Botspeak.ai

Diagnose in Botspeak Loop:

Making AI Decisions Transparent

Anusha Prakash
002306070

www.northeastern.edu

Table of Content

01 Diagnose in Botspeak

02 Philosophical Foundation

03 Case Study

04 Acceptance Test

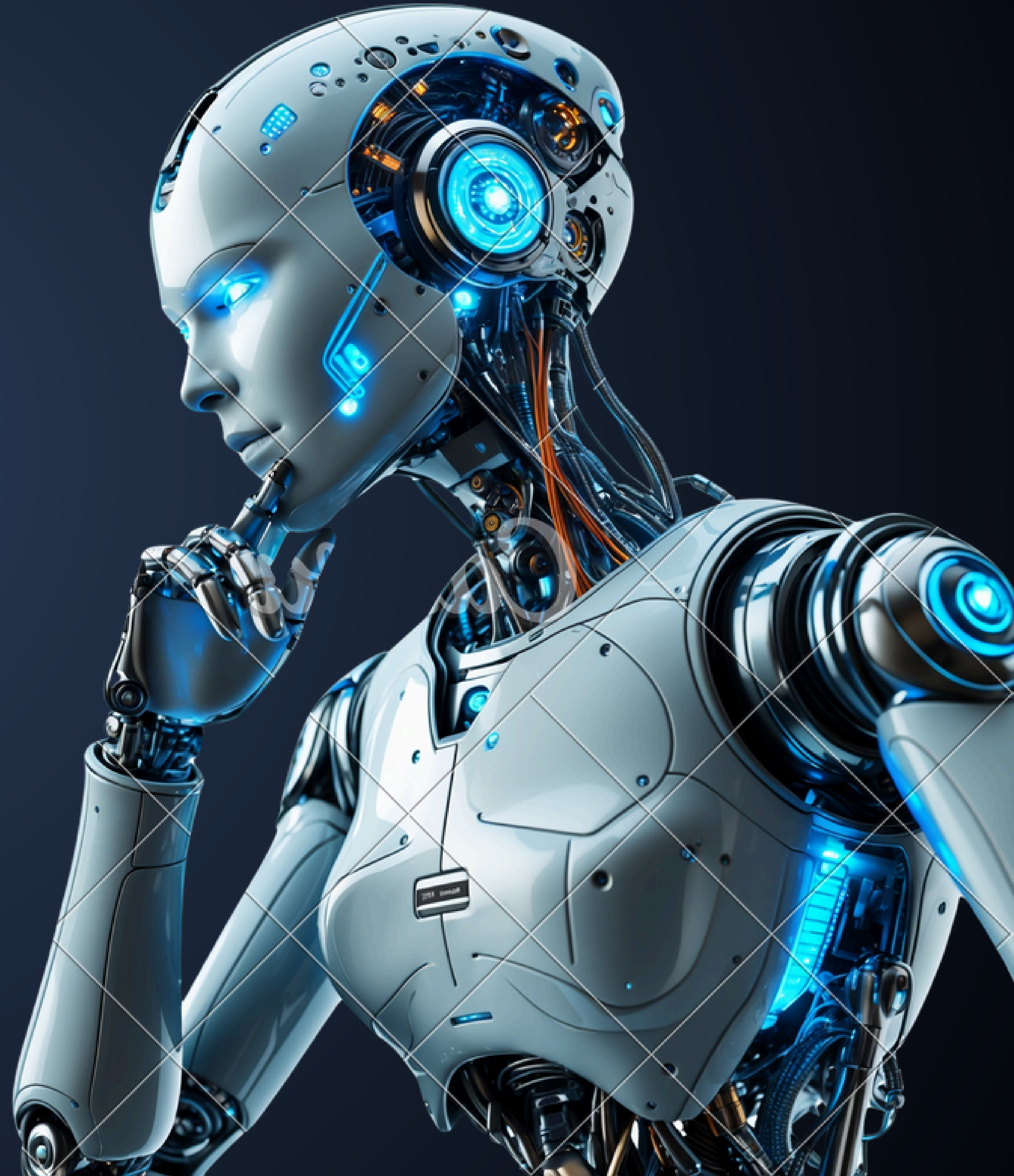
05 Human - AI Validation

06 Consistency Analysis

07 Bias Detection

08 Conclusion

09 References



Diagnose in Botspeak

The critical evaluation phase, where systematic skepticism is applied to AI outputs before acceptance.

Key Components:

- Acceptance Testing
- Uncertainty Assessment
- Bias Detection
- Adversarial Testing
- Evidence Documentation

Philosophical Foundation

Objetivo 01

Descartes - Systematic Doubt

"Question AI claims until proven reliable."
Application: Never accept AI outputs at face value

Objetivo 02

Hume - Problem of Induction

"Past performance doesn't guarantee future results."
Application: Test AI reliability in new contexts

Objetivo 03

Popper - Falsifiability

"Actively seek evidence that AI might be wrong."
Application: Design tests to find AI limitations

Case Study – AI Resume Screening

Problem: AI may discriminate against qualified candidates

Diagnostic Solution:

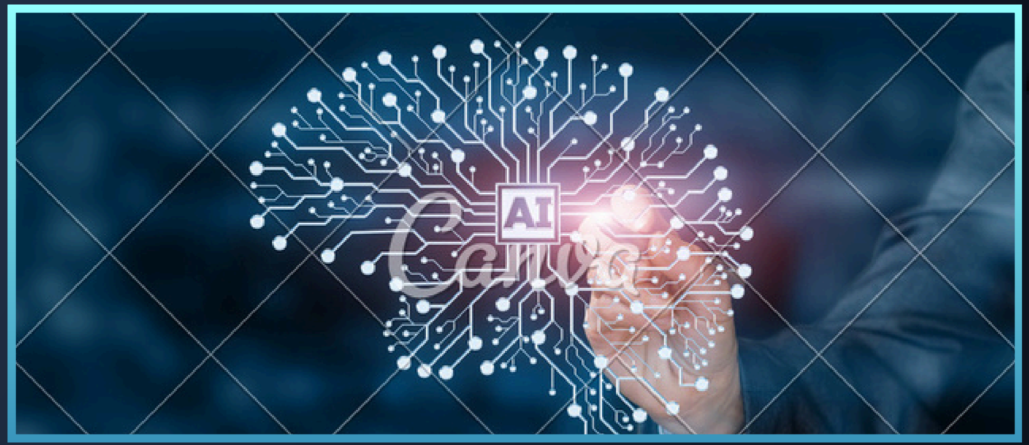
01

Human Review: Human experts check AI decisions before they become final. This catches mistakes that automated systems miss.

02

Bias Testing: We compare how AI treats different groups of people. If one group gets rejected more often, that signals unfair treatment.

03

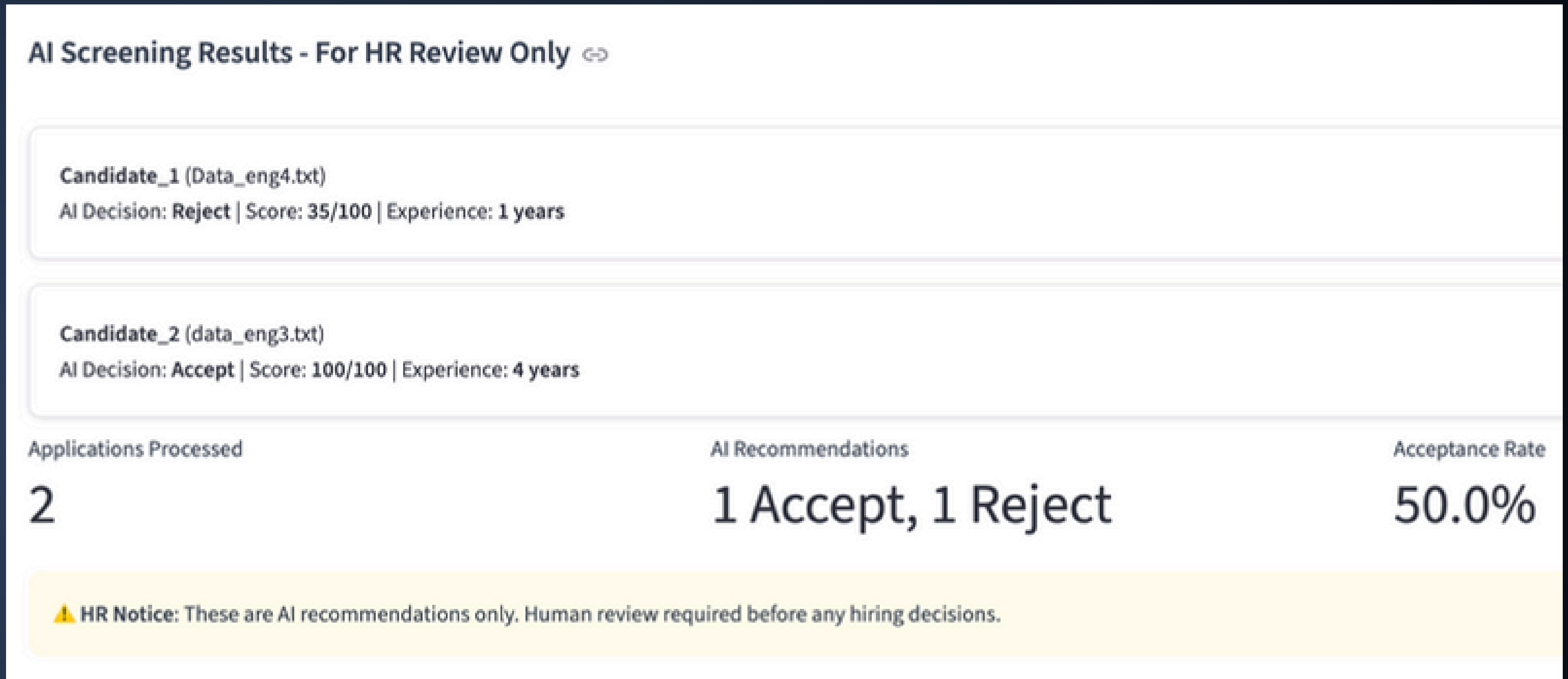


Override Protocol: When humans disagree with AI, the human decision wins. This keeps people in control of important hiring choices.

Acceptance Test : Threshold-Based Decisions

A systematic check to confirm whether AI outputs meet predefined performance criteria before deployment.

- Minimum experience requirement (≥ 3 years for Data Engineer)
- Required skills presence (Python, SQL, ETL, Spark, Airflow)
- Education qualification verification (Bachelor's/Master's in relevant field)



AI Screening Results - For HR Review Only

Candidate_1 (Data_eng4.txt)
AI Decision: **Reject** | Score: **35/100** | Experience: **1 years**

Candidate_2 (data_eng3.txt)
AI Decision: **Accept** | Score: **100/100** | Experience: **4 years**

Applications Processed	AI Recommendations	Acceptance Rate
2	1 Accept, 1 Reject	50.0%

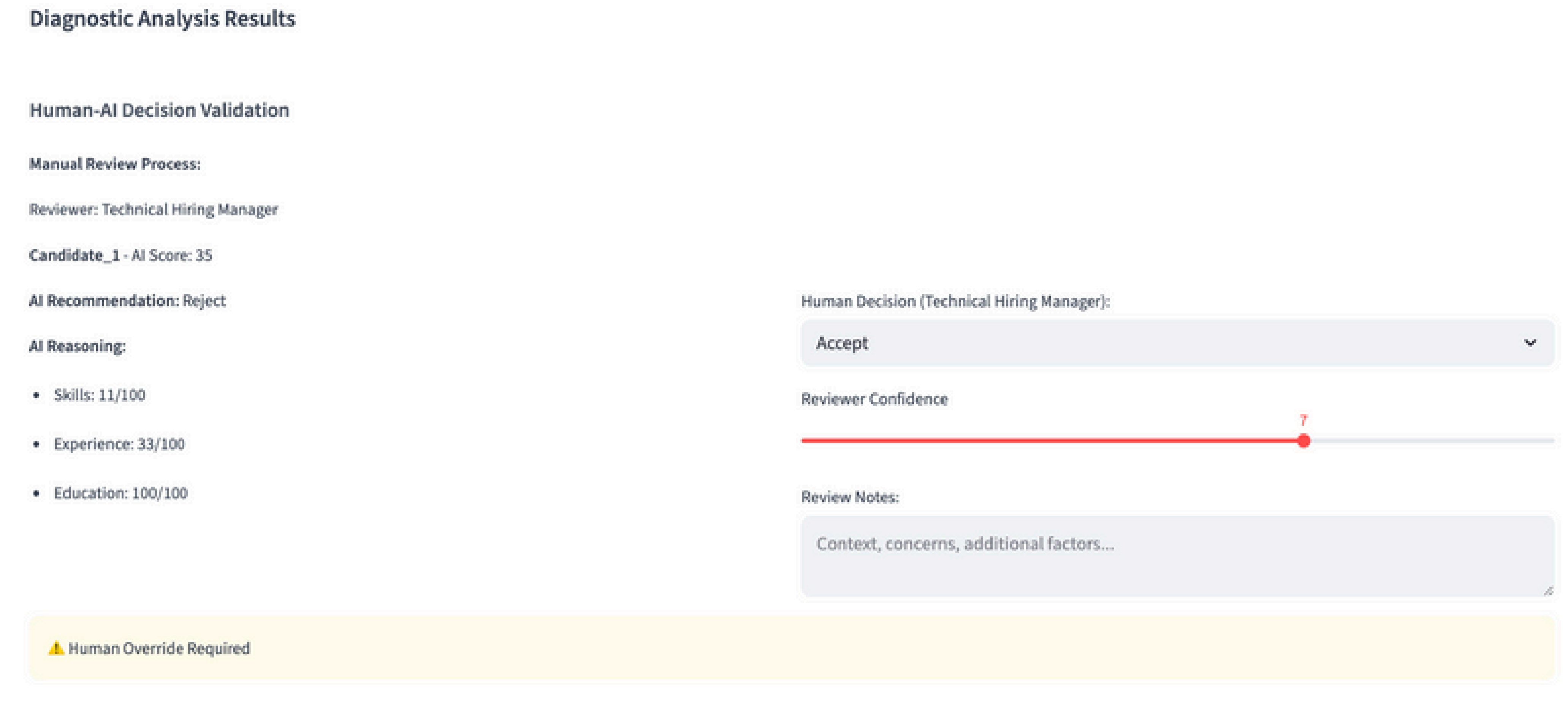
⚠ **HR Notice:** These are AI recommendations only. Human review required before any hiring decisions.

Total Score = (Skills × 0.5) + (Experience × 0.3) + (Education × 0.2)
Decision = Accept if Score ≥ 70, else Reject

Human-AI Validation: Agreement Analysis

Comparing human expert decisions with AI recommendations to measure system reliability.

- Candidate 1: AI says "Reject" (35 points), Human says "Accept" - DISAGREEMENT
- This creates 50% agreement rate and 1 human override
- Low agreement (50%) triggers alert: "AI system needs significant improvement"



Diagnostic Analysis Results

Human-AI Decision Validation

Manual Review Process:

Reviewer: Technical Hiring Manager

Candidate_1 - AI Score: 35

AI Recommendation: Reject

AI Reasoning:

- Skills: 11/100
- Experience: 33/100
- Education: 100/100

Human Decision (Technical Hiring Manager): Accept

Reviewer Confidence: 7

Review Notes: Context, concerns, additional factors...

⚠ Human Override Required

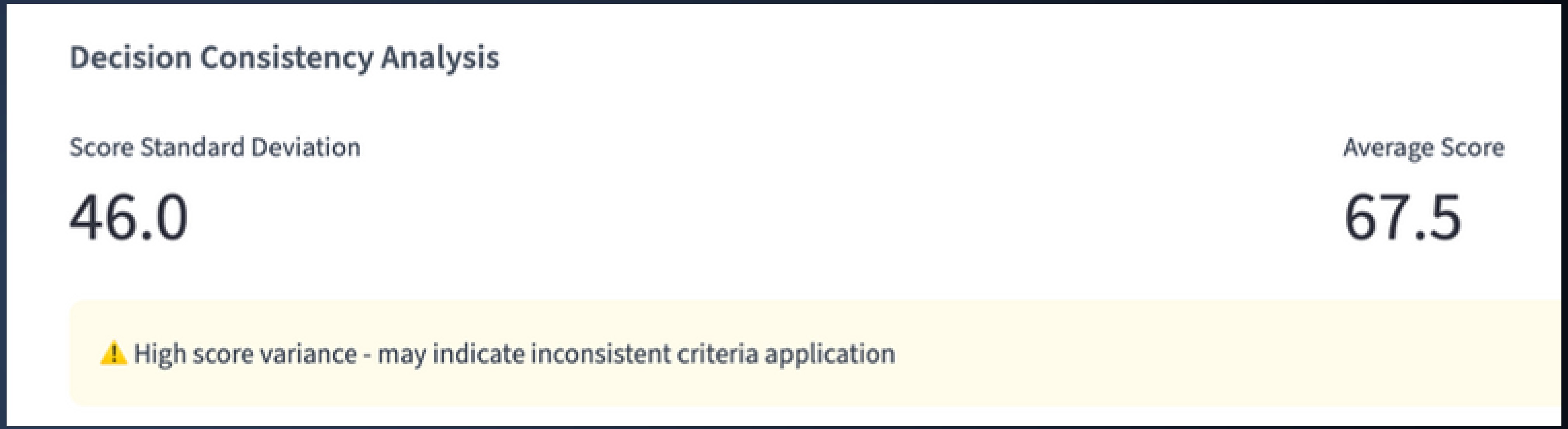
Agreement Rate = (Cases where Human = AI) / Total Cases
Override Rate = (Cases where Human ≠ AI) / Total Cases
Inter-rater Reliability = Statistical measure of consistent decision-making

Consistency Analysis: Score Variance

Measuring how much AI scores vary to identify potential system problems or inconsistent criteria application.

- Score Standard Deviation: 46.0 (high variance)
- Average Score: 67.5
- Warning: "High score variance - may indicate inconsistent criteria application"

If all candidates get very different scores, AI is applying criteria inconsistently. If all candidates get identical scores, AI might be broken. Moderate variance suggests normal, healthy scoring patterns.



Decision Consistency Analysis

Score Standard Deviation

46.0

Average Score

67.5

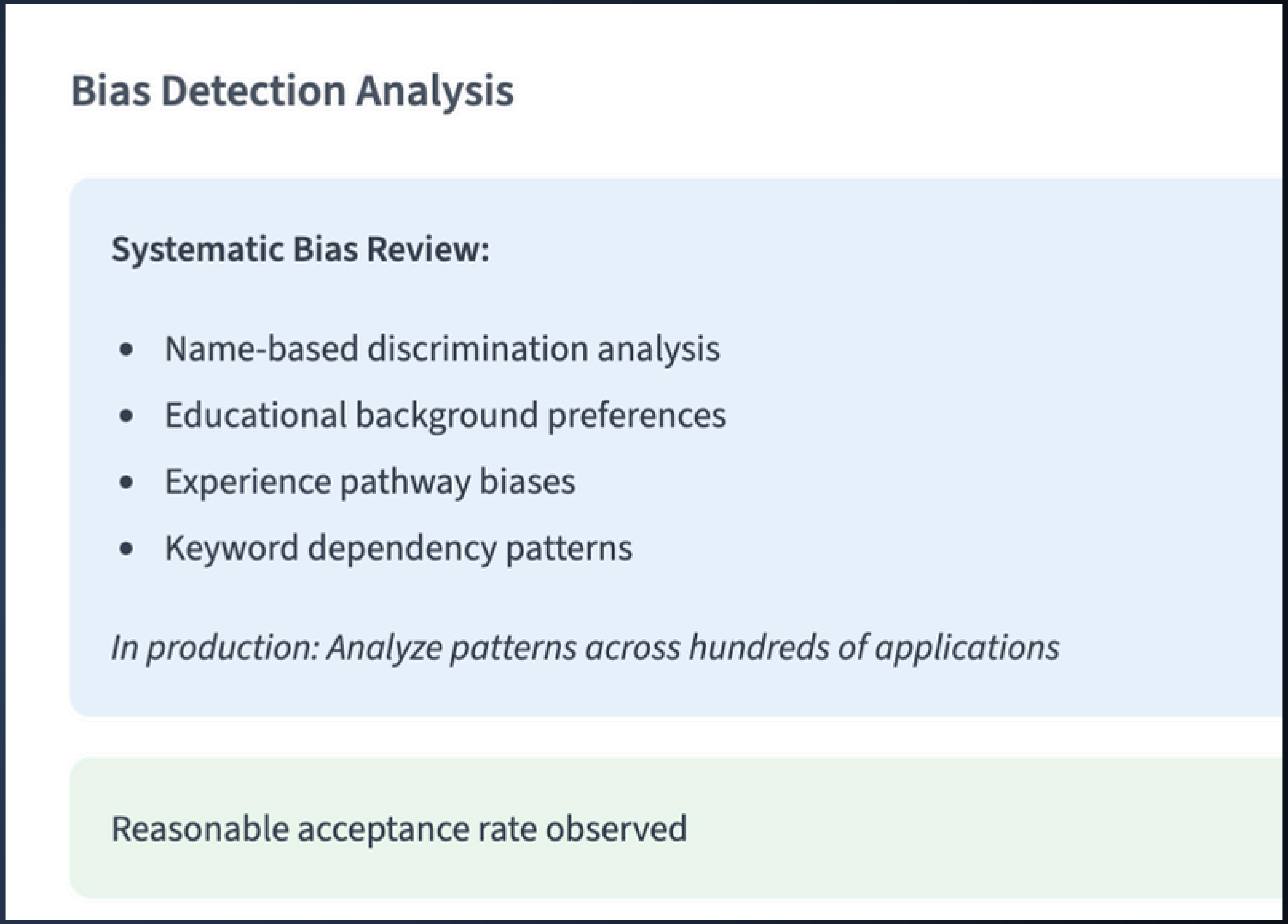
⚠ High score variance - may indicate inconsistent criteria application

Standard Deviation = √(Σ(xi - μ)² / n)
High Variance = Inconsistent scoring patterns
Low Variance = Consistent criteria application
Zero Variance = Potential system malfunction

Bias Detection: Demographic Fairness

Using statistical methods to identify if AI treats different demographic groups unfairly.

- Systematic bias review covering name-based, educational, experience, and keyword biases
- Analysis framework for identifying discrimination patterns
- "Reasonable acceptance rate observed" indicates no major bias detected



Bias Detection Analysis

Systematic Bias Review:

- Name-based discrimination analysis
- Educational background preferences
- Experience pathway biases
- Keyword dependency patterns

In production: Analyze patterns across hundreds of applications

Reasonable acceptance rate observed

Group Acceptance Rate = Accepted in Group / Total in Group
Bias Difference = |Group A Rate - Group B Rate|
Statistical Significance = Chi-square test ($p < 0.05$)
Practical Significance = Difference > 5% threshold

Conclusion

1 Philosophical Foundations Enable Practical Implementation

2 The quantitative rigor that data science methods provide to AI validation

3 The essential role of human expertise in the diagnostic process

4 The business case for systematic diagnosis through cost-benefit analysis

5 The operational feasibility of implementing diagnostic methodology at scale

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13. https://doi.org/10.1145/3290605.3300233

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229. https://doi.org/10.1145/3287560.3287596

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Thank You