Case Study Analysis: Biased Hiring Tool (Amazon’s AI Recruiting Tool)

By Nolin Masai

1. Source of Bias

The bias in Amazon’s AI recruiting tool stemmed primarily from **biased training data**. The model was trained on historical resumes submitted to Amazon over a 10-year period, most of which came from male applicants. As a result, the algorithm learned to favor male-dominated language and penalize resumes containing indicators of female identity (e.g., “women’s chess club”).
Additional contributing factors included:

- **Feature selection bias:** The model used features correlated with gender rather than job performance.
- **Lack of fairness constraints:** The model optimization focused solely on predictive accuracy, not equitable outcomes.

2. Proposed Fixes

1. **Balanced and Representative Training Data:**
   Rebuild the dataset to ensure gender balance and remove gender-identifying features. Include resumes from diverse backgrounds and industries to reduce skew.
2. **Fairness-Aware Model Design:**
   Integrate fairness constraints or debiasing algorithms (e.g., reweighing, adversarial debiasing) during model training to minimize disparate impact across gender groups.
3. **Human-in-the-Loop Oversight:**
   Implement a hybrid system where AI recommendations are reviewed by trained HR professionals who apply fairness guidelines before final decisions are made.

3. Fairness Evaluation Metrics

To assess fairness after implementing corrections, the following metrics can be used:

- **Demographic Parity:** Measures whether selection rates are similar across gender groups.
- **Equal Opportunity Difference:** Evaluates whether qualified candidates from all genders have equal chances of being selected.
- **Disparate Impact Ratio:** Compares the hiring rate of one gender to another; a ratio between 0.8 and 1.25 is generally considered fair.