
✅ Case Study: Amazon's Biased AI Recruiting Tool

📌 Question Breakdown:

1. **Identify the source of bias**
2. **Propose 3 ways to make the tool fairer**
3. **Suggest fairness metrics to evaluate the tool after correction**

🔷 1. Source of Bias

The bias in Amazon's AI recruiting tool came from the **training data**. The system was trained on 10 years of resumes submitted to Amazon, during which most successful applicants were **male**. As a result, the model learned to **favor male-associated language** and penalize resumes that contained indicators of being female, such as:

- Mention of "women's" organizations
- Graduation from all-women colleges
- Lack of male-coded verbs like "executed" or "captured"

This is a classic case of **historical bias being embedded into the model via biased data** — not a technical flaw, but a **data-driven discrimination** issue.

🔷 2. Three Ways to Make the Tool Fair

✅ a) Audit and Clean the Training Data

- Remove or neutralize gender-indicative terms like names, pronouns, or organizations linked to gender (e.g., "Women's Chess Club")
- Ensure balanced representation of male and female candidates in training samples

✅ b) Introduce Fairness Constraints During Model Training

- Use fairness-aware machine learning algorithms (e.g., re-weighting, adversarial debiasing) to reduce discriminatory patterns in the model
- Implement rules that penalize models for biased outputs during training

✅ c) Human-in-the-loop Oversight

- Reintroduce human review into the shortlisting process, especially when the model is unsure
- Recruiters should have visibility into **why** the AI made its decision and be able to override it

---

🔷 3. Fairness Metrics to Use Post-Correction

After applying corrections, it's critical to measure whether fairness has improved. Suggested metrics:

📏 a) Demographic Parity

- Checks if selection rates are similar across gender groups (e.g., % of male vs. female candidates shortlisted)

📏 b) Equal Opportunity Difference

- Measures whether qualified candidates from all groups have equal chances of being selected (True Positive Rate difference between males and females)

📏 c) Disparate Impact Ratio

- A ratio of selection rates across groups. A ratio below 0.8 usually indicates discrimination (the 80% rule)

📌 Summary

Amazon's AI tool failed because it **replicated historical gender bias** baked into past hiring data. To fix this:

- Clean the training data
- Add fairness constraints
- Keep humans involved in critical decisions

Use metrics like **Demographic Parity**, **Equal Opportunity**, and **Disparate Impact** to monitor fairness after correction. These steps ensure the system treats all candidates fairly, regardless of gender.
