AI Ethics Assignment

Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Algorithmic Bias
Algorithmic bias occurs when an AI system produces unfair outcomes due to flawed assumptions or skewed data. For example:

1. Hiring algorithms may favor male candidates if trained on historical data where men dominated tech roles, as happened with Amazon's recruitment tool.
2. Loan approval models might deny applications from minority neighbourhoods due to biased ZIP code correlations in training data.

Q2: Transparency vs. Explainability

- Transparency means openly sharing how an AI system is built (e.g., disclosing data sources and model architecture).
- Explainability focuses on making individual decisions understandable (e.g., showing which features led to a medical diagnosis).
  Both are vital: Transparency builds public trust (e.g., revealing when facial recognition is used), while explainability lets doctors challenge AI errors in patient care.

Q3: GDPR Impact on EU AI Development
GDPR imposes strict rules:

- Right to Explanation: Users can demand why an AI decision affected them (e.g., loan denial).
- Data Minimization: AI must use only essential data (e.g., limiting biometric collection).
- Fines: Violations cost up to 4% of global revenue, forcing ethical design from the start.

---

2. Ethical Principles Matching

Principle	Definition
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 (Amazon)

- Source of Bias: Training data from 10 years of male-dominated engineering hires, causing the model to penalize resumes with words like "women's chess club."
- Fixes:
  - Debias training data by oversampling female applicant profiles.
  - Remove gender proxies (e.g., pronouns, club names) using adversarial learning.
  - Human-AI collaboration: Flag uncertain cases for HR review.
- Fairness Metrics:
  - Disparate impact ratio (ensure selection rates for women/men differ by <20%).
  - False positive rate equality (equal interview chances for qualified candidates across genders).

Case 2: Facial Recognition in Policing

- Ethical Risks:
  1. Wrongful arrests: Misidentifying minorities (e.g., 35% higher error rates for darker skin tones).
  2. Mass surveillance: Tracking individuals without consent, eroding public trust.

- Deployment Policies:
    1. Legislative bans on real-time facial recognition in public spaces (as in EU's AI Act).
    2. Bias audits: Mandatory third-party testing for demographic fairness.
    3. Purpose limitation: Restrict use to serious crimes (e.g., terrorism), not petty offenses.

Part 4: Ethical Reflection

In my smart agriculture project, I'd ensure ethical AI through:

1. Justice: Partner with small farms in developing countries to avoid bias toward high-tech industrial data.
2. Non-maleficence: Validate soil sensors to prevent crop failure recommendations (e.g., testing in drought simulations).
3. Autonomy: Let farmers opt out of data sharing and override AI suggestions.
4. Sustainability: Use solar-powered edge devices to minimize carbon footprint.
   Example: If predicting crop yields, I'd balance training data across farm sizes and regions to prevent favoring wealthy landowners—addressing bias before deployment with tools like IBM AIF360.