ERICK WAMBUGU
POWER LEARN PROJECT AFRICA
AI FOR SOFTWARE ENGINEERING
WEEK 7 – ETHICAL AI

Part 2: Case Study Analysis (40%)

Case 1: Biased Hiring Tool — Amazon's Recruiting System

- **Scenario:** Amazon developed a hiring algorithm that unintentionally penalized female candidates because the training data came from 10 years of male-dominated tech hiring.

1. Identify the Source of Bias

- Biased Training Data
  - The historical CVs used for training were mostly from men.
  - The model learned that male-associated patterns (e.g., "executed", "captain", "software engineer") were "better", and downgraded CVs containing "women's…" activities or women-centric schools.

- Proxy Variables for Gender
  - Even when gender labels were removed, the model detected indirect signals (e.g., women-only colleges, female-coded words).

- Lack of Fairness Constraints in Model Design
  - The model was optimized purely for prediction accuracy — not fairness. No mechanisms prevented discriminatory outcomes.

2. Propose Three Fixes to Make the Tool Fairer

- Balanced, Representative Training Data
  - Include equal numbers of male and female Cvs.
  - Remove historical patterns that reflect discrimination.
  - Curate a gender-neutral dataset with diverse educational and professional backgrounds.

- Use Fairness-Aware Algorithms from Libraries like AIF360
  - Apply reweighing, disparate impact remover, or adversarial debiasing.
  - Add fairness constraints so the model cannot penalize gender-linked features.

- Human-in-the-Loop Decision-Making
  - AI should rank candidates, not judge them.
  - Final decisions should involve trained HR officers to catch errors or unfair patterns.

3. Suggest Metrics to Evaluate Fairness Post-Correction

- Disparate Impact Ratio (DIR)
  - Measures whether selection rates for women vs. men are proportionate.
  - (80% rule: ratio ≥ 0.8 is acceptable.)

- Equal Opportunity Difference
  - Ensures qualified male and female candidates have equal chances of being recommended.

- Demographic Parity
    - Checks whether the system recommends protected groups at similar rates.

Case 2: Facial Recognition in Policing

Scenario: Facial recognition misidentifies minorities at significantly higher rates, leading to wrongful arrests.

1. Ethical Risks

- Wrongful Arrests and Misidentification
    - Minority groups face higher risk of being falsely matched, leading to:
        - false criminal records
        - loss of freedom
        - psychological harm
        - erosion of community trust in law enforcement

- Privacy Violations
    - Mass surveillance may track individuals without consent, violating privacy rights and civil liberties.

- Discrimination and Inequality
    - Biased systems reinforce systemic racism and unequal treatment in policing.

- Lack of Transparency
    - Victims cannot understand or challenge AI-based decisions because algorithms are opaque.

- Overreliance on Technology
    - Police may assume the system is always correct, reducing human judgement.

2. Recommend Policies for Responsible Deployment

- Mandatory Accuracy & Bias Audits
    - Test models across racial, gender, and age groups.
    - Require minimum accuracy thresholds for each subgroup before deployment.

- Strict Human Oversight
    - AI should be used only as a supporting tool.
    - A positive match must be verified by trained human analysts.

- Transparency & Documentation Requirements
    - Agencies must disclose:
        - datasets used
        - model accuracy

        - limitations and risks
        - results of ongoing audits

- Consent, Privacy & Legal Safeguards
    - ✔ Enforce GDPR-style protections where citizens can contest automated decisions.
    - ✔ Ban real-time public surveillance unless legally justified.
    - ✔ Require warrants for facial recognition use.

- Community Engagement & Public Accountability
    - ✔ Involve civil rights groups in policy creation.
    - ✔ Publish annual fairness and impact reports.

- Use Only Fairness-Certified Vendors
    - ✔ Ensure systems meet standards such as:
        - Equal Opportunity
        - Demographic Parity
        - False Positive Rate Parity