Volume 8, No. 4 | July–August 2025

RAIL

The Journal of Robotics, Artificial Intelligence & Law

fastcase FULL COURT PRESS

RAIL
The Journal of Robotics, Artificial Intelligence & Law
Volume 8, No. 4 | July–August 2025

THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW (ISSN 2575-5633 (print) /ISSN 2575-5617 (online) at $495.00 annually is published six times per year by Full Court Press, a Fastcase, Inc., imprint. For customer support, please contact Fastcase, Inc., 729 15th Street, NW, Suite 500, Washington, D.C. 20005, 202.999.4777 (phone), or email customer service at support@fastcase.com.

Publishing Staff
Publisher: Leanne Battle
Production Editor: Sharon D. Ray
Cover Art Design: Juan Bustamante

Cite this publication as:

The Journal of Robotics, Artificial Intelligence & Law (Fastcase)

A Full Court Press, Fastcase, Inc., Publication

Editorial Office

729 15th Street, NW, Suite 500, Washington, D.C. 20005
https://www.fastcase.com/

POSTMASTER: Send address changes to THE JOURNAL OF ROBOTICS, ARTIFICIAL INTELLIGENCE & LAW, 729 15th Street, NW, Suite 500, Washington, D.C. 20005.

Articles and Submissions

Direct editorial inquiries and send material for publication to:

Steven A. Meyerowitz, Editor-in-Chief, Meyerowitz Communications Inc., 26910 Grand Central Parkway, #18R, Floral Park, NY 11005, smeyerowitz@meyerowitzcommunications.com, 631.291.5541.

Material for publication is welcomed—articles, decisions, or other items of interest to attorneys and law firms, in-house counsel, corporate compliance officers, government agencies and their counsel, senior business executives, scientists, engineers, and anyone interested in the law governing artificial intelligence and robotics. This publication is designed to be accurate and authoritative, but neither the publisher nor the authors are rendering legal, accounting, or other professional services in this publication. If legal or other expert advice is desired, retain the services of an appropriate professional. The articles and columns reflect only the present considerations and views of the authors and do not necessarily reflect those of the firms or organizations with which they are affiliated, any of the former or present clients of the authors or their firms or organizations, or the editors or publisher.

QUESTIONS ABOUT THIS PUBLICATION?

For questions about the Editorial Content appearing in these volumes or reprint permission, please contact:

Leanne Battle, Publisher, Full Court Press at leanne.battle@vlex.com or at 202.999.4777

For questions or Sales and Customer Service:

Customer Service
Available 8 a.m.–8 p.m. Eastern Time
866.773.2782 (phone)
support@fastcase.com (email)

Sales
202.999.4777 (phone)
sales@fastcase.com (email)

ISSN 2575-5633 (print)
ISSN 2575-5617 (online)

Five Human Best Practices to Mitigate the Risk of AI Hiring Tool Noncompliance with Antidiscrimination Statutes

Justin R. Donoho*

In this article, the author identifies human best practices to mitigate the risk of companies' artificial intelligence hiring tools violating various statutes.

While artificial intelligence (AI) hiring tools can improve efficiencies in human resource functions, such as candidate sourcing, resume screening, interviewing, and background checks, AI has not replaced the need for humans to ensure that AI-assisted human resources (HR) practices comply with a wide range of antidiscrimination laws such as Title VII of the Civil Rights Act of 1964 (Title VII), the Americans with Disabilities Act (ADA), the Age Discrimination in Employment Act (ADEA), the sections of Colorado's AI Act setting forth developers' and deployers' "duty to avoid algorithmic discrimination" (CAI), New York City's law regarding the use of automated employment decision tools (NYC's AI Law), the Illinois AI Video Act (IAIVA), and the 2024 amendment to the Illinois Human Rights Act to regulate the use of AI (IHRA).

This article identifies human best practices to mitigate the risk of companies' AI hiring tools violating the foregoing statutes, according to the statutes and the additional sources cited in note 1 to this article.[1]

Human Best Practice #1—Implement an AI Governance Structure

A "guiding principle" for managing AI hiring tool risk is to put in place a governance structure that enables humans to effectively challenge the model by selecting, auditing, and tuning it.[2] Effective challenge requires competence, independence, influence,

Robotics, Artificial Intelligence & Law / July–August 2025, Vol. 8, No. 4, pp. 259–267.

ISSN 2575-5633 (print) /ISSN 2575-5617 (online).

incentives, and, as necessary, external consultants, which should all be part of an overall risk management policy and program, as follows:

- *Competence*. Employers should "hire or train employees who have sufficient knowledge and experience with HR management and AI development."[3] Employers should also employ or retain attorneys to stay current with the rapidly evolving legal landscape regarding AI legislation as "a growing number of new state laws trigger notice, disclosure, and informed consent considerations" (such as those discussed in Human Best Practice #4, below).[4]
- *Independence and Influence*. A model validation department should exist separately and be "independent from the HR department and any in-house department creating the AI tools," and be headed by a chief risk officer or chief model risk officer with sufficient senior leadership to influence those engaged in model validation.[5]
- *Incentives*. "There is no greater incentive for corporations than avoiding government investigations, large fines, class action lawsuits, negative press, and adverse financial impacts on their businesses."[6] Thus, model validation personnel's financial incentives, such as annual bonuses and promotional opportunities, should be based on the validity of the model—e.g., "the accuracy and number of defects identified as well as the level of risk related to each defect."[7]
- *External Consultants*. Employers are ultimately responsible for employment decisions made by their AI hiring tools, including vendor-provided tools.[8] Thus, employers must also evaluate vendor-provided AI hiring tools, not just their own; and if they lack resources to do so, then they "should consider hiring external consultants to assess [such tools]."[9] Under the CAI, AI hiring tool vendors must, to the extent feasible, supply their customers with sufficient information for them or their external consultants "to complete an impact assessment."[10]
- *Risk Management Policy and Program*. The CAI requires companies deploying AI hiring tools to "implement a risk management policy and program," including specifying and incorporating "the principles, processes, and personnel that the deployer uses to identify, document, and mitigate

known or reasonably foreseeable risks of algorithmic discrimination" (such as those described in the bullets above and below).[11]

Human Best Practice #2—Understand and Prepare the Algorithm's Inputs

Companies should take several steps to understand and prepare the inputs going into their AI hiring tools, so as to mitigate the risk of the outputs violating antidiscrimination statutes, as follows:

- *Understand the Inputs*. Companies need to "know their data" in order to design inputs to AI hiring tools—e.g., to perform the remaining bullets below.[12] Knowing your company's data may require asking vendors about any vendor-provided technology being used.[13]
- *Use Unbiased Training Data*. Companies must ensure that AI hiring tools are not trained on data consisting of protected characteristics—e.g., race, color, religion, sex, national origin, disability, age—or data correlated with protected characteristics—e.g., ZIP code, first name, alma mater, credit history, and participation in hobbies or extracurricular activities.[14]
- *Use Representative Training Data*. "Where possible, developers should seek to utilize training data features that are well distributed across protected groups as a means of mitigating the risks of bias in the tool's outcomes."[15]
- *Track Protected Characteristics Associated with Training Data*. Although protected characteristics should not be used to train AI hiring tools, they should be tracked in a relational database in order to audit the tools and identify any adverse impacts to people with those protected characteristics.[16]
- *Perform Debiasing Techniques at the Model Input's Origins*. "Models trained on biased data generally reproduce and amplify those biases."[17] Thus, "the reliability and lawfulness of the AI's output is only as good as the inputs."[18] For example, an interview tool that analyzes interviews can only avoid disparate impact to the extent disparate impact was avoided by the resume screening tool used to source

the interviews. In turn, the resume screening tool can only avoid disparate impact to the extent disparate impact was avoided by the advertising or preemployment assessment tools used to source the resumes. Therefore, companies should apply bias mitigation techniques, such as those described in the bullets above and below, at the first tool in the chain and work forward, in order to minimize the bias of all inputs.[19]

Human Best Practice #3—Select, Audit, and Tune the Algorithm to Comply with Statutes

Companies can mitigate the risk of their models violating anti-discrimination statutes by taking the following steps to select, audit and tune their models to comply with those statutes:

- *Select Model Type(s)*. A wide selection of models underlying AI hiring tools is publicly available and comes in varieties that may be either intrinsically interpretable (e.g., linear or logistic regression, discriminant analysis, naive Bayes), or black box (e.g., K-nearest neighbors, decision tree ensembles, and neural networks).[20] Candidate tools should be modeled, audited, tuned, and put in competition with each other to select the best model among them in terms of disparate impact or other metrics.[21]
- *Audit Model for Adverse Treatment*. An AI hiring tool must not deny jobs or benefits "because of" the individual's race, color, religion, sex, or national origin (collectively, protected characteristics)[22]; disability[23]; or age.[24] Such adverse treatment could occur, for example, if the algorithm were either explicitly programmed to make decisions on the basis of a protected characteristic or trained with "biased training data to recognize and discriminate against protected characteristics without being explicitly programmed to do so."[25]
- *Audit Model for Adverse Impact*. An AI hiring tool should be analyzed to determine whether it "deprive[s] or tend[s] to deprive any individual of employment opportunities or otherwise adversely affect[s] his status as an employee, because of such individual's race, color, religion, sex, or national origin"[26] or such individual's "age,"[27] or "ha[s] the

effect of discrimination on the basis of disability."[28] These effects are known as "adverse impact" or "disparate impact." Relatedly, the CAI requires deployers of AI hiring tools to perform an "impact assessment" of those tools.[29] Adverse impact occurs where there is a "statistically significant" lesser selection rate for a protected group.[30] Four-fifths difference in selection rate is generally regarded as evidence of an adverse impact but an adverse impact may be found on lesser differences and, moreover, no adverse impact may be found on greater differences.[31] AI hiring tools may also present a "'whack-a-mole' problem" with respect to the impacts they present, where reengineering an algorithm to have less adverse impact on members of one protected group may increase adverse impact on another protected group.[32]

- *Audit Model for Validity*. If an AI hiring tool has an adverse impact on people with protected characteristics, it nevertheless may potentially comply with antidiscrimination statutes if its selection methods are shown to be "job-related" and "consistent with business necessity."[33] This is known as the "validity defense."[34] Validity studies should conform with the Uniform Guidelines on Employee Selection Procedures (UGESP), §§ 1607.5,[35] 1607.14,[36] and 1607.7.[37] Among many other things, this includes making a study of the AI hiring tool's "unfairness where technically feasible," with "[t]he concept of fairness or unfairness of selection procedures [being] a developing concept."[38] Using modern data science techniques, it is technically feasible to measure a variety of fairness metrics, including equal opportunity, negative predictive value, true positive rate, true negative rate, false positive rate, false negative rate, accuracy, statistical parity, conditional use accuracy, error rate quality, and demographic parity.[39]
- *Audit Model for Available Alternatives*. Even if an AI hiring tool with an adverse impact is valid, an employer should investigate any substantially equally valid, but less discriminatory alternatives or reasonable accommodations and, if available, use those instead.[40]
- *Perform Audits at Least Annually*. In terms of frequency regarding the above audits, "[s]ome experts recommend conducting an audit once a year while others emphasize

that audits should be continuous."[41] NYC's AI Law requires that "bias audits" be performed at least annually.[42]

- *Tune Model to Mitigate Disparate Impact*. As a result of all the foregoing audits, employers should apply bias mitigation techniques to the model's training data, the model itself, or the predictions generated by the model.[43] Examples of bias mitigation techniques include oversampling, data augmentation, debiasing of word embeddings, and considering alternative candidate model types.[44] Bias mitigation techniques should be limited to neutral techniques, however, and should exclude techniques that may be subject to claims of "reverse discrimination," such as the use of quotas or race norming.[45]

Human Best Practice #4—Explain the Algorithm

"Transparency and explainability are two very important concepts that foster algorithmic reliability, trust, credibility, and a general understanding of AI systems."[46] Companies should take the following steps toward transparency and explainability in order to comply with antidiscrimination statutes:

- *Comply with Statutory Documentation Requirements*. Title VII requires companies to maintain information "which will disclose the impact" of its employment selection procedures on people by identifiable protected characteristic.[47] The CAI requires (1) developers of AI hiring tools to disclose summaries of the training data types, data governance measures, algorithmic discrimination risk mitigation measures, intended outputs, and sufficient information, if feasible, to enable a deployer to perform an impact assessment;[48] and (2) deployers of AI hiring tools to maintain a risk management policy and impact assessment. The IAIVA requires employers relying on AI interview analysis to making hiring determinations to disclose the race and ethnicity of applicants who (1) are hired, and (2) are and are not afforded the opportunity for an in-person interview after the use of AI.[49]
- *Use Interpretability Methods*. Deep neural networks and other model types present a "'black box' problem, which

results from the difficulty, or impossibility, of explaining why AI tools produced a particular outcome."[50] Various interpretability methods exist that can shine at least a little light inside the black box to enable the auditing practices described in Human Best Practice #3 above, however. These include various global interpretability methods that measure how individual features impact predictions globally (on average) across the model, and local interpretability methods that measure how individual model predictions are influenced by their individual feature values.[51]

Human Best Practice #5—Intervene to Provide Reasonable Accommodations

- *Reasonable Accommodations*. "[C]ompliance with federal antidiscrimination law often requires human intervention, especially when workplace accommodations for disabled and religious employees are at issue."[52] That is because "the ADA and Title VII still require an interactive process to determine the reasonable accommodations when an employer uses AI."[53] Accommodations may be "especially useful and necessary during the interview process," including allowing human interviews in exceptional cases and providing alternative test formats.[54]

Conclusion

AI hiring tools designed to comply with antidiscrimination statutes will comply.[55] Moreover, "by eliminating some human decision-making and replacing it with carefully designed algorithms, AI holds the potential to substantially reduce the kind of bias that has been unlawful in the United States since the civil rights movement of the mid-twentieth century."[56]

This article identified human best practices to assist with such compliance and, relatedly, such potential substantial reduction of bias, as follows: (1) implement an AI governance structure; (2) understand and prepare the algorithm's inputs; (3) select, audit, and tune the algorithm to comply with the statutes; (4) explain the algorithm; and (5) intervene to provide reasonable accommodation.

Notes

* Justin R. Donoho is special counsel at Duane Morris and a Certified Information Systems Security Professional. He may be contacted at jrdonoho@duanemorris.com.

1. The Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §§ 1607.1-1607.18 (UGESP); EEOC Commissioner Keith E. Sonderling et al., The Promise and the Peril: Artificial Intelligence and Employment Discrimination, 77 U. Miami L. Rev. 1 (2022) (Sonderling I); Keith E. Sonderling et al., A New Approach to Measuring AI Bias in Human Resources Functions: Model Risk Management, 54 Seton Hall L. Rev. 965 (2024) (Sonderling II), EEOC Former Chief Analyst Kelly Trindel et al., Fairness in Algorithmic Employment Selection: How to Comply with Title VII, 35 ABA J. Lab. & Emp. Law 241 (2021) (Trindel); Grant Fleming et al., Responsible Data Science (Wiley 2021) (Fleming); and Stuart Russell et al., Artificial Intelligence: A Modern Approach (Pearson 2021) (Russell).
2. Sonderling II at 980.
3. Id. at 981.
4. Sonderling I at 85-86.
5. Sonderling II at 982.
6. Id. at 984.
7. Id.
8. See Mobley v. Workday, Inc., 2024 WL 3409146 (N.D. Cal. July 12, 2024) (finding that vendor-provided AI hiring tool was employer's "agent" under Title VII, the ADA, and ADEA).
9. Sonderling II at 986-97.
10. CAI § 6-1-1702(3)(a).
11. CAI § 6-1-1703.
12. Sonderling I at 75.
13. Id.
14. Trindel at 273; see also IHRA (employers may not "use ZIP codes as a proxy for protected classes" when employing AI hiring tools).
15. Trindel at 274.
16. UGESP, 29 C.F.R. § 1607.4; Fleming at 157-59.
17. Fleming at 261.
18. Sonderling I at 22.
19. See 29 C.F.R. § 1607.4 ("If . . . the total selection process for a job has an adverse impact, the individual components of the selection process should be evaluated for adverse impact").
20. Fleming at 12-19.
21. Fleming at 173-263.
22. Title VII, 42 U.S.C. § 2000e-2(a)(1)).
23. ADA, 42 U.S.C. § 12112(b)(1).
24. ADEA, 29 U.S.C. § 623(a)(1).

25. Sonderling I at 23-24.
26. Title VII, 42 U.S.C. § 2000e-2(a)(2).
27. ADEA, 29 U.S.C. § 623(a)(1).
28. ADA, 42 U.S.C. § 12112(b)(3)).
29. CAI § 16-1-1703(3).
30. UGESP, 29 C.F.R. § 1607.14(D).
31. Id.
32. Sonderling I at 22.
33. Title VII, 42 U.S.C. § 2000e-2(k)(1)(A) & (C); ADA, 42 U.S.C. § 12113(a)); see also ADEA, 29 U.S.C. § 632(f)(1) ("reasonably necessary to the normal operation of the particular business").
34. Trindel at 245.
35. "General standards for validity studies."
36. "Technical standards for validity studies."
37. "Use of other validity studies."
38. UGESP, 29 C.F.R. § 1607.14(8).
39. See Fleming at 183 (defining these fairness metrics and providing formulas); Russell at 992-93 (similar).
40. Title VII, 42 U.S.C. § 2000e-2(k)(1)(A) & (C); UGESP, 29 C.F.R. § 1607.3; Ricci v. Stefano, 557 U.S. 557, 578 (2009)); ADA, 42 U.S.C. § 12113(a)); ADEA, 29 U.S.C. § 632(f)(1).
41. Sonderling I at 79-80.
42. NYC Admin. Code § 20-871.
43. Fleming at 185-86.
44. Id. at 174, 214-18, 261.
45. See, e.g., Title VII's anti-race norming provision, 42 U.S.C. § 2000e-2(l) ("It shall be an unlawful employment practice for a respondent, in connection with the selection or referral of applicants or candidates for employment or promotion, to adjust the scores of, use different cutoff scores for, or otherwise alter the results of, employment related tests on the basis of race, color, religion, sex, or national origin").
46. Sonderling I at 77.
47. 29 C.F.R. § 1607.4.
48. CAI § 6-1-1702.
49. IAIVA, 840 ILCS 42/20.
50. Sonderling I at 22.
51. Fleming at 107-21.
52. Sonderling I at 82.
53. Id.
54. Id. at 82-84.
55. Trindel at 270.
56. Sonderling II at 966.