Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

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

1. Source of Bias:

- The AI model was trained on resumes submitted to Amazon over a 10-year period, most of which came from men.

- It learned to favor male-dominated language and penalized resumes that included the word "women's," as in "women's chess club captain."

2. Three Fixes to Make the Tool Fairer:

a) Use balanced training data with equal representation across genders.

b) Remove or neutralize gender-related terms using preprocessing and NLP fairness techniques.

c) Incorporate fairness constraints during model training and validate using fairness metrics.

3. Fairness Evaluation Metrics:

- Disparate Impact Ratio
- Equal Opportunity Difference
- Statistical Parity Difference

Case 2: Facial Recognition in Policing

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

1. Ethical Risks:

- Wrongful arrests due to false positives, especially for people of color.
- Violations of privacy, surveillance without consent, and potential chilling effects on civil liberties.

2. Recommended Policies for Responsible Deployment:

a) Mandatory bias audits before deployment.

b) Human oversight in identification processes - facial recognition should not be the sole basis for arrest.

c) Transparency reports and accountability structures.

d) Public consultation and opt-in mechanisms in affected communities.