INVESTIGATING AI SYSTEMS IN REAL-WORLD SCENARIOS: A RESPONSIBLE AI PERSPECTIVE

AI SAFARI, PLP ACADEMY

JULY 29, 2025

PLP ACADEMY

RESPONSIBLE AI INSPECTOR REPORT

TITLE: INVESTIGATING AI SYSTEMS IN REAL-WORLD SCENARIOS: A RESPONSIBLE AI PERSPECTIVE

AUTHOR: SAMUEL EMMANUEL KIMARO

ROLE: RESPONSIBLE AI INSPECTOR

PLP ACADEMY

CASE 1: RESUME ROULETTE – BIAS IN HIRING ALGORITHMS

What's Happening:

In a growing effort to reduce hiring costs and increase efficiency, companies are turning to AI-powered applicant tracking systems (ATS). These systems automate resume screening, ranking candidates using predefined criteria based on past successful hires.

In this scenario, the AI is filtering thousands of resumes to identify top applicants. However, a disturbing pattern emerges: candidates with career gaps especially women returning to the workforce are disproportionately rejected. These gaps are automatically treated as negative signals by the system.

What's Problematic

1. Inherited Bias from Historical Data

The core issue stems from the **training data**: if the organization's past hiring decisions were biased favoring men, penalizing career breaks, or overlooking diverse backgrounds the AI will replicate those same biases at scale. This is known as **historical bias**, and it turns prejudice into a systematic filter.

2. Proxy Discrimination

Career gaps may inadvertently become **proxy variables** for gender, disability, or caregiving responsibilities. While the AI might not explicitly use gender as a variable, career interruptions often correlate with gendered life experiences, leading to **indirect discrimination**.

3. Lack of Explainability

The algorithm acts as a black box. Candidates receive no insight into why they were rejected, and hiring managers often trust the system blindly. This violates **transparency** and undermines **accountability** in HR practices.

4. Exclusion of Human Judgment

By over-automating initial screening, the process removes essential human context—ignoring nuanced explanations for employment gaps such as maternity leave, illness, or upskilling periods. The result is a shallow and mechanical judgment of human potential.

Suggested Improvement

- **Bias Auditing & Debiasing Data:** The AI model should undergo routine fairness audits, and the training data should be **rebalanced** to remove discriminatory patterns. Synthetic or augmented data can help simulate underrepresented groups.
- **Context-Aware Design:** Introduce a feature allowing candidates to provide **contextual explanations** for career gaps, which are then reviewed by a human recruiter.
- **Explainability Tools:** Implement AI interpretability tools (e.g., LIME or SHAP) to generate transparent feedback such as "Resume filtered due to lack of keyword match" or "Shortlisted based on leadership experience."
- **Human-in-the-Loop (HITL):** Final shortlisting should involve human review, particularly for non-linear career paths or high-potential candidates not flagged by the AI.

CASE 2: THE EYE SPY SCANDAL – FLAWED SCHOOL PROCTORING AI

What's Happening

In response to the shift to remote learning, many educational institutions have adopted AI-based proctoring tools to detect cheating during online exams. These tools typically use facial recognition and eye movement tracking to monitor students' behavior.

In this case, the system automatically flags students as suspicious if they look away from the screen frequently, their face is partially obscured, or their body language seems "nervous." The flagged sessions are then reviewed for potential misconduct.

What's Problematic

1. Discrimination Against Neurodivergent Students

Students with autism, ADHD, anxiety disorders, or tics may not behave in line with neurotypical expectations. For example, they may fidget, avoid eye contact, or look away to process thoughts all of which can be flagged as suspicious behavior. The system is inadvertently **penalizing cognitive diversity**.

2. Cultural and Environmental Insensitivity

The AI may misinterpret gestures across different cultures or socio-economic contexts. For example, students from lower-income backgrounds may have shared spaces, causing movement or audio triggers unrelated to cheating. These variances are rarely accounted for.

3. Privacy and Surveillance Risks

Recording students' faces, environments, and movements in their private spaces raises **serious privacy issues**. If footage is stored insecurely, or used beyond the exam context, it could lead to **data misuse** or psychological harm.

4. Lack of Appeals and Transparency

Students often have no way to appeal a flag, or even understand what caused it. This undermines trust in academic integrity processes and can affect student outcomes unfairly.

Suggested Improvement

- **Inclusive AI Design:** Involve neurodivergent students, accessibility experts, and educators in the **design and testing** of proctoring systems. Build models that understand a **range of normal behaviors**, not just statistical averages.
- **Multimodal Cheating Detection:** Use a combination of indicators (e.g., browser behavior, keyboard activity) rather than relying solely on eye movement or facial expressions. This reduces false positives while maintaining exam integrity.
- **Privacy-First Architecture:** Ensure **video footage is encrypted, access-controlled, and deleted** after a defined period. Inform students how data is handled, and let them consent to its collection.
- **Appeal and Review Processes:** Any AI flag should be reviewed by a human, and students must be given a clear appeals process, including a breakdown of what behavior triggered the alert.

Conclusion: AI Is Not Neutral Responsibility Is Key

These two cases demonstrate that while AI systems promise efficiency and objectivity, they are not inherently fair or inclusive. Left unchecked, they can reinforce existing societal inequalities and erode trust in critical systems like hiring and education.

To build **responsible AI**, organizations must integrate ethics, transparency, and human oversight at every stage from data collection to deployment. Bias mitigation isn't just a technical task; it's a social responsibility.