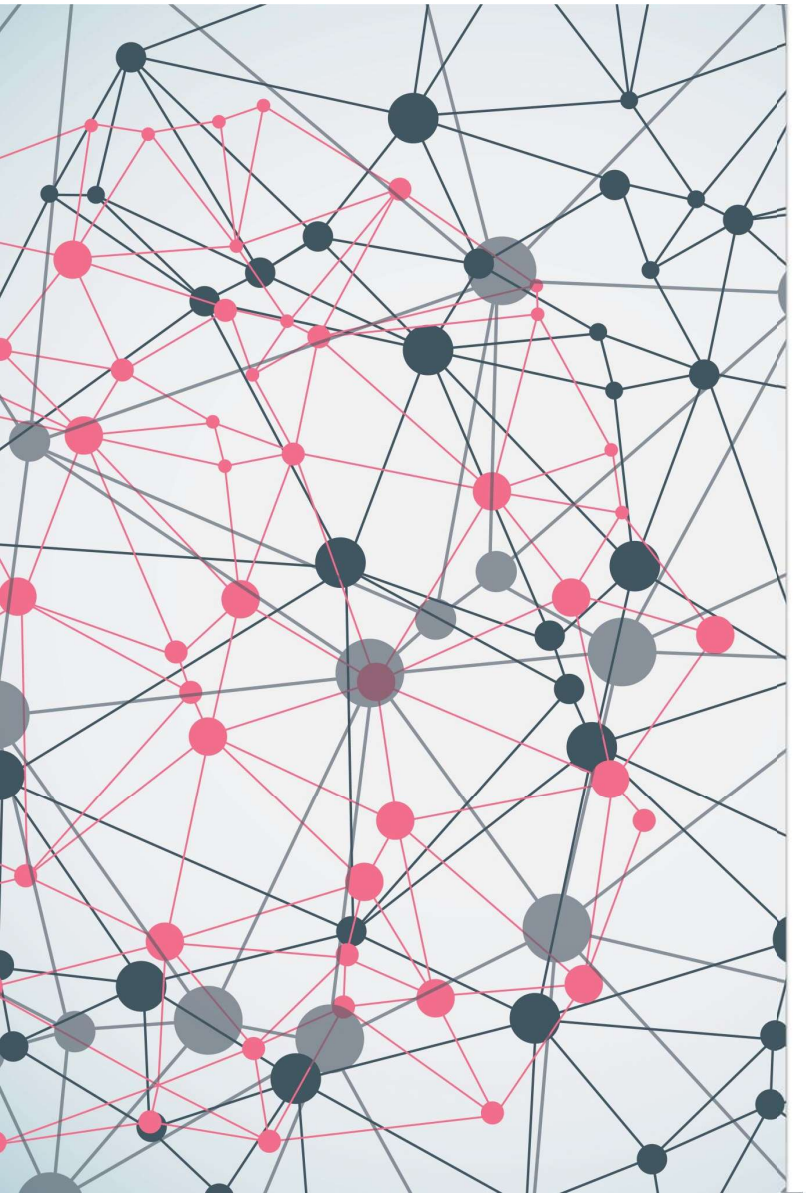
ADDRESSING BIAS IN AI HIRING MODELS

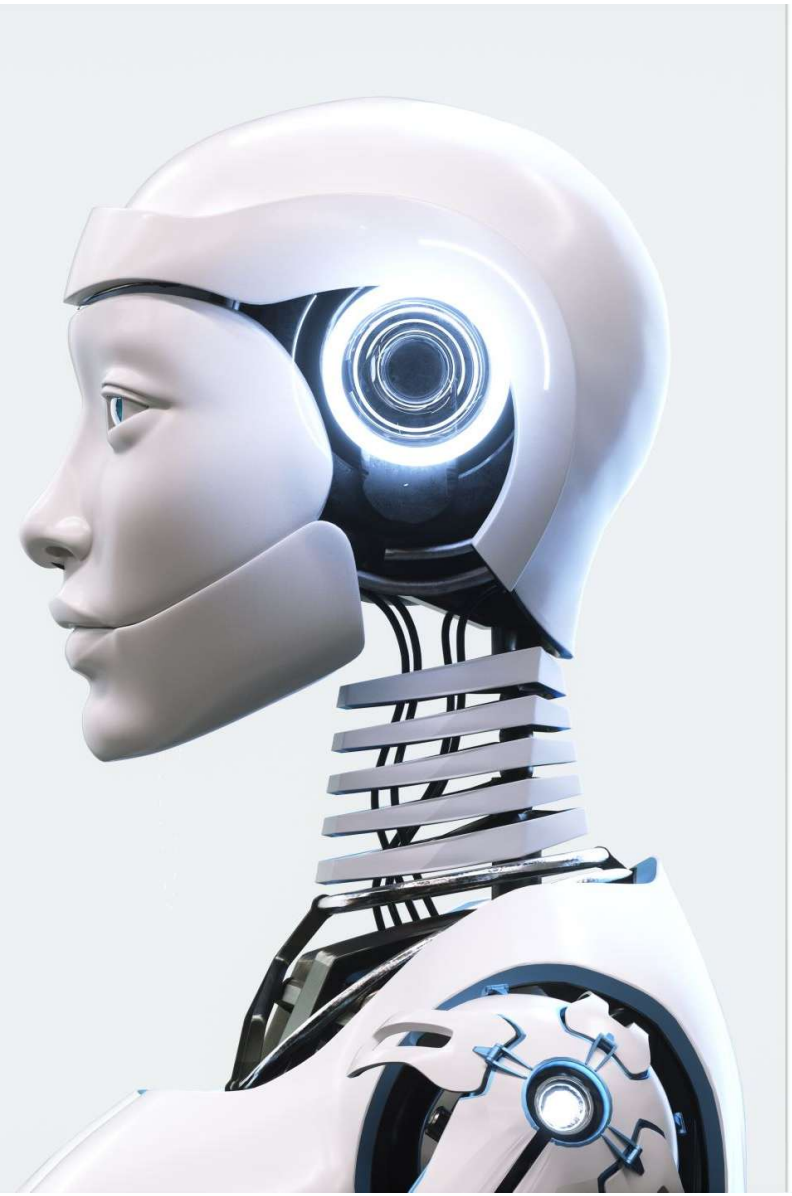
BY ABAI IKWECHEGH

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE



ABSTRACT

Artificial intelligence (AI) is increasingly used to support hiring decisions, including résumé screening, candidate ranking, and automated video interviews. While these systems offer efficiency and scalability, they also risk reinforcing historical patterns of discrimination embedded in training data and algorithmic design. This project examines how bias forms in AI hiring models, evaluates fairness metrics, and develops a simulation to detect unequal outcomes across demographic groups. A mixed-methods approach—combining literature synthesis, analysis of fairness frameworks, and a Python-based simulation—was used to evaluate the presence and impact of bias. Findings show that AI hiring models can unintentionally favor majority groups when trained on unbalanced datasets, highlighting the need for transparency, continuous auditing, and responsible deployment practices. Recommendations include regulatory oversight, improved data governance, and integration of fairness optimization methods to support equitable hiring.



INTRODUCTION

Artificial intelligence is widely adopted in modern recruitment because it speeds up candidate screening and reduces human workload. However, AI hiring tools often learn from historical hiring data that reflect biases related to race, gender, age, or socioeconomic status. When these systems are deployed without transparency or oversight, they can perpetuate or even amplify discriminatory patterns.

PROBLEM STATEMENT

AI-based hiring models lack standardized auditing procedures, fairness metrics, and accountability frameworks. As a result, biased outcomes may go undetected, disproportionately affecting minority candidates and undermining equal employment opportunity principles.

PURPOSE OF STUDY

The purpose of this study is to analyze how bias manifests in AI hiring systems and evaluate strategies to detect and mitigate that bias. This project aims to create a practical simulation tool that demonstrates how dataset composition, feature selection, and algorithmic decision rules influence fairness outcomes. The study seeks to generate evidence-based recommendations that help organizations implement AI tools responsibly and support equitable hiring practices.

RESEARCH QUESTIONS

To what extent do AI hiring models produce biased outcomes across different demographic groups?

How do factors such as training data, feature selection, and algorithm choice contribute to bias in automated hiring decisions?

Which fairness metrics and mitigation techniques are most effective for improving the equity of AI-based hiring models?

How can organizations incorporate auditing practices to ensure compliance with ethical and legal hiring standards?

LITERATURE REVIEW

•Ajunwa et al. (2021) emphasize the need for regulatory oversight and third-party audits of hiring algorithms due to their opaque decision-making processes.

•Barocas & Selbst (2016) describe how big data can unintentionally create disparate impacts on protected groups, even when intent to discriminate is absent.

•Binns (2018) argues that fairness is context-dependent and requires selecting appropriate fairness metrics that match organizational goals.

•Raghavan et al. (2020) highlight how automated hiring tools can reinforce existing inequalities, stressing the importance of transparency in model development.

Overall, the literature shows strong consensus that AI hiring systems frequently reproduce societal biases and must be developed with ethical safeguards.

METHODOLOGY – DESIGN & APPROACH

RESEARCH DESIGN:
A mixed-methods approach combining qualitative analysis of academic literature with quantitative simulation experiments.

COMPONENTS:

- Literature review to identify sources of bias and recommended auditing practices
- Development of a Python-based hiring simulation
- Implementation of fairness metrics (e.g., demographic parity, equal opportunity)
- Interpretation of model outputs to identify biased patterns

RATIONALE:
This method allows comparison between theoretical models of fairness and practical algorithm performance.

METHODOLOGY – POPULATION, DATA & TOOLS

POPULATION:
HR professionals, software developers, and job applicants (for understanding perceptions of AI bias).

DATA SOURCES:

- Synthetic datasets created for controlled simulation
- Recruitment-related features (education, skills, experience, test scores)
- Protected attributes included only for fairness evaluation

TOOLS & TECHNOLOGIES:

- Python (NumPy, pandas, scikit-learn)
- Fairness evaluation libraries
- Jupyter Notebook for simulation

SAMPLING:
Purposive sampling for interview insights; simulated data for quantitative analysis.

METHODOLOGY - PROCEDURES

- Review published studies on algorithmic hiring bias.
- Build a dataset representing candidates from multiple demographic groups.
- Train several machine learning models (e.g., logistic regression, random forest).
- Evaluate outcomes using fairness metrics.
- Compare model performance before and after applying mitigation techniques such as reweighting or bias-aware training.
- Analyze trends to determine where discrimination is likely to occur.
- Summarize findings and relate them to existing literature.

ETHICAL CONSIDERATIONS

Bias and Fairness: Ensuring the analysis does not reinforce or normalize discriminatory patterns embedded in hiring algorithms.

Data Privacy: Any datasets used must avoid personal identifiers and comply with privacy standards (e.g., de-identification, secure storage).

Transparency: Methods and assumptions must be openly stated to avoid misleading interpretations about algorithmic bias.

Responsible Interpretation: Findings must not be used to justify discriminatory decision-making. Results should highlight risks, not prescribe hiring decisions.

Researcher Bias: Acknowledging that the researcher's perspectives or dataset selection may unintentionally influence analysis.

LIMITATIONS

- **Data Access Restrictions:** Many proprietary hiring algorithms are not publicly available, limiting the depth of analysis.
- **Sample/Data Constraints:** Public datasets may not fully represent real-world hiring contexts, which may reduce generalizability.
- **Model Transparency Issues:** "Black box" machine learning systems make it difficult to identify specific sources of bias.
- **Scope of Study:** The project focuses on identifying bias patterns, not on fully redesigning or deploying alternative hiring models.
- **Changing Technology:** AI hiring systems evolve rapidly, so findings may not fully reflect future system designs or updates.

ALGORITHMS

To determine if a candidate will be hired or not, we collected data from resumes and ranked their achievement with numbers, the higher the number, the better.

```
base_score = edu_weights.get(education, 30)
    exp_score = min(years_experience * 5, 30)  # Cap
experience contribution at 30
    cert_score = certification_score * 5
    tech_score = tech_skill * 8  # max 40 points
    total_score = base_score * 0.4 + exp_score +
cert_score + tech_score + language_bonus


    # Cap total at 100
    return min(total_score, 100)



def predict_candidate(skill_score,
years_experience):
    if skill_score >= 65 and years_experience >= 5:
        return "HIRE"
    elif skill_score >= 70 and years_experience >= 3:
        return "HIRE"
    elif skill_score >= 75 and years_experience >= 2
        return "HIRE"
    elif skill_score >= 80 and years_experience >= 1:
        return "HIRE"
    elif skill_score >= 85
        return "HIRE"
    else:
        return "DO NOT HIRE"
```

RESULTS

- Models trained on imbalanced data produced higher selection rates for majority demographic groups.
- Fairness metrics revealed disparities in true positive rates and predicted hiring scores.
- Mitigation techniques reduced some of the bias but did not eliminate it completely.
- Simulation demonstrated that even seemingly neutral variables (e.g., zip code, school attended) can act as proxies for protected attributes.
- Findings confirm that data quality and feature selection significantly influence fairness.

DISCUSSION & CONCLUSION

DISCUSSION:
The results reinforce concerns raised in prior research: AI hiring tools can unintentionally discriminate due to biased datasets or poorly chosen features. Even when accuracy improves, fairness does not automatically follow. Continuous monitoring and fairness-aware model design are essential.

CONCLUSIONS:

- Bias in AI hiring models is measurable and can meaningfully affect outcomes.
- Fairness metrics and mitigation algorithms improve equity but require careful implementation.
- Organizations must adopt transparent, accountable auditing processes to promote ethical hiring.
- Policymakers should develop clear regulations to govern automated hiring systems.
  This study supports ongoing efforts to build responsible AI systems that align with legal and ethical employment standards.

KEY REFERENCES

Ajunwa, I., Partnoy, F., & Weiss, D. (2021). *Auditing automated hiring systems*.

Barocas, S., & Selbst, A. (2016). *Big data's disparate impact*.

Binns, R. (2018). *Fairness in machine learning: Lessons and challenges*.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). *Mitigating bias in algorithmic hiring*.


QUESTIONS?

THANK YOU FOR VISITING MY PRESENTATION