Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

Scenario: Amazon’s AI recruiting tool penalized female candidates.

1. Identify the Source of Bias

- **Training Data Bias:** The model was trained on historical hiring data predominantly consisting of male candidates, reflecting past gender imbalances. This caused the AI to learn biased patterns, favoring male resumes.

- **Feature Selection Bias:** Certain features (e.g., keywords associated with male-dominated fields or male pronouns) may have disproportionately influenced the model's decisions.

- **Lack of Diversity in Model Design:** The algorithm may not have included fairness constraints or human oversight to detect and correct gender bias.

2. Propose Three Fixes to Make the Tool Fairer

1. **Balanced and Representative Training Data:** Collect and use a more balanced dataset that includes equal representation of female and male candidates or synthetically augment data for underrepresented groups.

2. **Bias Mitigation Techniques:** Implement algorithmic fairness methods such as:

    - Reweighting training samples

    - Using adversarial debiasing

    - Adding fairness constraints to the optimization objective

3. **Human-in-the-Loop Review:** Incorporate human oversight at key decision points to audit AI recommendations, especially for protected groups, and to override biased decisions.

3. Suggest Metrics to Evaluate Fairness Post-Correction

- **Demographic Parity (Statistical Parity):** Check if the selection rate is similar across genders.
- **Equal Opportunity:** Ensure true positive rates (qualified candidates selected) are equal for male and female candidates.
- **Disparate Impact Ratio:** Measure the ratio of favorable outcomes for female vs. male candidates; aim for a ratio close to 1.
- **False Negative Rate by Gender:** Monitor if qualified female candidates are wrongly rejected more often.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

---

1. Discuss Ethical Risks

- **Wrongful Arrests and Detentions:** Misidentification can lead to innocent minority individuals being falsely accused, arrested, or detained, causing severe personal and legal consequences.
- **Privacy Violations:** Deployment without consent or transparency infringes on individuals' privacy rights, especially marginalized communities.
- **Reinforcement of Systemic Biases:** Using biased facial recognition deepens racial profiling and discrimination by law enforcement.
- **Loss of Public Trust:** Such errors undermine confidence in policing and technology, potentially causing social unrest.

---

2. Recommend Policies for Responsible Deployment

1. **Rigorous Bias Testing:** Require independent audits for accuracy and bias across all demographic groups before deployment.

2. **Transparency and Accountability:** Police departments must disclose when and how facial recognition is used and provide mechanisms for redress in case of errors.

3. **Limit Use Cases:** Restrict facial recognition to high-risk, serious crimes only, with strong legal oversight.

4. **Human Oversight:** Ensure that AI results are reviewed by trained human officers who consider context before taking action.

5. **Community Engagement:** Engage with affected communities to build trust, understand concerns, and co-develop guidelines.

6. **Data Privacy Protections:** Enforce strict data storage, access, and retention policies compliant with privacy laws.