Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

Scenario: Amazon’s AI recruiting tool was found to penalize female candidates, particularly in male-dominated technical roles.

a) Source of Bias:

The primary source of bias was the **training data**. The AI model was trained on ten years of historical resumes submitted to Amazon, which reflected existing gender imbalances in tech hiring. As a result, the model learned to favor resumes that resembled those of previously hired (mostly male) candidates, penalizing terms commonly associated with women, such as “women’s chess club captain.”

b) Proposed Fixes for Fairness:

1. **Balanced Training Data:** Re-train the model using a dataset that ensures gender representation across roles and industries. Use techniques like oversampling underrepresented groups or synthetic data generation.

2. **Bias Mitigation Algorithms:** Apply algorithmic fairness techniques such as re-weighting, adversarial debiasing, or fairness constraints to reduce gender-based disparities during model training.

3. **Feature Auditing and Removal:** Identify and remove features correlated with gender that contribute to biased predictions, such as indirect gender indicators or gendered language.

c) Fairness Evaluation Metrics:

- **Demographic Parity:** Compare selection rates across gender groups to ensure similar proportions are recommended.

- **Equal Opportunity:** Measure the true positive rate for qualified male and female candidates.
- **Disparate Impact Ratio:** Ensure the ratio of favorable outcomes for different groups meets fairness thresholds (e.g., 80% rule).

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system deployed in law enforcement misidentifies individuals from minority groups at higher rates, leading to ethical and legal concerns.

a) Ethical Risks:

- **Wrongful Arrests:** False positives disproportionately affect minorities, leading to arrests based on incorrect identification, which undermines the presumption of innocence and due process.
- **Privacy Violations:** Continuous surveillance and unauthorized biometric data collection infringe on individuals' privacy rights, especially in public spaces.
- **Discrimination and Social Harm:** Systematic misidentification exacerbates existing racial inequalities, undermines community trust in law enforcement, and may reinforce biased policing patterns.

b) Recommended Policies for Responsible Deployment:

1. **Pre-deployment Bias Audits:** Conduct thorough fairness assessments using racially diverse test sets to evaluate model accuracy across demographic groups before use.
2. **Human Oversight Requirements:** Ensure facial recognition results are always reviewed by trained human officers who must verify matches before any enforcement

action.

3. **Usage Transparency and Regulation:** Implement strict guidelines for when and how facial recognition can be used, including public disclosure, independent oversight bodies, and data retention limits.

4. **Opt-out and Consent Mechanisms:** Where feasible, inform citizens and allow them to opt out of facial recognition data processing, especially in non-criminal contexts.