GROUP TASK (MODULE 3)

ML Ethics & Bias Case Study

Case Analysis: Amazon AI Recruiting Tool (2014–2017)

Introduction

Machine Learning systems are increasingly used in high-stakes decision-making processes such as hiring, credit scoring, healthcare, and law enforcement. While these systems improve efficiency and scalability, they can unintentionally reproduce historical biases present in training data. This case study analyzes the Amazon AI Recruiting Tool developed between 2014 and 2017, which demonstrated gender bias during resume screening.

The objective of this report is to examine the system architecture, identify observed bias, analyze technical root causes, evaluate fairness metrics, and propose technical as well as governance-based solutions.

1. System Overview

Amazon developed an AI-based resume screening system to automate candidate evaluation. The goal was to rank applicants efficiently by predicting their suitability for technical roles.

Key Components of the System

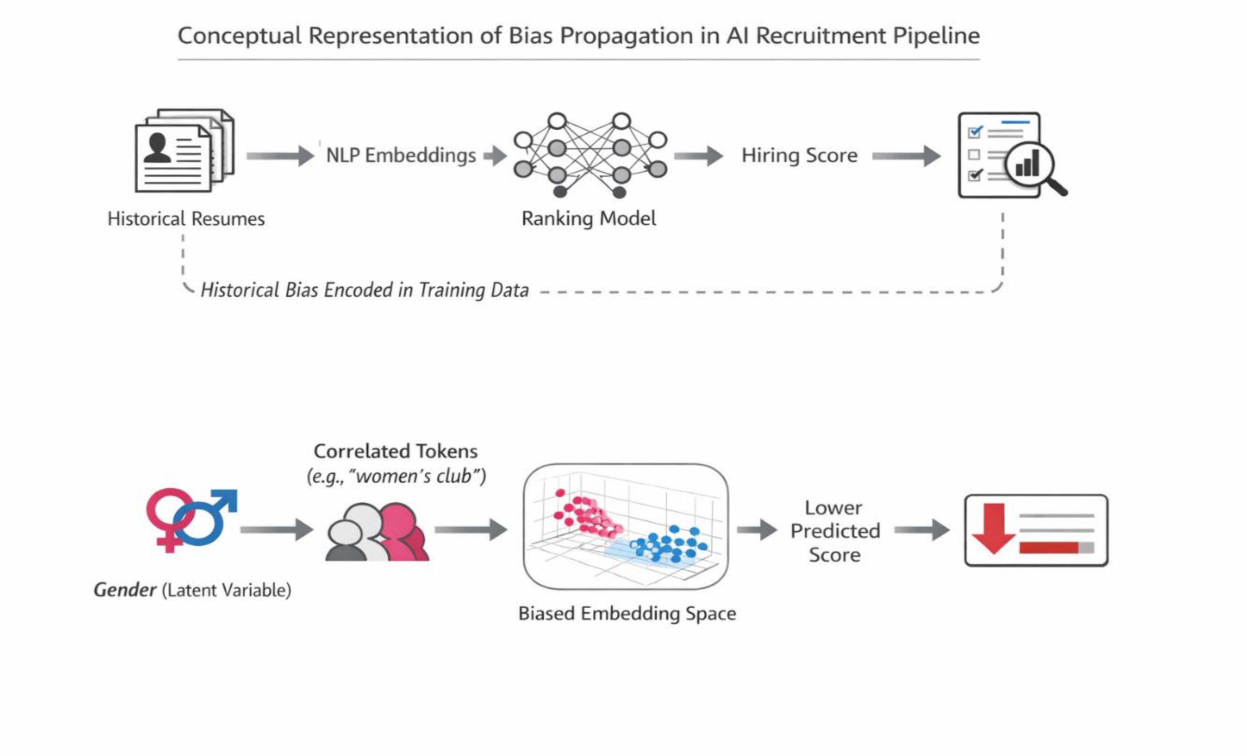
- Resume screening model trained on historical hiring data
- NLP embeddings extracted from resumes' textual content
- Supervised ranking model predicting candidate desirability scores
- Training dataset reflecting decades of male-dominated technology hiring

The model was trained using approximately ten years of past hiring decisions. Since the tech industry historically hired more men than women, this imbalance became embedded in the training data.

Simplified Pipeline Diagram

Historical Resumes
→ Text Embeddings (NLP Processing)

→ Ranking Model
→ Hiring Score



However, historical patterns influenced learning:

Historical Bias
→ Encoded in Training Data
→ Learned by Model
→ Biased Output

This demonstrates how past discrimination can become automated through machine learning.

2. Observed Bias

During internal evaluation, Amazon discovered significant gender-based bias in the model's predictions.

Evidence of Bias

- Resumes containing the word "women's" were penalized
- Candidates from women-only colleges were down-weighted
- The model learned proxy variables correlated with gender
- The system amplified existing workforce gender imbalance

Although gender was not directly included as an input variable, indirect signals within resume text influenced outcomes.

Bias Propagation Graph (Conceptual)

Gender (Latent Variable)
↓
Correlated Tokens (e.g., "women's chess club")
↓
Embedding Space Representation Shift
↓
Lower Predicted Hiring Score

This phenomenon is known as proxy discrimination, where protected attributes affect predictions indirectly through correlated features.

3. Root Causes (Technical Analysis)

The bias emerged due to several technical design flaws.

3.1 Dataset Imbalance

The dataset contained disproportionately more male applicants. Since past hiring favored male candidates:

- Positive labels were skewed toward male-associated features
- Female-associated patterns were underrepresented
- Model learned gender-correlated success indicators

This resulted in skewed label distribution.

3.2 Absence of Fairness Constraints

The optimization objective focused solely on minimizing prediction error.

Objective Function:
Minimize Prediction Loss

The model was not constrained to ensure fairness or equal treatment across groups.

3.3 Proxy Feature Retention

The NLP feature extraction process preserved gender-related signals such as:

- Organizational affiliations
- Linguistic style differences
- Educational background indicators

High mutual-information features correlated with gender were not removed.

3.4 Objective Misalignment

The model minimized classification error instead of minimizing disparate impact.

Example Metric Disparity:

Selection Rate (Male) = 0.45
Selection Rate (Female) = 0.18

Disparate Impact Ratio:

DI = 0.18 / 0.45 = 0.40

Since DI < 0.80 threshold, the system demonstrated adverse impact.

4. Quantitative Fairness Metrics

Evaluating fairness requires structured statistical measurement.

4.1 Statistical Parity Difference (SPD)

SPD = P(selected | male) − P(selected | female)

A large difference indicates imbalance in selection rates.

4.2 Equal Opportunity Gap

TPR Gap = TPR(male) − TPR(female)

Ensures equally qualified candidates have equal probability of selection.

4.3 Disparate Impact Ratio (DI)

DI = Selection Rate (Protected Group) / Selection Rate (Reference Group)

If DI < 0.80, potential discrimination exists.

4.4 Calibration Across Subgroups

Checks whether predicted probabilities are equally reliable across demographic groups.

4.5 Counterfactual Fairness Testing

Procedure:

- Modify gender-related features
- Keep other qualifications constant
- Observe change in predicted score

If prediction changes significantly due to gender-related perturbation, the model lacks fairness.

Fairness Evaluation Flow

Model Predictions
↓
Group Partitioning (Gender)
↓
Compute Metrics:
- SPD
- TPR Gap
- DI Ratio
- Calibration Error

This systematic evaluation helps detect hidden discrimination

5. Proposed Technical Mitigations

Bias mitigation must occur at multiple stages of the ML lifecycle.

5.1 Data-Level Mitigation

- Reweigh samples using inverse propensity scoring
- Balance representation of underrepresented groups
- Remove highly correlated proxy features

5.2 In-Processing Methods

- Adversarial debiasing to remove protected attribute influence
- Fairness-constrained optimization objectives
- Regularization techniques to reduce disparate impact

Modified Objective Example:

Minimize (Prediction Loss + $\lambda \times$ Fairness Penalty)

5.3 Post-Processing Correction

- Apply equalized odds correction
- Adjust decision thresholds across demographic groups

These methods reduce bias in final predictions.

6. Governance & Operational Controls

Technical interventions must be supported by organizational safeguards.

6.1 Pre-Deployment Validation

- Conduct subgroup performance testing
- Evaluate fairness metrics before deployment
- Document results using model cards

6.2 Continuous Monitoring

- Monitor distributional shifts in data
- Track fairness metrics over time
- Implement human-in-the-loop review

6.3 External Oversight

- Conduct periodic third-party audits
- Ensure compliance with anti-discrimination laws
- Establish ethical review committees

Governance ensures accountability and transparency.

7. Generalized Ethical Guidelines

This case demonstrates broader ethical principles in ML systems:

- Align optimization with fairness-aware objectives
- Explicitly encode protected attribute constraints
- Evaluate intersectional subgroup performance
- Prefer interpretable models in high-stakes domains
- Ensure transparency in automated decision systems

Machine learning systems must balance accuracy with equity.

Conclusion

The Amazon AI Recruiting Tool case highlights how machine learning systems can unintentionally amplify historical discrimination when trained on biased data. Although gender was not directly included as a feature, the system learned proxy correlations that disadvantaged female candidates.

This case reinforces several critical lessons:

- Data reflects historical inequality
- Accuracy alone is insufficient in high-stakes decisions
- Fairness must be intentionally engineered
- Governance and monitoring are essential

Ethical AI development requires a combination of technical safeguards, quantitative fairness evaluation, operational oversight, and transparency. Only through responsible design and governance can machine learning systems support equitable and trustworthy decision-making.