Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

- Scenario Overview: Amazon developed an AI recruiting tool designed to automate candidate screening for technical roles. However, the tool was found to systematically disadvantage female candidates by scoring resumes with words associated with women (e.g., "women's chess club captain") lower.
- Source of Bias:
    - Historical Data Bias: The tool was trained on resumes submitted over a decade, during which male candidates dominated the tech industry. The AI learned these historical patterns and equated male-associated experiences with higher competence.
    - Feature Selection Bias: The system learned to favor certain words and patterns more common in male resumes.
    - Feedback Loop: Without interventions, continued use of such a model would reinforce gender imbalances in hiring.
- Three Fixes to Make the Tool Fairer:

    1. Curate a Gender-Balanced Dataset: Train the model on an equal number of male and female resumes to eliminate representational bias.
    2. Implement Bias Detection Algorithms: Use bias detection tools like IBM's AI Fairness 360 to detect gender bias in predictions before deployment.
    3. Feature Neutralization: Remove or reduce the weight of gender-related proxies in the features used for prediction.
- Metrics to Evaluate Fairness Post-Correction:
    - Disparate Impact Ratio: Ensures selection rates across genders are proportionate.
    - Equal Opportunity Difference: Measures the difference in true positive rates between male and female candidates.
    - Calibration by Group: Checks whether predicted probabilities of success are accurate for both genders.

Case 2: Facial Recognition in Policing

- Scenario Overview: Several studies, including those by MIT Media Lab and the ACLU, revealed that facial recognition systems, particularly those used by law enforcement, misidentify individuals of darker skin tones at a significantly

higher rate than lighter-skinned individuals. This has led to instances of wrongful arrests and detentions.

- Ethical Risks:
    1. Wrongful Arrests and Misidentification: The increased false positive rates for minorities can lead to innocent people being arrested, facing legal and social consequences.
    2. Violation of Privacy: Widespread surveillance using facial recognition can infringe on privacy rights, particularly when done without informed consent.
    3. Systemic Discrimination: If not properly regulated, these systems can reinforce existing racial biases in the criminal justice system.
    4. Loss of Public Trust: Misuse can lead to erosion of trust in law enforcement and public institutions.
- Policies for Responsible Deployment:

    1. Bias Auditing and Transparency: Require independent, third-party audits for bias in facial recognition algorithms before and during deployment.
    2. Regulatory Oversight: Enact legislation that limits the use of facial recognition technology in sensitive areas like policing, with strict judicial review.
    3. Data Diversity Requirements: Ensure that training datasets are diverse and represent all ethnicities, genders, and age groups adequately.
    4. Human-in-the-Loop Systems: Mandate that any facial recognition match be verified by human experts before action is taken.
    5. Public Disclosure: Agencies should publicly report on the accuracy and demographic performance of their facial recognition systems.
    6. Right to Redress: Implement mechanisms for individuals to contest and seek redress for wrongful identification.
