🌍⚖️ AI Ethics Assignment Guide

Theme: Designing Responsible and Fair AI Systems

🧠 Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: What is Algorithmic Bias?

Algorithmic bias occurs when an AI system produces systematically unfair outcomes due to flawed data, design, or assumptions. This bias often stems from flawed data or the way the algorithm is designed.

- For example loan approval bias - a credit scoring model may deny loans to certain ethnic groups due to biased historical data.
- Another example includes hiring bias - an AI trained on male-dominated resumes may penalize female applicants.

Q2: Transparency vs. Explainability

Transparency refers to openness about how an AI system is built, trained, and deployed.

Explainability is the ability to understand and interpret how an AI system arrives at its decisions. Importance of both:

- Transparency builds trust and accountability.
- While explainability enables users and regulators to challenge or validate decisions.

Q3: GDPR's Impact on AI

GDPR has a significant impact on AI development by placing a high value on individual rights and data privacy. Key impacts include:

- Requires data minimization and purpose limitation.
- Enforces user consent and right to explanation. Article 22 gives individuals the right not to be subject to decisions based solely on automated processing.
- Limits automated decision-making that significantly affects individuals.

2. Ethical Principles Matching

- **Justice:** Fair distribution of AI benefits and risks.
- **Non-maleficence:** Ensuring AI does not harm individuals or society.

- **Autonomy:** Respecting users' right to control their data and decisions.
- **Sustainability:** Designing AI to be environmentally friendly.

📚 Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

- **Source of Bias:**

  The primary source of bias was **data bias**, specifically **historical data bias**. The tool was trained on a decade of hiring data from Amazon's tech department, which was largely male-dominated. The model learned to associate certain words and actions more common among men with success and penalized resumes that contained gender-specific keywords like "women's" or were from all-women's colleges.

- **Three Fixes:**

1. **Data Re-balancing and Augmentation:** Actively collect or synthesize a more representative dataset that includes a balanced number of successful resumes from all genders. This can involve oversampling underrepresented groups or creating synthetic data.
2. **Feature Engineering:** Systematically remove or de-prioritize features that act as proxies for gender. This includes removing the names of all-women's colleges, or flagging and removing any gendered keywords.
3. **Human-in-the-Loop Oversight:** Instead of fully automating the process, the AI should be used as a pre-screening tool that presents a diverse pool of qualified candidates to a human recruiter, who makes the final decision.

- **Fairness Metrics Post-Correction**:

1. **Disparate Impact Ratio:** This metric compares the selection rate of the unprivileged group (e.g., female candidates) to that of the privileged group (e.g., male candidates). A fair system would have a ratio close to 1.
2. **Equal Opportunity Difference:** This metric measures the difference in true positive rates (e.g., the proportion of actual qualified candidates who were correctly selected) between the unprivileged and privileged groups. A value of 0 indicates equal opportunity.

Case 2: Facial Recognition in Policing

- **Ethical Risks:**
  - **Wrongful Arrests:** Due to lower accuracy rates for minorities, these systems can generate **false positives**, leading to innocent people being misidentified and arrested. This can have devastating real-world consequences.
  - **Privacy Violations:** The widespread use of facial recognition enables mass surveillance, eroding the right to privacy and creating a chilling effect on free speech

  - and assembly.
  - **Reinforcement of Systemic Bias:** If a system is used predominantly in certain neighborhoods, it can create a feedback loop where more arrests are made in those areas, leading to more data for the system to learn from, thus reinforcing the biased perception of crime distribution.

- **Policy Recommendations for Responsible Deployment:**
  - **Mandatory Fairness Audits:** Before any system is deployed, an independent, public audit must be conducted to prove that the system performs with high and equitable accuracy across all racial and gender groups.
  - **Human-in-the-Loop Requirement:** The AI's output should never be the sole basis for an arrest or any significant legal action. A human officer must always review and independently verify the evidence.
  - **Limited and Specific Use Cases:** The technology's use should be limited to specific, high-stakes scenarios (e.g., matching a suspect to a photo from a crime scene), and its use in real-time, untargeted surveillance should be prohibited.

🧪 Part 3: Practical Audit (25%)

COMPAS Dataset Bias Audit Report Summary:

Racial Disparities in Risk Scores

The audit highlights a notable disparity in how the COMPAS algorithm predicts recidivism for different racial groups. For both general and violent recidivism, **African-American defendants are assigned high-risk scores at a higher false positive rate than Caucasian defendants**. This means that when an individual does not re-offend, the model is more likely to incorrectly predict a high risk for an African-American person than for a Caucasian person. The distribution of decile scores further reinforces this finding, showing that African-American individuals are disproportionately assigned higher risk scores across the board, regardless of the outcome.

Implications for the Justice System

These disparities have serious implications for the justice system. Because COMPAS scores can influence decisions regarding bail, sentencing, and parole, a biased model can lead to unequal outcomes. A higher false positive rate for African-American individuals means they are more likely to face harsher penalties, longer periods of pre-trial detention, and greater surveillance based on a flawed risk assessment. This unequal application of the score can erode public trust and perpetuate systemic inequities within the legal system.

Bias Mitigation Recommendations

To address the identified biases, several technical mitigation strategies can be employed. **Reweighting**, which involves assigning different weights to data points to achieve a more balanced representation, can help reduce the disparity. Alternatively, **adversarial debiasing** can be used to train a model that not only predicts recidivism but also simultaneously tries to hide any sensitive attribute information (like race), thereby forcing the model to make predictions without relying on discriminatory patterns. Implementing these methods is crucial to improving the fairness and ethical use of algorithmic risk assessments in the justice system.

💭 Part 4: Ethical Reflection

In my future project on predictive healthcare analytics, I will ensure ethical compliance by:

- Using diverse and representative datasets.
- Implementing fairness-aware algorithms.
- Ensuring transparency through model documentation.
- Obtaining informed consent from users.
- Regularly auditing for bias and unintended consequences

🏥A Guideline for Ethical AI Use in Healthcare

Purpose: This document outlines a framework for the responsible development and deployment of Artificial Intelligence (AI) systems within healthcare settings. Our goal is to ensure that AI enhances patient care while upholding core ethical principles, protecting patient autonomy, and mitigating potential harms.

1. Patient Consent Protocols

- **Informed and Explicit Consent:** All patients must provide explicit, written consent for their data to be used in AI models. This consent must be fully informed, detailing:
  - The specific purpose for which their data will be used.
  - The types of AI systems the data will train.
  - The potential risks and benefits to the patient.
  - How their data will be anonymized and protected.
- **Opt-Out Mechanism:** Patients must have an easy-to-use and transparent mechanism to revoke their consent at any time, with clear instructions on how their data will be removed from future AI training.
- **Transparency on AI-Assisted Decisions:** Patients must be informed if an AI system was used to assist in their diagnosis or treatment plan. The human clinician's role as the final decision-maker must be clearly communicated.

2. Bias Mitigation Strategies

- **Pre-Deployment Bias Audits:** All AI models must undergo a rigorous, independent audit before deployment to assess for bias. This audit will use standardized fairness metrics (e.g., Equal Opportunity Difference) to ensure the model performs with comparable accuracy and safety across all demographic groups, including race, gender, age, and socioeconomic status.
- **Diverse and Representative Datasets:** Data used for training AI models must be representative of the patient population the system will serve. Data scientists must actively audit and balance datasets to prevent the over- or under-representation of specific groups.
- **Regular Monitoring:** Post-deployment, AI systems must be continuously monitored for performance drift and emerging biases. Any significant disparities in outcomes must trigger an immediate review and re-calibration of the model.

3. Transparency Requirements

- **Documenting AI Lifecycle:** The entire AI system lifecycle, from data collection to deployment, must be thoroughly documented. This includes:
  - Detailed descriptions of the training data sources.
  - Model architecture and key parameters.
  - Fairness and performance metrics used during development.
  - A "model card" explaining the model's intended use, known limitations, and potential risks.
- **Explainable AI (XAI) Mandate:** For high-stakes applications like diagnostics, AI models must be designed to be explainable. The system should be able to provide a human-understandable rationale for its recommendations, allowing clinicians to verify the logic and make an informed final judgment.
- **Accountability:** Clear lines of responsibility must be established. The hospital, the developer, and the clinician must all be accountable for the outcomes of an AI-assisted decision.