
PART 1: Theoretical Understanding (30%)

Q1: What is Algorithmic Bias? Give Two Examples.

Answer:

Algorithmic bias occurs when an AI system produces unfair, prejudiced, or discriminatory outcomes due to flaws in the data, model, or decision-making process.

Examples:

1. **Hiring Systems:** An AI recruiting tool trained on biased past data may penalize women or minority applicants.

2. **Loan Approvals:** AI models may deny loans to certain races or ZIP codes due to historical biases in financial datasets.

Q2: Transparency vs. Explainability

Term	Definition
Transparency	Refers to the openness of an AI system—what data was used, how decisions are made.
Explainability	Describes how well a human can understand *why* the AI made a decision.

Importance:

- **Transparency** builds **trust** in AI systems.

- **Explainability** allows developers and users to **identify errors or biases**, improving accountability.

Q3: How GDPR Affects AI in the EU

- **Right to Explanation:** Users can ask *why* an algorithm made a decision.
- **Data Protection:** Limits what data can be used and requires consent.
- **Accountability:** Forces companies to ensure fairness and document their AI processes.

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 (40%)

Case 1: Amazon's Biased Hiring Tool

- **Bias Source:**
    - Historical data was biased—mainly male applicants.
    - Model learned to down-rank resumes with "women" or female-associated terms.
- **3 Fixes:**
    - **Debias training data** (e.g., balance male/female examples).
    - **Remove gender-indicating features** from input.
    - **Human-in-the-loop reviews** to monitor fairness.
- **Fairness Metrics:**

- **Disparate Impact Ratio**
- **Demographic Parity**
- **Equal Opportunity Difference**

Case 2: Facial Recognition in Policing

- **Ethical Risks:**
  - High **false positive rates** for minorities → wrongful arrests.
  - **Privacy violations** through mass surveillance.
  - **Lack of consent** from individuals being scanned.
- **Policy Suggestions:**
  - **Ban use in public spaces** without consent.
  - Require **bias audits** before deployment.
  - Use **transparent model reporting** and human oversight.

Summary Report (300 words)

The COMPAS dataset is widely used to assess risk of recidivism but has faced criticism for racial bias. Using the AI Fairness 360 toolkit, we audited the dataset and observed significant disparities in predicted outcomes between black and white individuals.

The **Disparate Impact Ratio** showed that black defendants were more likely to be labeled as high-risk than white defendants, despite similar criminal records. Additionally, visualizations of false positive rates confirmed that the algorithm often wrongly labeled black individuals as reoffenders.

To mitigate this, we recommend:

- Applying **Reweighing** or **Adversarial Debiasing** during training.

- Using **Equal Opportunity Difference** to evaluate fairness.
- Involving **human reviewers** to cross-check automated decisions.

Ultimately, the audit shows that fairness in AI must be continuously monitored, especially in high-stakes domains like criminal justice.

PART 4: Ethical Reflection (5%)

In my future AI project—a health chatbot for low-income users—I will ensure ethical AI by:

- Using **representative data** that includes different genders, ethnicities, and age groups.
- Ensuring **explainability** of diagnosis and suggestions.
- Getting **informed consent** before collecting any data.
- Being transparent about data usage and algorithm design.
  I believe ethical AI is not a feature—it's a foundation.

BONUS TASK (10%): Policy for Ethical AI in Healthcare

1-Page Policy Draft

Title: Ethical AI Use in Healthcare — PLP Academy Proposal

1. **Patient Consent:**
   - All AI systems must receive informed, voluntary consent before collecting data.
   - Patients must be able to opt-out at any time.
2. **Bias Mitigation:**
   - Datasets should be reviewed for diversity (age, race, gender).

- Apply fairness algorithms (e.g., reweighing, debiasing).
- Models must be retrained periodically to avoid drift.

3. **Transparency & Accountability:**
    - Clearly explain how the AI model works in simple language.
    - Include a human-in-the-loop for critical decisions.
    - Document the data sources and modeling process.