**Analysis of Bias in Amazon’s AI Recruitment Tool**

Amazon’s AI recruitment tool was found to exhibit gender bias, disproportionately favoring male candidates over female candidates. The bias stemmed from the training data, which consisted of resumes submitted to Amazon over a 10-year period—a predominantly male-dominated pool. The AI learned to penalize resumes containing words like "women’s" or affiliations with women’s colleges, while favoring terms more commonly found in male applicants' resumes (e.g., technical jargon from male-dominated fields).

This case highlights a critical issue in AI fairness: **historical biases in training data can perpetuate discrimination**. Since AI models learn patterns from past hiring decisions, they may reinforce existing inequalities rather than mitigate them. Additionally, the lack of diversity in the training dataset exacerbated the problem, as the model was not exposed to enough examples of successful female candidates in technical roles.

### ****Proposed Fairness Improvements****

1. **Debiasing the Training Data**
   * **Action:** Use synthetic data augmentation to balance gender representation or apply reweighting techniques to underrepresented groups.
   * **Justification:** Ensuring the dataset reflects a fair distribution of candidates prevents the model from inheriting historical biases.
2. **Adversarial Fairness Training**
   * **Action:** Implement adversarial debiasing, where a secondary model penalizes the AI for making gender-biased predictions.
   * **Justification:** This forces the model to learn features that are predictive of job performance rather than demographic attributes.
3. **Human-in-the-Loop Oversight**
   * **Action:** Introduce human reviewers to audit AI-generated shortlists, with explicit fairness checks before final decisions.
   * **Justification:** Humans can catch subtle biases that the AI may miss, ensuring accountability.

### ****Recommended Fairness Metrics****

1. **Demographic Parity**
   * Measures whether selection rates are equal across gender groups.
   * **Why?** Ensures no group is disproportionately favored.
2. **Equalized Odds**
   * Checks if the model has similar true positive and false positive rates for male and female candidates.
   * **Why?** Prevents the AI from being more accurate for one group over another.
3. **Disparate Impact Ratio**
   * Compares selection rates between protected and non-protected groups (e.g., women vs. men).
   * **Why?** A ratio close to 1.0 indicates fairness; significant deviations suggest bias.