The Limits of the Algorithm-A Case Study in HCAI

Hunjun Shin

1 Identify the Flaws

1.1 Flaw 1: Inherently Biased Training Data

The foundation of the "Optimal Hire" AI is a decade of Innovate Inc.'s hiring data, a dataset that reveals a stark historical preference for "men from a handful of elite universities." Rather than being an objective measure of merit, this training data is a codification of the company's own past biases in hiring and promotion.

Peter Norvig highlights the issue of proxy data and bias in data collection, as seen with the COMPAS recidivism prediction system, where there was "no access to ground truth" on who commits crime, only proxy data of arrests and convictions, which can be biased. Similarly, Sasha Luccioni demonstrates how AI models can encode and perpetuate stereotypes and racism/sexism due to biased training data, leading to systems that disproportionately perform worse for certain demographics, such as facial recognition systems for women of color. The example of the UK school grading AI penalizing poorer students due to inadvertent bias in its training is also relevant.

1.2 Flaw 2: The "Black Box" Problem : Lack of Transparency

The system's "black box" nature directly contravenes core HCAI principles of transparency. Because the logic behind its "Success Score" is "incredibly complex and not fully understood by its creators," users are left unable to comprehend "what the AI system is doing and why," creating a critical deficit in accountability.

HCAI principles emphasize transparency, meaning users should understand "what the AI system is doing and why". Sasha Luccioni refers to these systems as "black boxes" and states that "even their creators can't say exactly why they work the way they do". Ben Shneiderman's "Prometheus Principles" advocate for "informative feedback to acknowledge each user action" and progress indicators to show status", which are absent here. IBM's AI guidelines warn that "imperceptible AI is not ethical AI"

1.3 Flaw 3: Narrow and Flawed Objective Function

The primary goal is to predict "top performers" based on historical data, and the company "celebrates the tool for increasing 'hiring efficiency' by 400%". This objective function is

too narrow, focusing on internal metrics (efficiency, historical "top performers") that are themselves products of past biases, rather than broader societal values or true long-term company success, such as diversity or innovation potential.

Peter Norvig stresses that defining the objective is "the hard part" in machine learning, and maximizing an internal measure like "accuracy" doesn't necessarily mean success. He argues that fairness is a "societal values question," not a mathematical or software one, and that focusing on internal metrics like "well calibrated" scores (as with COMPAS) can obscure real-world harm. The "Human Imperative in AI" source states that "what is technically optimal may be socially unjust". Goodhart's Law suggests that optimizing too much for a metric can lead to ignoring the real goal.

2 Explain the Problems

2.1 Impact on Individual Applicants

- **Systematic Exclusion and Denied Opportunities:** Applicants falling outside the narrow historical profile—such as women or graduates from non-elite universities—are systematically penalized with lower "Success Scores" and rejected before human review. The harm caused by this automated gatekeeping is analogous to the well-documented failures of the COMPAS system, which inflicted false positives on Black defendants at twice the rate of their white counterparts.

- **Perpetuation of Discrimination:** The AI would effectively codify and amplify existing gender and socioeconomic biases, denying fair opportunities based on demographics rather than individual merit, leading to what Sasha Luccioni describes as AI discriminating against "entire communities"

2.2 Impact on Innovate Inc.

- **Homogenization and Stifled Innovation:** The system's preference for a narrow demographic will inevitably lead to a homogenous workforce. This lack of diversity stifles the very creativity and innovation Innovate Inc. presumably seeks, as it misses out on the varied perspectives critical for identifying new market opportunities and solving complex problems.

- **Legal and Reputational Risk:** Implementing a system that systemically discriminates based on gender or other protected characteristics exposes Innovate Inc. to class-action lawsuits, similar to those faced by Apple for false advertising of AI features, or legal challenges for bias, like those filed against AI companies for copyright infringement. This also harms the company's reputation, making it less attractive to diverse talent and customers.

2.3 Impact on Society at Large

- **Exacerbated Inequality:** If similar "Optimal Hire" systems become widespread, they could reinforce and deepen existing societal inequalities in the job market, limiting social mobility and concentrating economic power among already privileged groups. This echoes concerns about AI's impact on mass incarceration and bias in the justice system, as discussed regarding the COMPAS system.

- **Erosion of Trust in AI:** The deployment of biased "black box" systems erodes public trust in AI technologies, making it harder to gain acceptance for beneficial AI applications. James Landay stresses the importance of designing AI systems "at the user community and Society level" to ensure a positive societal impact.

3 Why Code Isn't Enough

3.1 Why Technical Fixes are Insufficient:

- **Fundamental Data Bias:** Simply "adding more data" to "Optimal Hire" would likely perpetuate or even amplify existing biases if that data still reflects historical discriminatory patterns. As Peter Norvig noted with the COMPAS system, the problem wasn't the data quantity but the bias in the proxy data and the "failure to bring in multidisciplinary teams up front" to challenge underlying assumptions.

- **The Mathematical Trap of Fairness Metrics:** The belief that the algorithm can simply be "fixed" ignores a fundamental constraint: as research on systems like COMPAS has shown, it is "mathematically impossible" to satisfy all definitions of fairness simultaneously. Consequently, any technical attempt to "improve" the algorithm by optimizing for one fairness metric will likely degrade its performance on another, trading one form of bias for another.

- **A Societal, Not Technical, Problem:** Ultimately, the system's core flaw is not technical but philosophical. It attempts to solve a "societal values question"—what constitutes a fair hiring process—with a purely computational tool. An algorithm can optimize for predefined metrics, but it possesses no capacity to understand or navigate the nuanced, evolving human definitions of "good" or "fair," a task that requires human judgment.

3.2 Necessity of Non-Algorithmic Solutions and a Broader Human-Centered AI Framework:

3.2.1 Human Oversight and Intervention in the Hiring Process:

- **HCAI Principle:** HCAI emphasizes systems that augment rather than replace humans, ensuring "appropriate control to the humans". Automation should handle routine tasks,

but humans must retain meaningful oversight and decision-making power, especially for complex and consequential applications like hiring.

- **Application to Optimal Hire:** Instead of automatically rejecting candidates who score below 85, the system should serve as a decision support tool for HR professionals. It can analyze resumes to provide a concise summary of how a candidate's background aligns with the role and where there might be a mismatch. This allows for a more nuanced review, flagging promising candidates and highlighting diverse applications that might be overlooked by a purely automated filter. This approach allows human judgment to override algorithmic recommendations, much like how a patient-controlled analgesia (PCA) device empowers a patient to manage their own medication while automation prevents overdosing.

3.2.2 Diverse Development Teams and Inclusive Design Principles:

- **HCAI Principle:** HCAI systems must be transparent, allowing users to understand "what the AI system is doing and why".

- **Application to Optimal Hire:** The "black box" nature must be addressed. The system should provide explanations for its "Success Score", indicating which factors (e.g., specific skills, experiences, academic background) contributed to a high or low score, rather than just giving a number. This allows HR to understand potential biases and challenge flawed reasoning, fulfilling the need for "explainability".

3.2.3 Ongoing Auditing and Ethical Review of AI Systems:

- **HCAI Principle:** HCAI systems must be accountable and ensure fairness, requiring "continuous review of data quality & bias testing to cope with shifting contexts of use" and "independent oversight structures".

- **Application to Optimal Hire:** Innovate Inc. needs to implement regular, independent audits of "Optimal Hire" to continuously monitor its impact on hiring diversity, evaluate for disparate outcomes, and assess fairness metrics for different demographic groups, beyond just "hiring efficiency." This oversight, perhaps by an internal review board or external agencies, would ensure that the system's performance aligns with ethical and societal values, not just technical ones.