Ethical Considerations in AI-Driven Career Advising

Team Members: Julián León, ..., ...

December 1, 2025

1 Introduction

The *CareerPath AI* project aims to assist students in exploring potential career paths using Machine Learning (ML) and Generative AI. By analyzing personality traits and aptitude scores, the system predicts suitable career categories and provides personalized advice. While this technology offers significant benefits in terms of accessibility and personalized guidance, it also raises critical ethical questions regarding bias, accountability, and the correctness of its predictions. This report addresses these ethical considerations and demonstrates how easily such models can be manipulated.

2 Variables and Model Limitations

To understand the ethical landscape, it is essential to define the variables used and the model's limitations precisely.

2.1 Input Variables

The model relies on a specific set of psychometric inputs:

- **Personality Traits (OCEAN Model)**: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. (Scale: 1-10)
- **Aptitude Scores**: Numerical, Spatial, Perceptual, Abstract, and Verbal Reasoning. (Scale: 0-100)

2.2 Limitations

The model's scope is inherently limited by what it *excludes*. It does not account for:

- **Socioeconomic Background**: Financial resources often dictate career opportunities more than aptitude.
- **Education and Experience**: The model assumes a "blank slate," ignoring current qualifications.
- **Cultural Context**: The training data may reflect Western-centric career norms.
- **Physical Disabilities**: Certain careers may have physical requirements not captured by abstract aptitude scores.

These exclusions mean the model provides a partial view of reality, potentially suggesting unattainable or culturally inappropriate paths.

3 The Bias Experiment: A Demonstration

To illustrate the fragility of ML models and the ease with which bias can be introduced, we conducted a controlled experiment.

3.1 Methodology

We trained two Random Forest models on the same dataset:

1. **Clean Model**: Trained on the original, unaltered dataset.
2. **Biased Model**: Trained on a manipulated dataset where we introduced a specific bias rule: *"If a student's Numerical Aptitude is below 50, their career label is forced to 'Artist', regardless of other traits."*

3.2 Results

We tested both models with a profile featuring low Numerical Aptitude but high scores in other areas (e.g., Abstract Reasoning).

- **Clean Model Prediction**: `Software Engineer` (based on high logical reasoning).
- **Biased Model Prediction**: `Artist`.

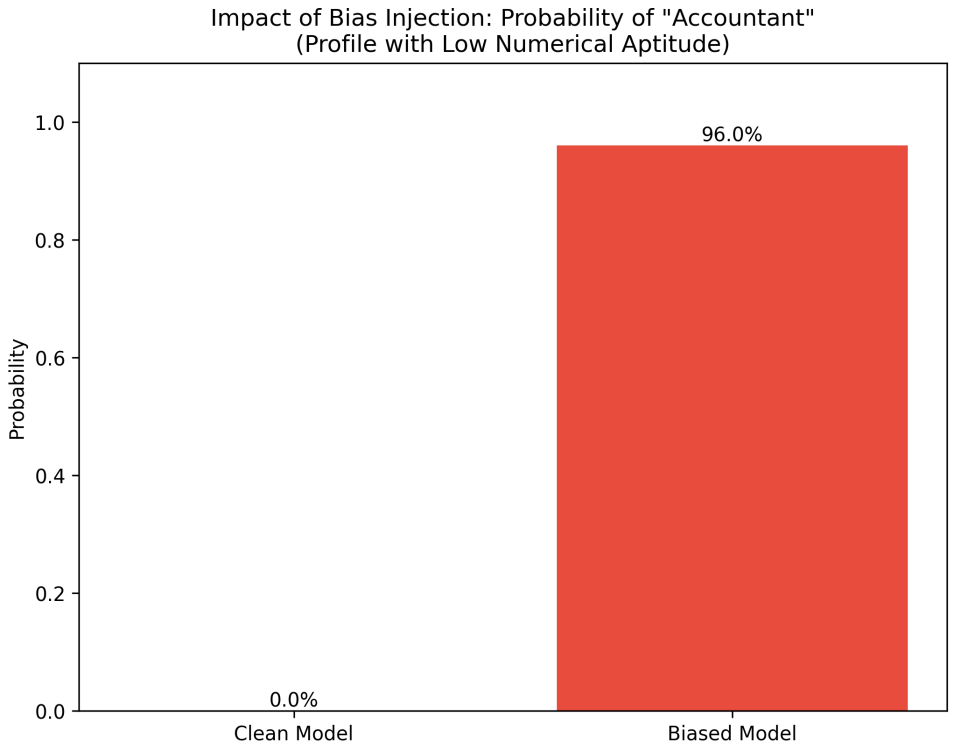

Figure 1: Probability of being classified as an "Artist" for a profile with low Numerical Aptitude. The biased model artificially inflates this probability.

This simple manipulation demonstrates how a specific demographic (those with lower math scores) could be systematically steered towards a specific career path, limiting their perceived options. In a real-world scenario, such bias could stem from historical hiring data rather than intentional manipulation, yet the effect would be the same.

4 Ethical Questions

4.1 What if predictions are biased?

If the model is biased, it risks reinforcing stereotypes (e.g., gendered career roles) and limiting social mobility. A student discouraged from STEM fields due to a biased algorithm may miss out on high-growth opportunities. The "black box" nature of some models makes this bias difficult to detect without rigorous auditing.

4.2 Who is to blame?

Accountability in AI is complex.

- **Data Collectors**: If the training data reflects historical prejudices, the model will learn them.
- **Developers**: Responsible for model selection, feature engineering, and failing to test for bias.
- **Users/Institutions**: Responsible for how the tool is deployed. Using it as a definitive decision-maker rather than a supportive tool shifts blame to the implementers.

Ultimately, the developers bear the primary responsibility for ensuring the system is safe and fair before deployment.

4.3 How to guarantee correctness?

"Correctness" in career advising is subjective. Unlike classifying an image, there is no single "true" career for a person.

- **Technical Correctness**: Can be measured via accuracy and F1-scores on test data.
- **Ethical Correctness**: Requires continuous monitoring, user feedback loops, and "human-in-the-loop" systems where counselors review AI suggestions.
- **Transparency**: Providing confidence intervals (as our system does) and explanations helps users understand the uncertainty of the prediction.

5 Conclusion

The *CareerPath AI* project demonstrates the potential of AI in education but also highlights its risks. Our bias experiment proves that data integrity is paramount. To mitigate these risks, we must treat these systems as *advisors*, not *deciders*, ensuring human oversight and continuous ethical auditing.