1. Source of Bias

The primary source of bias in Amazon's AI recruiting tool was **biased training data**. Specifically:

- The model was trained on **resumes submitted to Amazon over a 10-year period**, the majority of which came from **male applicants** due to historical gender imbalance in the tech industry.
- As a result, the AI learned to **downgrade resumes that included terms or experiences more commonly associated with women**, such as references to all-women colleges, or mentions of "women's" organizations or clubs.
- The model design did not include mechanisms to detect or mitigate **gender-related patterns** in its predictions, leading to systemic gender bias.

2. Three Fixes to Make the Tool Fairer

Fix 1: Debias the Training Data

- **Action**: Clean and balance the dataset by ensuring equal representation of male and female candidates.
- **Approach**: Augment the dataset with qualified female candidate resumes and remove or reduce reliance on gender-correlated features.

Fix 2: Implement Fairness Constraints in Model Design

- **Action**: Incorporate fairness-aware algorithms or constraints during model training.
- **Approach**: Use techniques like **demographic parity**, **equalized odds**, or **adversarial debiasing** to minimize discriminatory outcomes.

Fix 3: Strip or Mask Sensitive Attributes and Proxies

- **Action**: Remove explicit and implicit gender indicators from resumes before model processing.
- **Approach**: Mask terms like names, gendered pronouns, or organizations that could act as proxies for gender (e.g., "Women in Tech").

3. Metrics to Evaluate Fairness Post-Correction

Metric 1: Demographic Parity

- Measures whether candidates from different gender groups are selected at **equal rates**, regardless of their actual qualifications.

Metric 2: Equal Opportunity

- Checks if the **true positive rate** (i.e., correctly identifying qualified candidates) is equal across gender groups.

Metric 3: Disparate Impact Ratio

- Ratio of selection rates between groups (e.g., women vs. men). A commonly used threshold is **80% rule** (the ratio should be no less than 0.8).

Simulation Overview

Let's simulate a hiring model that evaluates 1,000 candidates:

- 500 male candidates
- 500 female candidates
  Each candidate has a "qualification score" (0 to 100), and the model outputs a binary decision: **hire (1)** or **reject (0)**.

Before Fix: Biased Model

Gender	Qualified (score ≥ 70)	Hired	Hiring Rate	True Positive Rate
Male	300	240	48%	80% (240/300)
Female	280	140	28%	50% (140/280)

Bias Observed:

- **Disparate Impact Ratio**: 28% / 48% = **0.58** (less than 0.8 → bias)
- **Equal Opportunity Violation**: Males have a higher TPR than females

After Fix: Fair Model

We apply the three fixes:

- Balance dataset
- Strip gendered terms
- Use fairness-aware model training

Gender	Qualified (score ≥ 70)	Hired	Hiring Rate	True Positive Rate
Male	300	240	48%	80%
Female	280	224	44.8%	80%

Fairness Improvements:

- **Disparate Impact Ratio**: 44.8% / 48% = **0.93** ✓
- **Equal Opportunity**: TPR for both = **80%** ✓
- **Demographic Parity Gap** reduced: Hiring rates are more balanced

Visualization (Conceptual)

Before Fix:

Hired Candidates (out of 500)

Men: 240

Women: 140

After Fix:

Men: 240

Women: 224