Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon) Scenario: Amazon’s AI recruiting tool penalized female candidates.

Identify the Source of Bias:

Training Data: The model was trained on historical resumes, which reflected a male-dominated tech workforce, embedding gender biases (e.g., favoring terms like “engineer” associated with male candidates).

Model Design: The algorithm likely prioritized features correlated with male hires, such as specific job titles or keywords, indirectly penalizing female candidates. Propose Three Fixes:

Anonymize and Diversify Data: Remove gender-identifying information (e.g., names, pronouns) from training data and include diverse resumes to balance representation.

Apply Fairness Algorithms: Use techniques like adversarial training (e.g., AI Fairness 360’s Adversarial Debiasing) to minimize gender-based disparities in predictions.

Conduct Regular Audits: Implement ongoing fairness checks using metrics like disparate impact ratio, adjusting the model if biases are detected.

Suggest Metrics to Evaluate Fairness Post-Correction:

Disparate Impact Ratio: Measures whether selection rates (e.g., shortlisting) are equitable across genders (target: ratio ≈ 1).

Equal Opportunity Difference: Ensures true positive rates (e.g., correctly identifying qualified candidates) are similar for male and female candidates.

Demographic Parity: Ensures the proportion of selected candidates is balanced across gender groups.

Case 2: Facial Recognition in Policing Scenario: A facial recognition system misidentifies minorities at higher rates.

Tasks:

Discuss Ethical Risks:

Wrongful Arrests: Higher false positive rates for minorities increase risks of unjust detentions or convictions, exacerbating systemic inequities.

Privacy Violations: Mass surveillance disproportionately targets marginalized communities, eroding trust and violating personal autonomy.

Bias Amplification: Misidentifications feed back into training data, perpetuating and worsening existing biases. Recommend Policies for Responsible Deployment:

Pre-Deployment Bias Testing: Mandate audits using metrics like false positive rate disparity across racial groups before system deployment.

Human-in-the-Loop Oversight: Require human review of AI outputs before actions like arrests, with clear criteria for intervention.

Transparency and Accountability: Publicly disclose system performance metrics and usage policies; establish independent oversight boards.

Restricted Use: Ban facial recognition in high-stakes contexts (e.g., arrests) until biases are mitigated, and require community consent for deployment.