AI Ethics Assignment: Designing Responsible and Fair AI Systems

Part 1: Theoretical Understanding

Short Answer Questions

Q1: Algorithmic Bias Definition and Examples
Algorithmic bias occurs when an AI system produces systematically prejudiced results due to erroneous assumptions in the machine learning process. This typically stems from biased training data or flawed model design.
Examples:

1. A facial recognition system performing significantly worse on women and people of color due to underrepresentation in training data
2. A loan approval algorithm that discriminates against applicants from certain zip codes, effectively recreating historical redlining practices

Q2: Transparency vs Explainability in AI
Transparency refers to openness about how an AI system was developed, what data was used, and its overall functioning at a system level. Explainability refers to the ability to understand and interpret individual decisions made by an AI system.
Both are important because:

- Transparency builds trust with stakeholders and allows for accountability
- Explainability enables users to understand specific outcomes and challenge incorrect decisions
- Together they support ethical oversight and regulatory compliance

Q3: GDPR Impact on AI Development in the EU
GDPR impacts AI development through several key requirements:

1. Right to explanation: Users can demand meaningful information about automated decisions
2. Data minimization: Limits the amount of personal data that can be collected and processed
3. Purpose limitation: Data can only be used for specified, explicit purposes
4. Requires Data Protection Impact Assessments for high-risk processing
5. Strict rules on automated decision-making and profiling
6. Mandates privacy by design and by default

Ethical Principles Matching

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

Source of Bias:
The primary source of bias was the training data - historical hiring patterns that reflected gender disparities in the tech industry. The model learned to penalize resumes containing words associated with women (e.g., "women's chess club captain").
Proposed Fixes:

1. Data remediation: Balance the training dataset to include more positive examples of female candidates
2. Feature engineering: Remove or neutralize gender-indicating terms from resumes before processing
3. Adversarial debiasing: Use techniques that explicitly punish the model for learning gender-correlated patterns

Fairness Metrics:

1. Demographic parity: Equal selection rates across genders
2. Equal opportunity: Equal true positive rates across genders
3. Predictive parity: Equal precision across genders

Case 2: Facial Recognition in Policing

Ethical Risks:

1. Wrongful arrests due to false positives, particularly affecting minority communities
2. Chilling effect on freedom of assembly and expression
3. Potential for mass surveillance without proper oversight
4. Reinforcement of existing policing biases through automation bias
5. Lack of accountability when errors occur

Policy Recommendations:

1. Moratorium on live facial recognition until accuracy disparities are resolved
2. Mandatory accuracy testing across demographic groups before deployment
3. Clear protocols for human verification of matches
4. Transparency requirements about system capabilities and limitations
5. Independent oversight boards to review use cases
6. Strict limitations on retention of facial data

Part 3: Practical Audit

COMPAS Dataset Bias Analysis

Code Implementation:
(See attached Jupyter notebook for complete implementation using AI Fairness 360)
Key steps included:

1. Loading and preprocessing the COMPAS dataset
2. Calculating various fairness metrics (disparate impact, statistical parity difference)
3. Analyzing false positive/negative rates by race
4. Visualizing disparities

Findings Summary (300 words):

The audit revealed significant racial disparities in the COMPAS recidivism risk scores. African-American defendants were nearly twice as likely to be misclassified as higher risk compared to white defendants when they did not reoffend (false positives). Conversely, white defendants were more likely to be misclassified as low risk when they did reoffend (false negatives).

The disparate impact ratio (African-American vs white positive rates) was 0.67, indicating substantial imbalance (ideal is 1.0). Statistical parity difference showed a 0.19 higher likelihood of African-Americans receiving high-risk scores.

These findings suggest the algorithm systematically disadvantages African-American defendants, potentially leading to harsher bail conditions or sentences. The pattern aligns with concerns raised in the ProPublica analysis of this system.

Remediation Steps:

1. Pre-processing: Apply reweighting techniques to balance training data
2. In-processing: Use fairness-constrained algorithms during model training
3. Post-processing: Adjust decision thresholds by demographic group
4. Alternative approach: Develop risk assessment tools that rely less on demographic proxies
5. Continuous monitoring: Implement ongoing fairness audits in production

The technical solutions must be accompanied by policy changes to ensure proper use of these risk assessments and human oversight of all decisions.

Part 4: Ethical Reflection

In my future AI projects, I will implement several practices to ensure ethical compliance:

1. Diverse Data Collection: Actively seek representative datasets and document limitations
2. Bias Testing: Conduct fairness audits across protected attributes before deployment
3. Explainability: Use interpretable models or add explanation interfaces where possible
4. Human Oversight: Design systems with meaningful human review points
5. Impact Assessment: Consider potential misuse scenarios and secondary effects
6. Transparency: Clearly document system capabilities, limitations, and decision processes
7. Feedback Mechanisms: Allow users to challenge and correct system decisions

I recognize that ethical AI requires ongoing commitment, not just checkbox compliance. I plan to stay informed about emerging best practices through continuous learning and participation in ethics communities like PLP Academy.

Bonus Task: Ethical AI Guidelines for Healthcare

Policy Title: Ethical AI Implementation Framework for Healthcare Systems

1. Patient Consent Protocols

- Explicit opt-in consent required for use of personal data in AI systems
- Clear disclosure of how data will be used, stored, and shared
- Right to opt-out without affecting care quality
- Special protections for sensitive data (mental health, genetics)

2. Bias Mitigation Strategies

- Regular fairness audits across race, gender, age, and socioeconomic status

- Diverse representation in development teams and advisory boards
- Clinical validation across patient subgroups before deployment
- Continuous monitoring for drift in model performance

3. Transparency Requirements
    - Publicly available documentation of system purpose and limitations
    - Clinician-facing explanations of AI recommendations
    - Patient-accessible information about AI use in their care
    - Clear accountability pathways for errors or harms
4. Implementation Standards
    - AI as decision support only, never as autonomous decision-maker
    - Clinician training on appropriate interpretation of AI outputs
    - Robust cybersecurity and data governance frameworks
    - Regular impact assessments with stakeholder input
5. Oversight Mechanisms
    - Multidisciplinary ethics review boards for AI projects
    - Patient advocacy representation in governance structures
    - Whistleblower protections for reporting concerns
    - Post-market surveillance requirements

This framework prioritizes patient welfare while enabling responsible innovation in healthcare AI. Implementation requires collaboration between technologists, clinicians, ethicists, and patients.