SEBRIS | BUSTO | JAMES

425-454-4233
sbj.law

Employment Law Note

APRIL 2023

Reducing Harmful Bias in Hiring Algorithms

By Jennifer Hohnstein, jhohnstein@sbj.law

Companies are increasingly using algorithms in their candidate selection process either directly, or indirectly when they employ a vendor that uses algorithms. The use of algorithms is intended to remove unintended selection bias from the hiring process. The algorithmic tools also have the potential for increasing efficiency and saving costs. Although well-intentioned, reliance on algorithms or artificial intelligence ("AI") can have unintended consequences because any algorithm is only as good as the information used to train the algorithm. This means any biases and/or skewed trends in the underlying data set can be mimicked or in some cases amplified by the algorithm. Companies must be wary because they can be legally liable for discrimination regardless of whether it originates from a biased algorithm or a person.

One comical example of algorithmic biases occurred when an audit of a hiring algorithm revealed that it had selected two factors as the most indicative of job performance: whether the candidate's name was Jared and whether they played high school lacrosse. In a less humorous example, Amazon machine learning specialists discovered that their new recruiting engine did not like women. Essentially, when training the search engine to recognize top candidates, the developers provided resumes submitted to the company over a ten-year period. Because most of the resumes came from men, the algorithm taught itself that male candidates were preferable and downgraded resumes that included the word "women's" (as in "women's sports") along with graduates of two all-women's colleges. Amazon ended the project.

The legal risks to companies, though, are not theoretical. On February 21, 2023, a class action was filed against Workday, Inc., in the Northern District Court of California alleging that the company had engaged in race, age, and disability discrimination through its use of applicant-screening tools that use biased AI algorithms. Similarly, in the beginning of 2023, the Equal Employment Opportunity Commission ("EEOC") announced that it would increase enforcement efforts aimed at discrimination resulting from the use of AI-assisted employment-related decision tools. This follows the EEOC's previous guidance issued in May 2022 for employers to avoid discrimination while using AI tools.

Given the consequences of using an algorithm with a harmful bias, and the prevalence of these algorithms in the hiring process, employers should be familiar with the characteristics of trustworthy AI. One tool comes from the National Institute of Standards and Technology ("NIST"), which released its AI Risk Management Framework ("RMF") on January 26, 2023. NIST started work on its RMF in 2021 following a Congressional mandate set forth in the National Artificial Intelligence Initiative Act of 2020. The RMF is designed to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time.

NIST sets out seven trustworthiness characteristics for AI:

1. **Valid and Reliable.** Validation is the confirmation through objective evidence that the requirements for a specific intended use or application have been fulfilled and reliability is the ability of an item to perform as required, without failure, for a given time interval, under given conditions. Both validity and reliability can be assessed through ongoing monitoring and testing to confirm that a system is performing as intended.
2. **Safe.** AI systems should not lead to a state in which human life, health, property or the environment is endangered. While different types of safety risks require different approaches, on a practical level, AI safety includes having the ability to test, monitor, and

(if needed) shut down systems that have deviated from their intended or expected functionality.

3. **Secure and Resilient.** These characteristics include both the ability to avoid, protect against, respond to, or recover from attacks as well as the ability to return to normal function after an unexpected adverse event or unexpected changes in their environment or use – or if they can maintain their functions and structure in the face of internal and external change and degrade safely and gracefully when this is necessary.
4. **Accountable and Transparent.** Trustworthy AI must be accountable while accountability presupposes transparency. While a transparent system is not necessarily an accurate, privacy-enhanced, secure, or fair system, it can be difficult to determine whether an opaque system possesses such characteristics.
5. **Explainable and Interpretable.** Whereas transparency can answer the question of "what happened" in the system, explainability can answer the question of "how" a decision was made and interpretability can answer "why" that decision was made. Together, these factors assist those operating or overseeing an AI system, as well as users of the AI system, in gaining deeper insights into a system's functionality and trustworthiness, including its outputs.
6. **Privacy-Enhanced.** Privacy values such as anonymity, confidentiality, and control generally should guide choices for AI system design, development, and deployment. Privacy-enhancing technologies ("PETs"), as well as data-minimizing methods such as de-identification and aggregation can support the design for privacy-enhanced AI systems.
7. **Fair - with Harmful Bias Managed.** Fairness in AI includes concerns for equality and equity by addressing issues such as harmful bias and discrimination. Three major categories of AI bias to be considered and managed are: systemic, computational and statistical, and human cognitive. Each of these can occur in the absence of prejudice, partiality, or discriminatory intent. Bias can become ingrained in the automated systems that help make decisions about our lives. While bias is not always negative, AI systems can increase the speed and scale of biases and perpetuate and amplify harm to individuals, groups, communities, organizations, and society.

The RMF recognizes that there is tension between the characteristics and when weighing the seven characteristics there can be tradeoffs and different risk prioritization depending on the AI's intended use. For example, AI systems interacting directly with job applicants and that handle sensitive or personal data will likely prioritize privacy and fairness considerations. However, there is no fixed approach and the harm/cost benefit tradeoffs will continue to be developed and debated. There is also a lack of consensus on robust and verifiable measurement methods.

Ultimately, the RMF is meant to be a living document and provides guidance as to how companies can govern, map, measure and manage AI. While use of the RMF is non-binding and voluntary, it will influence best practices for the use of trustworthy AI. Moreover, while using the RMF does not create a safe harbor, a company's use of the RMF could provide positive evidence that it has worked in good faith to mitigate potential harms should it discover that an algorithm has unintended harmful consequences.

Advice for Employers

Employers using AI to recruit for or screen candidates may want to check out the RMF at https://doi.org/10.6028/NIST.AI.100-1. Employers using a third-party vendor should do their due diligence and inquire as to the use of AI in the screening process and whether the algorithms used have been scrutinized for bias.

Anyone who has questions on this topic or would like to learn more is encouraged to contact Jennifer Hohnstein at jhohnstein@sbj.law.

For more information about this month's Employment Law Note
contact us at **425-454-4233**