Part 2: Case Study Analysis.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

Scenario Recap:

Amazon developed an AI-powered resume screening tool. However, the tool penalized resumes with the word “women’s” (e.g., “women’s chess club”), leading to **systemic gender bias** against female candidates.

1. Source of Bias:

- **Training Data Bias:**
  The model was trained on resumes submitted to Amazon over a 10-year period — data that reflected **male-dominated hiring patterns** in tech. As a result, the algorithm learned to associate male terms with success.
- **Feature Selection Bias:**
  Words related to female-associated activities were weighted negatively without context — an example of **proxy discrimination**.
- **Lack of Fairness Constraints in Model Design:**
  No fairness metrics or bias mitigation strategies were embedded into the training pipeline.

2. Three Fixes to Make the Tool Fairer:

1. **Balanced and Representative Training Data:**
   Actively curate training datasets to include equal representation of genders, ethnicities, and educational backgrounds. Use **reweighing techniques** from libraries like AI Fairness 360.
2. **Blind Sensitive Attributes During Feature Engineering:**
   Remove or obfuscate features that directly or indirectly encode gender, such as pronouns or gendered terms in activities.
3. **Incorporate Fairness-Aware Algorithms:**
   Use fairness-aware classifiers or post-processing tools (e.g., **Reject Option Classification**, **Equalized Odds**) to balance performance across groups.

3. Fairness Metrics Post-Correction:

- **Disparate Impact Ratio (DIR):**
  Measures whether selection rates between groups (e.g., male vs. female) are equitable. Ideal ratio ≈ 1.0.

- **Equal Opportunity Difference:**
  Compares true positive rates for different groups. A value close to 0 indicates fairness.

- **Statistical Parity Difference:**
  Measures difference in positive outcomes between groups, aiming for 0.

Case 2: Facial Recognition in Policing

Facial recognition tools used by law enforcement have shown **higher false positive rates** for people of color — leading to **misidentifications**, **privacy concerns**, and **trust issues** in marginalized communities.

1. Ethical Risks:

- **Wrongful Arrests & Legal Harm:**
  False matches can lead to arrest and prosecution of innocent individuals — violating principles of **justice** and **non-maleficence**.

- **Racial Discrimination:**
  Systemic inaccuracies disproportionately affect minority groups, exacerbating **bias** and reinforcing **historical inequities**.

- **Privacy Violations:**
  Continuous public surveillance without consent undermines **autonomy** and the **right to privacy**.

- **Loss of Trust in Law Enforcement:**
  Citizens may perceive AI policing tools as tools of oppression rather than protection.

2. Policies for Responsible Deployment:

1. **Mandatory Independent Bias Audits:**
   Require third-party evaluations of facial recognition systems for accuracy across racial, gender, and age groups before deployment.

2. **Human Oversight Protocols:**
   Use AI only as an assistive tool — not the sole basis for identification or arrest. Ensure a

human decision-maker validates all matches.

3. **Strict Consent and Transparency Rules:**
   Deploy facial recognition only with public knowledge, local government approval, and in compliance with data protection laws (e.g., GDPR-like standards).

4. **Ban in High-Risk Contexts:**
   Until accuracy and fairness improve, restrict use in sensitive areas like immigration enforcement, schools, or political protests.