# Amazon's AI Hiring Tool: A Case of Amplified Bias

Amazon's ambitious attempt to automate its recruitment process with an AI tool infamously backfired when it was discovered to be systematically biased against female candidates.[1] The incident serves as a critical case study in the ethical pitfalls of artificial intelligence, highlighting how even well-intentioned technological solutions can perpetuate and even amplify existing societal biases.

## The Root of the Problem: Biased Training Data

The primary source of bias in Amazon's recruiting tool was the **historical data** used to train the model. The AI was fed a decade's worth of resumes submitted to the company. Since the tech industry, and consequently Amazon's workforce at the time, was predominantly male, the algorithm learned to favor candidates with characteristics more commonly found on male resumes.[2]

The model incorrectly identified male-dominated language and patterns as indicators of job success.[3] For instance, it penalized resumes that included the word "women's," such as "women's chess club captain," and downgraded graduates of two all-women's colleges.[4] This demonstrates a classic case of "garbage in, garbage out"; the AI simply mirrored and codified the existing gender imbalance present in the data it was trained on.

## Three Proposed Fixes for a Fairer System

To rectify such a biased system, a multifaceted approach is necessary. Here are three potential fixes:

**1.Diverse and Representative Training Data:** The most crucial step is to curate a new training data-set that is balanced and representative of the gender diversity desired in the workforce. This would involve actively sourcing and including a significant number of resumes from qualified female candidates. The goal is to teach the AI what a successful candidate looks for, independent of gender.

**2.Adversarial Debiasing:** This technique involves building a second AI model that acts as an "adversary" to the primary hiring model. The adversary's goal is to predict the gender of a candidate based on the hiring

model's output. The primary model is then penalized if the adversary can successfully determine gender. This forces the hiring model to make decisions that are not correlated with gender, thus promoting fairness.

**3.Explainable AI (XAI) and Human-in-the-Loop:** Instead of a fully automated decision-making process, the AI tool should serve as a recommendation engine with transparent reasoning. Implementing XAI techniques would allow human recruiters to understand *why* the AI has flagged a particular candidate. This "human-in-the-loop" approach ensures that the final hiring decision rests with a person who can critically evaluate the AI's suggestions and override them if they appear biased.

## 4.Measuring Fairness: Key Metrics for Evaluation

- After implementing corrective measures, it is vital to continuously evaluate the fairness of the AI tool. Here are three key metrics to do so:

- Demographic Parity: This metric assesses whether the proportion of successful candidates from different gender groups is roughly equal. In a fair system, the selection rate for qualified male and female applicants should be comparable. The formula for demographic parity is:

$$P(\text{selected} \mid \text{female}) \approx P(\text{selected} \mid \text{male})$$

- Equal Opportunity: This metric goes a step further and checks if the model accurately identifies qualified candidates equally across different genders. It measures the true positive rate, ensuring that the probability of a qualified female candidate being selected is the same as that of a qualified male candidate. The formula is:

$$P(\text{selected} \mid \text{qualified}, \text{female}) \approx P(\text{selected} \mid \text{qualified}, \text{male})$$

- Predictive Equality: This metric focuses on the precision of the model for each gender. It ensures that among the candidates the model selects, the proportion of those who are actually qualified is the same for both men and women. The formula for predictive equality is:

$$P(\text{qualified} \mid \text{selected}, \text{female}) \approx P(\text{qualified} \mid \text{selected}, \text{male})$$

By implementing these fixes and continuously monitoring these fairness metrics, organizations can work towards developing AI hiring tools that are not only efficient but also equitable and free from the biases of the past.