AI Ethics & Bias Mitigation: Comprehensive Analysis & Implementation

Author: Odugbile Olalekan Ayodele

Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Algorithmic Bias Definition and Examples

Algorithmic Bias refers to systematic and repeatable errors in computer systems that create unfair outcomes, such as privileging one arbitrary group of users over others. This occurs when AI systems produce discriminatory results due to prejudiced assumptions in the machine learning process or biased training data.

Examples of Manifestation:

Hiring Discrimination: Amazon's recruiting tool penalized female candidates because it was trained on resumes submitted over a 10-year period, which came predominantly from male applicants. The system learned to downgrade resumes containing words like "women's" or graduates from all-women's colleges.

Criminal Justice Disparities: The COMPAS algorithm used in US courts to predict recidivism was found to be nearly twice as likely to falsely flag black defendants as future criminals (45% false positive rate) compared to white defendants (23% false positive rate), while being more likely to falsely predict low risk for white defendants.

Q2: Transparency vs Explainability in AI

Transparency refers to the openness about AI system development, including data sources, model architecture, and training methodologies. It answers "What are you doing and how are you doing it?"

Explainability refers to the ability to understand and articulate how an AI system reaches specific decisions or predictions. It answers "Why did you make that particular decision?"

Importance of Both:

Transparency builds trust with stakeholders and enables regulatory compliance

Explainability enables error detection, bias identification, and user trust in individual decisions

Together, they facilitate accountability, allow for system improvement, and ensure that AI decisions can be challenged and understood by humans

Q3: GDPR Impact on AI Development in EU

The General Data Protection Regulation significantly constrains AI development through:

Right to Explanation: Article 22 gives individuals the right not to be subject to solely automated decisions, including profiling, that produce legal or similarly significant effects

Data Minimization: AI systems can only collect data strictly necessary for specific purposes

Purpose Limitation: Data collected for one purpose cannot be repurposed for AI training without additional consent

Right to Erasure: "Right to be forgotten" requires AI models to be retrainable without specific individual data

Data Protection by Design: AI systems must incorporate privacy protections from the initial design phase

Impact Assessments: Mandatory Data Protection Impact Assessments for high-risk AI applications

2. Ethical Principles Matching

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

A) Justice → Fair distribution of AI benefits and risks

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool Analysis

Source of Bias:
The primary source of bias was historical training data bias. Amazon's AI was trained on resumes submitted to the company over a 10-year period, which predominantly came from male applicants, reflecting the male-dominated tech industry. The system learned to associate masculine patterns with successful candidates and penalize feminine indicators.

Three Proposed Fixes:

Debiased Training Data:

Collect balanced training data representing all demographic groups

Use synthetic data generation to create balanced datasets

Implement adversarial debiasing during model training

Feature Engineering & Removal:

Remove proxy variables for protected attributes (names, gender-indicating words, specific colleges)

Use only job-relevant qualifications and skills as features

Implement fairness-aware feature selection

Post-processing Fairness:

Apply different decision thresholds for different demographic groups

Use statistical parity constraints to ensure equal selection rates

Implement regular fairness audits with demographic parity testing

Fairness Evaluation Metrics:

Demographic Parity: Equal selection rates across gender groups

Equalized Odds: Similar false positive and false negative rates across groups

Predictive Parity: Equal precision across demographic groups

Adverse Impact Ratio: Selection rate ratios between protected and non-protected groups (should be >0.8)

Case 2: Facial Recognition Ethical Analysis

Ethical Risks:

Wrongful Arrests & Convictions: Higher false positive rates for minorities could lead to innocent people being detained, arrested, or convicted based on erroneous identifications

Privacy Violations: Mass surveillance capabilities threaten individual privacy and freedom of assembly without appropriate safeguards

Reinforcement of Systemic Bias: Automated systems could perpetuate and amplify existing racial disparities in law enforcement

Due Process Erosion: Over-reliance on "black box" AI systems without proper human oversight and validation

Responsible Deployment Policies:

Pre-deployment Validation:

Mandatory independent third-party bias auditing across demographic groups

Minimum accuracy thresholds (e.g., <1% false positive rate across all demographics)

Transparency requirements for model performance metrics

Operational Safeguards:

Prohibition of sole reliance on facial recognition for arrests or warrants

Requirement for human verification with clear standards

Use limited to serious crimes with judicial oversight

Ongoing Monitoring:

Regular performance audits with demographic breakdowns

Public reporting of system usage and error rates

Independent oversight committee with diverse representation

Legal & Regulatory Framework:

Clear liability frameworks for erroneous identifications

Data protection compliance and usage limitations

Community consultation requirements before deployment