Proskauer

AI at Work: Training Data Issues

The Proskauer Brief on **December 3, 2024**

In part two of our insightful artificial intelligence series, partner Guy Brenner, who leads Proskauer's D.C. Labor & Employment practice and is co-head of the Counseling, Training & Pay Equity Group, and Jonathan Slowik, senior counsel, Labor & Employment, in the firm's Los Angeles office, explore the critical issue of AI training data in employment decisions. We discuss how issues with training data can create risk with respect to employment discrimination laws, even when AI systems are not explicitly programmed to consider protected characteristics. We also highlight the potential for inaccuracies in AI models due to insufficient or unrepresentative training data. So be sure to tune in as the legal implications of these issues can be of importance to employers when understanding potential biases in AI systems.

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 and Counseling Group based in Washington, DC. And I'm joined by my colleague, Jonathan Slowik, senior counsel in the Labor and Employment Practice Group based in Los Angeles. This is part two of our multi-part series detailing what employers need to know about the use of artificial intelligence, or AI. When it comes to employment decisions such as hiring and promotions, Jonathan, thank you for joining me again today.

Jonathan Slowik: I'm excited to be here, guys.

Guy Brenner: If you haven't heard part one of this series, we encourage you to go back to listen to it, because we go through what we hope is some useful background about what AI is and what kind of solutions it provides to employers and HR departments. But, as we promised at the end of that episode, AI also presents some challenges and pitfalls for employers, and this episode is the first of a series that's going to go through those potential pitfalls. Today, we are going to talk about training data and the issues that can arise from that. So, Jonathan, can you begin by giving us a quick primer on what we mean by training data?

Jonathan Slowik: Certainly. So, to understand what training data is, it's an important first to understand that most AI systems involve some aspect of machine learning. By machine learning, I'm referring to a process by which the AI takes in a very large quantity of data, analyzes that data for patterns and connections, and then applies the lessons that it draws from that analysis to making predictions or drawing conclusions in a new context. This process is called training the AI, and the data that it analyzes is called the training data.

Guy Brenner: That's a really helpful synopsis, Jonathan, so. What do employers need to be aware of when it comes to machine learning? I mean, it may sound to some like a perfectly sensible approach that leads to, you know, appropriate results better than any human could do.

Jonathan Slowik: That's certainly the promise of it, and why a lot of employers are very excited about this new technology. There are some potential problems. One is that AI systems can be trained on data that's not representative, that incorporates historical biases or correlates with protected characteristics. I'll make this a little more concrete. Suppose that an AI model is being used to assist in hiring decisions, and this model's been trained on data about your company's workforce over a period of years. The AI analyzing this data and drawing lessons about who has been successful at your company in the past is using those lessons to try to predict who has the greatest likelihood of meeting success in the future. This seems like a perfectly reasonable way to go about it. It's obviously job related. The type of person who has been successful in the past will likely also be successful in the future. Do you see any problems with that approach, Guy?

Guy Brenner: I think the example itself illustrates what makes AI so attractive. If AI can predict and provide clear guidance about those who are most likely to succeed at a company based on the company's own data, its own track record, that would help the employer narrow its recruiting efforts, saving a ton of time and money, and improving results. So the appeal is obvious, right? But the issue is the data being used and whether it itself reflects biases. For example, what if your workforce historically has been predominantly male, or predominantly white, or people without primary caregiving responsibilities, et cetera, right? You could see how the data being used could perpetuate biases that were latent in the organization.

Jonathan Slowik: Right. So, in any of these cases, the AI would be making hiring decisions based on a biased training data set, and that may reflect a historical failure of the organization to promote a diverse range of candidates. It may reflect just, you know, trends in society over time. But if the training data set is predominantly representing people from certain demographics or failing to represent people from certain demographics, you might get results that reflect that bias.

Guy Brenner: So Jonathan, it seems to me that somebody might out there might be thinking, “Well, the easy thing to do is just make sure that the training data doesn’t include demographic information,” right? Just, you know, cleanse it of any, you know, race, sex, age — any protected characteristic data whatsoever. What’s, what’s wrong with that thing?

Jonathan Slowik: Well, that’s certainly a step in the right direction, but it might not solve the problem completely. But remember, there could be characteristics about your organization’s past and present employees that correlate with protected characteristics. So even if the AI model knows to ignore protected characteristics or is even kept in the dark about that information completely because you’ve removed it from the training data set, if it’s fed a training data set that, for example, is predominantly white and male, it could end up recommending a pool of job candidates that look like your prior and current workforce. In other words, being predominantly white and male, even if the AI model is not thinking about protecting characteristics at all.

Guy Brenner: I’ve been hearing a lot of buzz about something called algorithmic discrimination. Statutes such as those in New York City and Colorado — the one in Colorado isn’t in effect yet — both require periodic bias audits to root out and prevent algorithmic discrimination. Was what you’re talking about an example of that?

Jonathan Slowik: Yes, exactly

Guy Brenner: Okay, but I’ll point out that even if our listeners are not in a jurisdiction that has laws concerning algorithmic discrimination or things of the like, that doesn’t mean that they don’t have to worry about this. Existing laws already outlaw discrimination based on protected characteristic, whether it’s being done by a human being or machine. You can’t protect yourself by saying, “The AI made me do it.” Jonathan, what are some of the other ways unrepresentative data training can make an AI model go awry?

Jonathan Slowik: So, I think an underrated potential problem is that AI models can struggle with certain tasks if they don't have sufficient volume of training data. To illustrate what I mean, there's a, a simple and humorous example that was highlighted last year on the New York Times audio podcast Hard Fork. In this example, they gave the same prompts to two different generations of the engine that powers ChatGPT. On the one hand was GPT-4, which at the time was the most advanced version of this technology and, was the version available to paying subscribers of ChatGPT, and on the other hand there was GPT-2, which was a version that at the time was about three years old and had never been released to the public, and of an early prototype. So GPT-4 has some more bells and whistles than GPT-2. For example, you can use it to generate images and things like that. But perhaps the biggest difference between the two engine is just that GPT-4 was trained on much more data than GPT-2. According to some estimates, maybe 1000 times more data. So with that background, GPT-4 and GPT-2 were given the same prompts, and the prompt was this: Finish the Valentine's Day card that begins, "Roses are red, violets are blue." Guy, do you want to read how GPT-4 finished the prompts?

Guy Brenner: Sure. "Roses are red, violets are blue. Our love is timeless. Our bond is true." Not bad.

Jonathan Slowik: Not, not bad. It's pretty — you can see that on a hallmark card, maybe. It's maybe a little cheesy, but that may have more to do with the prompt than anything. So that's GPT-4. Now, Guy, do you want to do the honors for GPT-2? And remember, this is a, an engine that was trained on a much smaller data set.

Guy Brenner: Here you go. "Roses are red, violets are blue. My girlfriend is dead." Not quite as good.

Jonathan Slowik: Not quite. Maybe so bad it's good? Maybe less hallmark material, more novelty shop material. So, this is obviously a silly example, and if an AI ever gave, you know, in an HR context, gave, gave kind of a similarly absurd output, it shouldn't be and couldn't be taken seriously, and so, if you got something this ridiculous, you would ignore it. There's a lack of training data that's making an AI misfire in more nuanced ways, you can see how that would be more problematic. To give you an example of what I mean by that, there's a wealth of academic research indicating that facial recognition software can be less accurate when it comes to women and people of color. And one possible reason for that is that some of these platforms may have been trained on data sets that consist disproportionately of the faces of white males. So, if the same is true about a video interviewing platform, for example, that analyzes facial expressions, tone of voice, speech patterns of candidates, that tool might also be less accurate when it comes to women or people of color, and that bias might be harder to detect, and maybe only over time and, and in hindsight.

Guy Brenner: It's a lot like if someone comes to you with an answer and you don't probe how they came to the answer and just assumed it was correct, you could later learn it was flawed in some way. We're used to thinking that a machine gets things right, but with AI, you can't make those assumptions. It's not a calculator, right?

Jonathan Slowik: Exactly.

Guy Brenner: There's some thinking, so to speak, involved. So now that we've covered training data issues, what are we going to discuss in the next episode?

Jonathan Slowik: In the next episode, we'll discuss model opacity issues. So a lot of these advanced AI platforms are what are often referred to as black boxes, meaning that it can be difficult to understand their internal workings. As we'll discuss in the next episode, the fact that an AI system is a black box can have profound implications that employers using such a tool in hiring, promotions, or other employment decisions should be aware of.

Guy Brenner: Well, I'm looking forward to that discussion. And to those listening, thank you for 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, YouTube Music, and Spotify so you can stay on top of the latest hot topics in labor and employment law.

View original.

Related Professionals

- **Guy Brenner**
  Partner
- **Jonathan P. Slowik**
  Senior Counsel