Proskauer

AI at Work: Black Box Issues

The Proskauer Brief on **January 10, 2025**

In part three of our series on potential pitfalls in the use of artificial intelligence (or AI) when it comes to employment decisions, partner Guy Brenner and senior counsel Jonathan Slowik dive into the concept of "black box" systems—AI tools whose internal decision-making processes are not transparent. The internal workings of such systems may not be well understood, even by the developers who create them. We explore the challenges this poses for employers seeking to ensure that their use of AI in employment decisions does not inadvertently introduce bias into the process. Be sure to tune in for a closer look at the complexities of this conundrum and what it means for employers.

Guy Brenner: Welcome again to The Proskauer Brief: Hot Topics in Labor and Employment Law. I'm Guy Brenner, a partner in Proskauer's Employment Litigation & Counseling group, based in Washington, D.C. I'm joined by my colleague, Jonathan Slowik, a special employment law counsel in the practice group, based in Los Angeles. This is part three of a multi-part series on potential pitfalls in the use of artificial intelligence (or AI) when it comes to employment decisions, such as hiring and promotions. Jonathan, thank you for joining me today.

Jonathan Slowik: It's great to be here, Guy.

Guy Brenner: If you haven't heard the earlier installments of this series, we encourage you to go back and listen to them. In part one, we go through what we hope is some useful background about what AI solutions are out there for employers and HR departments, including tools like résumé scanners, chatbots, interviewing platforms, social media tools, job fit tests, and performance reviews. In part two, we discuss how issues with training data can lead to biased or otherwise problematic outputs. Today's episode is about what we call "black box" issues. Jonathan, what do we mean when we refer to an AI being a "black box"?

Jonathan Slowik: So a “black box” system draws conclusions without providing any explanations as to how those conclusions were reached. This is also sometimes referred to as “model opacity”—we can’t see what it’s doing under the hood. In fact, the internal workings of a black box system might not be clear even to the developer that built it. For example, the AI developer Anthropic has spent significant resources on research to better understand the workings of its large language model, Claude, and it was considered big news in the industry when they published a paper in May 2024 announcing some preliminary findings.

Guy Brenner: That’s pretty sobering – even the developers don’t know exactly how it works. So, putting aside the larger questions prompted by this fact, why should an employer care if an AI is a black box?

Jonathan Slowik: Well, if it’s difficult to understand why a system is doing what it’s doing, it can also be difficult to evaluate whether what it’s doing is unbiased or using inappropriate criteria. There was another interesting study that also came out this spring that examined what the researchers call the overt and covert biases of large language models, or LLMs, like Claude, or the many chat bots that that many of us have come to rely on for all kinds of things. There was another interesting study that also came out this spring that examined what the researchers called the overt and covert biases of large language models, or LLMs, like the chat bots that a lot of us have come to rely on for all kinds of things. LLMs are trained on our own speech and writing, and the most advanced versions of trained on, for example, a significant portion of the internet. And so this vast corpus of data and writing naturally includes some ugly stereotypes. And perhaps unsurprisingly then, early versions of this technology exhibited that bias in its responses. That’s obviously a huge problem. No one wants to be putting out a racist chat bot, and these are the public developers solve for this problem primarily through what’s called human feedback training.

This is a process kind of like sites like Reddit where people upvote or down vote things people say. In this process, a human being or a large number of human beings review a large number of outputs from the model and essentially good outputs and downvote bad outputs. This feedback trains the AI not to give racist outputs and to give more accurate outputs, but the researchers found that even advanced alums were exhibiting implicit bias against racial minorities. So over time, through this human feedback training, the LLMs had gotten very good at mimicking real people and for the most part, most of us don't say racist things out loud, thank goodness. But real people may harbor biases underneath the surface, consciously or not.

Guy Brenner: Jonathan, as I listen to you, the implications of this for employers are pretty evident and profound. If an employer cannot know for sure that under the hood, the AI is operating in a way that is unbiased, that means that biases might only become apparent over time and after the fact. In these contexts, bias audits could be critical to mitigating the risk of algorithmic discrimination.

Jonathan Slowik: That's exactly right. And that's one reason why lawmakers and regulators, as they grapple with the issues created by these new AI tools, have been especially focused on bias audits as they begin to craft implement specific regulations for AI.

Guy Brenner: So what are we going to discuss in the next episode, Jonathan?

Jonathan Slowik: In the next episode, we'll explore mismatches between a platform's design and its end use as we'll discuss. Even a purportedly unbiased system can produce biased results if it's used for an unintended purpose.

Guy Brenner: Well, thanks Jonathan. And to those listening and joining us on The Proskauer Brief today. As developments warrant, we'll be recording new podcasts to help you stay on top of this fascinating and ever- changing area of the law and technology. Also, please be sure to follow us on Apple Podcasts, Google Podcasts, and Spotify so you can stay on top of the latest hot topics in labor and employment law.

Related Professionals

- **Guy Brenner**
  Partner

- **Jonathan P. Slowik**
  Senior Counsel