AI ETHICS ASSIGNMENT

Assignment: Designing Responsible and Fair AI Systems

Author: Peter Kamau Mwaura

1 PART 1: THEORETICAL UNDERSTANDING (30%)

1.1 Short Answer Questions

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

Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others. This bias can emerge from various sources in the AI development pipeline, including biased training data, flawed model design, or unintended use cases.

Two examples of manifestation:

Hiring Discrimination: Amazon's recruiting tool was trained predominantly on resumes from male applicants, causing it to systematically downgrade resumes containing words like "women's" (as in "women's chess club") or graduates from women's colleges, thereby disadvantaging female candidates.

Criminal Justice Disparities: The COMPAS algorithm used for predicting recidivism risk was found to be disproportionately likely to falsely flag Black defendants as high-risk while incorrectly labelling White defendants as low-risk, perpetuating existing racial disparities in the justice system.

Q2: Explain the difference between *transparency* and *explainability* in AI. Why are both important?

Transparency refers to openness about the AI system's overall design, data sources, algorithms, and processes. It answers the question: "What is this system and how does it generally work?" This includes documenting the training data, model architecture, and potential limitations.

Explainability refers to the ability to understand and articulate the reasoning behind a specific decision or prediction made by an AI system. It answers the question: "Why did the system make this particular decision for this specific case?"

Why both are important:

Transparency builds institutional trust and enables regulatory compliance, allowing stakeholders to assess the system's overall fairness and reliability.

Explainability builds individual trust and enables accountability, allowing affected individuals to understand decisions that impact them and providing developers with tools to debug and improve the system.

Together, they create a comprehensive framework for trustworthiness - transparency at the system level, and explainability at the decision level.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

GDPR significantly constrains and guides AI development in several key ways:

1. **Right to Explanation (Article 22)**: Individuals have the right not to be subject to solely automated decisions that significantly affect them, and to obtain meaningful explanations of such decisions.
2. **Data Minimization Principle**: AI systems can only collect and process data that is strictly necessary for their specified purpose, limiting the scope of training data.
3. **Purpose Limitation**: Data collected for one purpose cannot be repurposed for AI training without additional consent.
4. **Right to Erasure**: Individuals can request deletion of their personal data, creating challenges for models already trained on that data.

5. **Privacy by Design and by Default**: Requires data protection to be integrated into AI systems from the initial design phase rather than being added as an afterthought.
6. **Data Subject Rights**: Includes rights to access, rectify, and port personal data, ensuring individuals maintain control over how their information is used in AI systems.

1.2 Ethical Principles Matching

1. Ensuring AI does not harm individuals or society. → B) Non-maleficence

2. Respecting users' right to control their data and decisions. → C) Autonomy

3. Designing AI to be environmentally friendly. → D) Sustainability

4. Fair distribution of AI benefits and risks. → A) Justice

2 PART 2: CASE STUDY ANALYSIS (40%)

2.1 Case 1: Biased Hiring Tool

2.1.1 Identify the source of bias

The primary source of bias was in the **training data**. The model was trained on resumes submitted to Amazon over a 10-year period, which predominantly came from male applicants. The AI learned patterns associated with successful candidates from this historically male-dominated dataset, causing it to penalize features associated with female candidates, such as attendance at women's colleges or participation in women-focused organizations.

2.1.2 Propose three fixes to make the tool fairer

Data Remediation: Actively source and incorporate resumes from a gender-balanced pool of candidates. Apply techniques like synthetic data generation for underrepresented groups and remove gender-identifying information from the training data.

Algorithmic De-biasing: Implement fairness constraints during model training, such as adversarial de-biasing that removes protected attributes (gender) from the learned representations, or use pre-processing techniques like reweighing the training data to balance representation.

Human-in-the-Loop Validation: Ensure human reviewers validate the AI's recommendations, with particular attention to gender balance in the shortlisted candidates, and establish a feedback mechanism to continuously improve the system.

2.1.3 Suggest metrics to evaluate fairness post-correction

Demographic Parity: Measure whether the selection rate is approximately equal across gender groups (ideally close to 1:1 ratio).

Equalized Odds: Ensure that true positive rates (qualified candidates correctly recommended) and false positive rates are similar for both male and female applicants.

Predictive Parity: Verify that the precision (percentage of truly qualified candidates among those recommended) is similar across groups.

2.2 Case 2: Facial Recognition in Policing

2.2.1 Discuss ethical risks

- **Wrongful Arrests and Convictions**: False positives can lead to innocent people being detained, arrested, or convicted based on erroneous AI identification, causing severe psychological trauma, financial loss, and permanent damage to reputation.
- **Amplification of Systemic Bias**: These systems risk automating and scaling existing racial biases in policing, potentially leading to increased surveillance and over-policing of minority neighbourhoods.
- **Privacy Erosion and Chilling Effects**: Widespread facial recognition creates a surveillance infrastructure that could deter lawful political protest, free assembly, and normal public activities.
- **Due Process Violations**: The "black box" nature of many AI systems makes it difficult for defendants to challenge evidence or understand the basis for their identification.

2.2.2 Recommend policies for responsible deployment

- **Pre-deployment Bias Audits**: Mandate independent third-party testing to measure and report performance disparities across demographic groups before any deployment.

- **Use Case Restrictions**: Prohibit use for real-time mass surveillance and restrict applications to generating investigative leads rather than serving as primary evidence for arrests.
- **Transparency and Accountability**: Require public reporting of accuracy statistics, usage policies, and incident reports. Establish clear lines of responsibility for errors.
- **Judicial Oversight**: Require warrants or similar judicial authorization for most uses, similar to requirements for other invasive surveillance techniques.
- **Regular Impact Assessments**: Conduct ongoing monitoring and periodic re-evaluation of the technology's effects on civil rights and liberties.

3 PART 4: ETHICAL REFLECTION (5%)

Prompt: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?

For a future project developing a personalized learning recommendation system for students, I would implement the following ethical framework:

1. **Justice and Fairness**: I would begin by conducting a comprehensive bias audit of the training data to ensure equal representation across different student demographics (socioeconomic background, gender, geographic location). I would implement regular fairness testing using metrics like equal opportunity and demographic parity to detect any performance disparities in course recommendations for different student groups.
2. **Transparency and Explainability:** The system would include an "Explain This Recommendation" feature that clearly articulates why specific learning materials are suggested (e.g., "Recommended because you excelled in mathematics and showed interest in programming basics"). I would maintain detailed documentation of the system's capabilities, limitations, and data sources for all stakeholders.
3. **Privacy and Autonomy**: I would implement strict data minimization, collecting only essential information needed for educational personalization. Students would have clear opt-out controls and the ability to view, correct, or delete their data. The system would be designed with privacy-preserving techniques like federated learning where possible.

4. **Non-maleficence**: I would establish guardrails to prevent filter bubbles and ensure the system exposes students to diverse perspectives rather than reinforcing existing biases. Regular ethical reviews would assess potential unintended consequences, such as inadvertently steering certain demographics away from challenging subjects.
5. **Human-in-the-Loop**: While the system would provide recommendations, final educational decisions would remain with human educators and students themselves, positioning AI as an augmentative tool rather than a replacement for professional judgment.

By embedding these principles throughout the development lifecycle—from initial design to deployment and monitoring—I aim to create an AI system that not only enhances learning outcomes but does so in a manner that is fair, accountable, and respectful of student rights and diversity.

---

4 POLICY GUIDELINE: ETHICAL AI IMPLEMENTATION IN HEALTHCARE

Effective Date: [Insert Date] **Scope:** All clinical AI tools, predictive algorithms, and diagnostic support systems.

4.1 Purpose

To establish a mandatory framework for the deployment of Artificial Intelligence (AI) within [Organization Name], ensuring that patient safety, autonomy, and equity remain the core priorities of clinical innovation.

4.2 Patient Consent Protocols

Informed consent is not merely a signature; it is a continuous process of understanding.

- **Plain Language Disclosure:** Patients must be explicitly informed when an AI tool is playing a significant role in their diagnosis or treatment plan. This disclosure must be provided in non-technical language, avoiding jargon.
- **Scope of Usage:** Consent forms must clearly distinguish between AI used for direct clinical care and AI used for research/secondary data analysis.
- **Opt-Out Mechanisms:** Unless the AI is integral to a standard-of-care device (e.g., MRI image reconstruction), patients shall have the right to request a human-only review or opt-out of data sharing for model training without compromising their quality of care.
- **Data Sovereignty:** Patients retain the right to request the deletion of their specific data points from active learning databases where technically feasible.

4.3 Bias Mitigation Strategies

Algorithmic fairness is a clinical safety requirement.

- **Representative Training Data:** Before deployment, all AI models must be validated against a dataset that reflects the demographic diversity of our specific patient population (race, gender, age, socioeconomic status).
- **Protected Class Audits:** Algorithms must undergo quarterly "stress tests" to identify disparate performance metrics across protected classes. Any model showing a variance in error rates >5% between groups must be suspended for recalibration.
- **Human-in-the-Loop (HITL):** AI recommendations regarding critical care (e.g., triage, organ allocation, medication dosage) must never be automated fully. A qualified clinician must review and approve the AI's suggestion, serving as the final decision-maker to catch context-blind errors.

4.4 Transparency and Explainability Requirements

We do not deploy "Black Boxes" in critical care settings.

- **Interpretability:** Clinical staff must have access to "explainability layers" (e.g., heatmaps, decision trees) that indicate *why* an AI model reached a specific conclusion. If the reasoning is opaque, the tool cannot be used as a primary diagnostic aid.
- **Confidence Scoring:** All AI outputs displayed to clinicians must include a confidence interval or uncertainty score.
- **Labelling and Notification:** Electronic Health Records (EHR) must clearly flag entries generated or auto-populated by AI.
- **Accountability Trail:** A tamper-evident log must be maintained for every AI-driven decision, recording the input data, the model version used, the AI output, and the final human decision.

Review Cycle: This policy shall be reviewed every 6 months to adapt to rapid technological advancements.