lighter-skinned faces.

Q2: Difference between transparency and explainability. Why are both important?

- Transparency means openness about how an AI system works like data sources, model design, decision processes, and limitations. It ensures visibility into the AI lifecycle.
- Explainability refers to the ability to interpret why an AI system made a specific decision. It focuses on producing human-understandable explanations.

They matter because:

- Transparency builds trust, allows audits, and supports accountability.
- Explainability helps users understand model behaviour, detect errors, contest decisions, and ensure fairness.
  These two promote ethical, safe, and reliable AI.

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

GDPR affects AI by imposing strict rules on how personal data is collected, stored, processed, and used. It requires:

- Explicit user consent before collecting data
- Right to explanation for automated decisions
- Data minimization by only collecting necessary data
- Right to access, delete, or correct data
- Penalties for misuse or unauthorized access

This encourages AI developers to design systems that are privacy-preserving, transparent, secure, and user-centric.

Ethical Principles Matching

Definition	**Principle**
Ensuring AI does not harm individuals or society.	B) Non-maleficence
Respecting users' right to control their data and decisions.	C) Autonomy
Designing AI to be environmentally friendly.	D) Sustainability
Fair distribution of AI benefits and risks.	A) Justice

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

1. Identify the source of bias

- The AI model was trained using 10 years of past hiring data, which reflected a male-dominated tech industry.
- The model learned to associate "male-related patterns" with success.
- Words like "women's sports" were penalized because they didn't appear in the resumes of previously hired candidates.

2. Three fixes to make the tool fairer

1. Balanced and representative training data
   Remove gendered terms and ensure resumes from diverse genders, backgrounds, and roles are included.
2. Feature filtering / debiasing
   Delete or neutralize features that encode gender (e.g., pronouns, gendered keywords, women's groups).
3. Fairness-aware algorithms
   Use fairness constraints such as:

    - Equal Opportunity
    - Demographic Parity
    - Disparate Impact Removal (AIF360)

3. Fairness metrics to evaluate after correction

- Disparate Impact Ratio which should be near 1
- Equal Opportunity Difference (true positive rates across gender groups)
- False Negative Rate difference
- Calibration differences between genders

Case 2: Facial Recognition in Policing

1. Ethical risks

- Wrongful arrests due to misidentification
- Racial profiling and reinforcement of discrimination
- Violation of privacy through mass surveillance
- Lack of consent for data collection
- Loss of trust in law enforcement
- Potential abuse of power

2. Policies for responsible deployment

1. Mandatory accuracy benchmarks
   Systems must meet high performance standards across all demographic groups.
2. Human-in-the-loop decision making
   Facial recognition should only be used as support, not as the final decision.
3. Independent third-party audits
   Regular bias testing to ensure equal error rates across races and genders.
4. Transparency and public oversight
   Public reporting on usage, errors, and complaints.
5. Strict data governance
   Limit data retention, ensure consent, and protect biometric data.

Audit Report

Bias Audit Report: COMPAS Recidivism Dataset

The COMPAS dataset contains criminal justice data used to predict the risk of reoffending. Previous investigations (e.g., by ProPublica) suggest that the algorithm exhibits racial disparities, particularly against African-American defendants. Using IBM's AI Fairness 360 toolkit, I conducted a fairness audit to evaluate error rates and prediction patterns across racial groups.

I trained a logistic regression model on the dataset and computed fairness metrics comparing two groups: Caucasian (privileged) and African-American (unprivileged). The results clearly show disparities. The false positive rate difference is significantly above zero, meaning African-Americans are more likely to be incorrectly labeled as “high risk” when they are not. Additionally, the equal opportunity difference is negative, indicating lower true positive rates for the unprivileged group. Lastly, the statistical parity difference shows that the unprivileged group receives high-risk labels more often.

Visualizations of the False Positive Rate by race show that Reweighing reduces the disparity, though it does not completely eliminate it.

These findings align with concerns raised by civil rights groups: the model amplifies historical biases present in the criminal justice system. The dataset itself contains structural inequalities, such as higher arrest rates in minority neighbourhoods due to policing patterns, not necessarily higher actual offenses.

To address these issues, I recommend implementing bias mitigation strategies such as Reweighing (pre-processing), Adversarial Debiasing (in-processing), or Reject Option Classification (post-processing). Additionally, the system should not be used as a standalone decision-making tool; instead, human review and transparent reporting should accompany all automated predictions.

Overall, the audit confirms that naive ML models built on COMPAS data can perpetuate racial harm, and fairness-aware algorithms are essential for more equitable outcomes.

Part 4: Ethical Reflection

In my future AI projects, I will prioritize ethical design by integrating transparency, fairness, and user rights from the beginning. First, I will adopt data minimization and collect only what is necessary. Second, I will apply fairness audits using tools like AI Fairness 360 to detect biases early. Third, I will ensure transparency by documenting model limitations and providing clear user explanations. Finally, I will maintain accountability by including human oversight and continuous monitoring. Ethical AI is not an add-on, it is a core requirement for building systems that are trustworthy and beneficial to society.

Ethical AI in Healthcare Policy

Ethical AI Guidelines for Healthcare Systems

1. Patient Consent Protocols

- Obtain explicit, informed consent before collecting patient data.
- Patients must have the right to withdraw consent at any time.

- Provide clear explanations of how AI processes their data.

2. Bias Mitigation Strategies

- Use diverse, representative datasets across gender, race, age, and socioeconomic backgrounds.
- Conduct regular fairness audits for diagnostic tools, treatment recommendations, and triage algorithms.
- Apply fairness-aware algorithms and remove sensitive attributes when appropriate.

3. Transparency Requirements

- Explain AI decisions in language patients can understand.
- Clearly document model limitations, accuracy rates, and applicable populations.
- Provide access logs showing when AI influenced clinical decisions.

4. Accountability & Safety

- Maintain human oversight, doctors must validate AI suggestions.
- Establish mechanisms to report incorrect predictions or ethical concerns.
- Regularly update models to reflect new medical evidence.

5. Data Security

- Enforce strong encryption, access control, and anonymization.
- Limit data sharing to authorized entities only.
- Ensure compliance with HIPAA, GDPR, and local health data laws.