# Part 2: Case Study Analysis

## Case 1: Biased Hiring Tool

### Scenario:

Amazon developed an AI recruiting tool intended to help streamline the hiring process. However, it was discovered that the tool penalized female candidates, showing bias against resumes that contained certain words associated with women, such as "women’s chess club" or all-female colleges.

### 1. Identify the source of bias

The primary source of bias was the **training data**. The AI model was trained on historical hiring data collected over a ten-year period. Since most of the successful candidates during that time were men—especially in technical roles—the model learned to associate male-related patterns and experiences with success. This means the system "learned" that male candidates were preferable, not because of actual merit, but because of biased historical hiring practices.

Another contributing factor was **model design**, which did not include bias detection or fairness constraints. The AI evaluated resumes using keywords and patterns without understanding context or intent, which allowed gender-based associations to influence its rankings unfairly.

### 2. Propose three fixes to make the tool fairer

1. **Train the model on a balanced and diverse dataset:**
   Rebuild the dataset with equal representation of male and female applicants, and ensure inclusion of diverse educational, cultural, and professional backgrounds. This will help the model learn patterns based on merit rather than historical bias.
2. **Integrate fairness constraints and regular auditing:**
   Include fairness-aware algorithms and tools (e.g., IBM AI Fairness 360) to detect and reduce bias during training and deployment. Regular audits should be conducted to evaluate the model’s performance across gender, race, and other demographics.
3. **Remove or de-emphasize gender-indicating features:**
   Features such as names, gender-specific pronouns, and words like “women’s” should either be excluded or neutralized in the feature engineering process, so they don’t skew the model’s results.

### 3. Suggest metrics to evaluate fairness post-correction

- **Demographic Parity:**
  Measures whether the model recommends candidates from different demographic groups at similar rates. For example, are male and female applicants selected at roughly equal proportions?
- **Equal Opportunity Metric:**
  Assesses whether the model gives equal chances of selection to candidates from different groups who are similarly qualified. This ensures fairness in outcomes based on merit.
- **Disparate Impact Ratio:**
  Compares the rate of positive outcomes (such as being shortlisted) between groups. A ratio below 0.8 often indicates potential bias under legal fairness standards.

These metrics should be reviewed consistently to ensure the tool performs equitably after each update or change.

---

## Case 2: Facial Recognition in Policing

### Scenario:

Several police departments have adopted facial recognition systems to identify suspects and monitor public spaces. However, studies have shown that these systems misidentify individuals from minority groups—especially Black and Asian individuals—at significantly higher rates than white individuals, leading to serious ethical concerns.

---

### 1. Discuss ethical risks

- **Wrongful Arrests:**
  Misidentification can lead to people being wrongfully accused or arrested for crimes they did not commit. This is particularly concerning in high-stakes criminal investigations, where the consequences can be life-altering.
- **Privacy Violations:**
  Facial recognition often operates without individuals’ consent. When used in public spaces, it can track people’s movements and behaviour continuously, creating a surveillance environment that threatens personal privacy.
- **Discrimination and Inequity:**
  When accuracy varies across racial or gender groups, it results in discriminatory treatment. Minority communities are more likely to be stopped, questioned, or detained due to faulty matches.
- **Loss of Public Trust:**
  Widespread errors and lack of transparency erode trust in law enforcement

and in the technology itself. If the public feels unfairly targeted, cooperation with authorities may decline.

---

## 2. Recommend policies for responsible deployment

1. **Mandatory Accuracy Audits Across Demographics:**
   Require vendors and law enforcement agencies to test and publish performance data for different races, genders, and age groups. Systems must meet a minimum accuracy standard for all groups before deployment.
2. **Human-in-the-loop Review:**
   AI matches should never lead directly to enforcement actions without human confirmation. Officers must review AI-generated matches and make independent judgments before taking action.
3. **Strict Use Policies and Consent Guidelines:**
   Limit facial recognition use to serious criminal investigations. Where possible, require court warrants or informed consent before using facial recognition systems on individuals.
4. **Data Protection and Transparency Measures:**
   All captured images and data must be securely stored, and retention periods should be clearly defined. The public must be informed about when and where facial recognition is used, and individuals should have the right to request access to their data or challenge its use.
5. **Community Engagement and Oversight:**
   Involve community representatives and ethics boards in the planning and evaluation process. Their input ensures diverse perspectives and helps align deployment with social values and legal expectations.