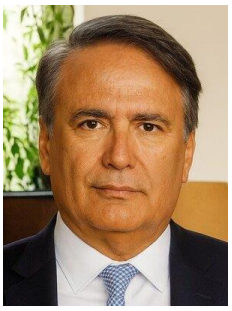
LAW360

Portfolio Media. Inc. | 230 Park Avenue, 7th Floor | New York, NY 10169 | www.law360.com
Phone: +1 646 783 7100 | Fax: +1 646 783 7161 | customerservice@law360.com

Workday Case Shows Auditing AI Hiring Tools Is Crucial

By **Hossein Borhani** (October 27, 2025, 11:37 AM EDT)


Hossein Borhani

The recent proceedings in Mobley v. Workday in the U.S. District Court for the Northern District of California highlight the emerging challenges and legal risks associated with utilizing artificial intelligence in employment decision-making.

In its May 16 decision, the court ruled in favor of Mobley, finding that Workday's AI recommendation system represented a unified policy affecting applicants across employers and justified collective consideration of the claims. In a subsequent ruling on July 29, the court rejected Workday's attempt to narrow the certified collective, further underscoring the legal complexities surrounding AI-driven hiring practices and employment decisions.

These developments have broad implications for the allocation of responsibility between employers and technology vendors. Employers can no longer assume that off-the-shelf algorithmic tools shift compliance obligations to vendors, and vendors face increasing scrutiny over model design and performance. In this environment, systematic evaluation of AI-based hiring systems has become even more essential.

This article outlines complementary empirical approaches for such evaluations: randomized internal experiments using firm-level data and internal matched-pair testing that simulates real applicants. Together, these methods provide a credible framework for assessing whether algorithmic systems operate based on legitimate job-related factors. By adopting these approaches, employers and vendors can enhance evidence-based oversight and ensure explainable algorithmic hiring.

Background

Artificial intelligence has become an increasingly prominent force in human resources, particularly in hiring. Employers are turning to AI-powered software to manage large applicant pools, streamline resume screening, and identify qualified candidates more efficiently than traditional methods allow.

These tools promise cost savings, consistency and speed — attributes that are especially appealing in competitive labor markets. Yet, the rapid adoption of AI in HR functions carries inherent risks.

Hiring databases commonly contain limited information on applicants, leaving out the key experience factors and skills that hiring managers use in decision-making. Algorithms trained on that limited

historical data may yield hiring recommendations that do not necessarily reflect legitimate job-related factors.

The recent proceedings in Mobley v. Workday illustrate how consequential these risks have become.

The plaintiff, Derek Mobley, alleges that Workday acted as employers' agent and, through its AI-powered applicant screening and recommendation tools, had a disparate impact on certain job applicants, namely those who were 40 or older. The court has so far allowed the proceeding to go forward and has certified a collective.

Although many legal questions remain unsettled, Mobley signals that software vendors can no longer assume immunity when their systems play a substantive role in their clients' employment decisions. For both employers and vendors, this emerging framework of shared accountability makes it essential to proactively assess whether recommendations from AI-driven hiring systems are explainable, i.e., related to legitimate factors associated with performance in the position.

From a labor economics perspective, Mobley reflects the growing recognition that algorithmic systems act as active participants in labor market matching, rather than neutral tools. The decision broadens the scope of accountability for employment outcomes beyond employers to technology vendors whose systems structure access to jobs. Both parties now share incentives to monitor how automated screening tools influence demographic hiring patterns and ensure that any observed disparities reflect legitimate job-related factors.

However, key questions remain unresolved. It is not yet clear whether Workday's algorithms produced measurable group disparities or whether human oversight mitigated — or potentially exacerbated — those effects. The scope of vendor liability and the effectiveness of safeguards such as audits, model documentation and review protocols also remain uncertain.

These open issues underscore the need for empirical frameworks to evaluate how algorithmic systems function, the extent to which observed differences reflect legitimate job-related factors, and how technological intermediation shapes equal employment opportunities. In this context, systematic auditing becomes essential.

Scientific Methodologies for Auditing AI Hiring Tools

The central challenge posed by Mobley v. Workday is how employers and vendors can demonstrate that AI-driven hiring systems operate impartially across groups, ensuring that any observed differences reflect legitimate, job-related factors. Because these systems often function as black boxes, assurances of neutrality are insufficient. Instead, employers and vendors must adopt rigorous, empirically grounded auditing practices.

Two complementary methodologies are particularly suited to this task: (1) randomized internal experiments using firm data and (2) matched-pair internal audits simulating real applicants.

Randomized Internal Experiments

One approach is to conduct controlled experiments using the employer's own historical data.[1] This involves applying randomized relabeling or counterfactual analysis to past applications and outcomes to test whether the algorithm recommendations systematically differ by group.

For instance, evaluators can modify a candidate's race, gender or age indicator while holding all other qualifications constant and observe whether predicted outcomes change. These randomized internal experiments — replay tests — rerun relabeled historical applications through the recruiting model, allowing the direct comparison of test results with the historical selection rates and score distributions to assess whether the system responds consistently.

This method offers several advantages: It is cost-effective, uses existing data, operates within the system's actual context and provides a statistically grounded measure of potential disparate impact. The primary limitation of this approach is its reliance on historical data, which may overlook differences that emerge only in live, dynamic environments — for example, how the system interacts with new job postings or evolving applicant pools.

Matched-Pair Audits

To address these limitations, organizations and vendors can conduct a simulated field experiment through matched-pair testing. In this approach, auditors design realistic resumes with randomly assigned group indicators to test whether the hiring system generates different selections or hiring probabilities across groups.[2]

Rather than relying on historical data, evaluators create matched pairs or sets of applications that are identical in qualifications, experience and skills but differ only in demographics. These applications are then processed through the organization's recruiting pipeline — ideally under conditions that mirror live job postings — so that the system's scoring, routing and, where applicable, human reviewer stages function as they would for real applicants. By comparing how the system screens and ranks these applications at each stage, evaluators can measure whether group-related differences arise.

Matched pair testing has a well-established pedigree in labor economics and other disciplines.[3] When adapted as a simulated internal audit of an organization's recruiting system, it provides a direct, real-world assessment of how AI-powered hiring tools operate. Conducting such audits internally, in collaboration with the vendor, allows for tight experimental control and avoids the ethical and legal concerns associated with submitting fictitious applications to external employers.

However, matched-pair audits require careful design, documentation and statistical rigor to be defensible. Their limitations include higher resource costs and the possibility that simulated postings may not fully capture real-world applicant behavior or job-market dynamics.

Nonetheless, when properly executed, matched-pair testing yields valuable insights into the behavior of AI systems across varying conditions and helps organizations to ensure that observed differences reflect legitimate, job-related factors.

The Case for a Combined Approach

Neither internal nor external auditing is sufficient on its own. Internal randomized experiments excel at diagnosing potential algorithmic inconsistencies within existing data, while external matched-pair audits measure how the system functions in practice with new applicants. Together, the two methods provide complementary insights: one is diagnostic and the other experiential.

For employers and vendors facing the potential prospect of joint liability after Mobley, adopting both

approaches creates a more comprehensive and defensible compliance framework. It is also useful to conduct a parallel review by human recruiters to assess whether human judgment aligns with the tool's recommendations.

Practical Guidance for Employers and Vendors

The doctrinal shift signaled by Mobley v. Workday suggests that both employers and software vendors now face heightened scrutiny and potential liability regarding the impartiality of their AI-driven systems. In this environment, auditing is no longer an optional best practice but a legal and strategic necessity. The following considerations outline how organizations can integrate auditing into their broader compliance frameworks.

Establishing an Audit Framework

Employers and vendors should design a clear independent audit protocol at the outset of their contractual relationship. This protocol should specify the scope of the audit (e.g., whether it covers screening, ranking, interview scheduling), the methodologies to be employed (randomized internal testing and external matched-pair audits), and the frequency of review. Building these requirements into vendor-client agreements ensures that compliance obligations are understood and enforceable.

Independent and Credible Review

Although internal compliance teams can initiate audits, the credibility of results often depends on independent oversight. An audit conducted with genuine independence is more likely to withstand judicial and regulatory scrutiny, and independence enhances not only the evidentiary weight of the findings but also public trust.

Documentation and Attorney-Client Privilege

From a practical standpoint, maintaining clear documentation is essential for understanding and improving system performance over time. Conducting audits under attorney-client privilege can also create space for more open evaluation and discussion of potential issues between the employer and the vendor.

Frequency and Triggers for Audits

Audits should not be one-off exercises. To align audits with the dynamic nature of AI systems, organizations should conduct audits with a regular cadence (e.g., annually), but additional audits should also be triggered whenever there are significant changes in the software's design, the company's organizational structure, training data or deployment context.

Integrating Audit Findings Into Governance

The ultimate purpose of an audit is identifying risk areas that require a deeper review, leading to improvements in the processes. Employers and vendors must establish governance mechanisms for implementing corrective actions when audits reveal areas that need further attention and further exploration. This may include retraining algorithms, revising selection criteria or reintroducing human review at key decision points. Governance structures should assign clear responsibility for follow-up, ensuring that audit results translate into systemic change.

Takeaways

In the wake of the evolving legal framework affecting the use of AI tools, employers can no longer outsource compliance responsibilities to their vendors, and vendors cannot rely on claims of technological neutrality. Joint accountability requires joint vigilance. By implementing structured audit protocols, both parties can not only mitigate risk but also strengthen fairness and transparency in the hiring process.

It is also important to note that statistical results may indicate group-related differences or disparate impact even when the observed disparities arise from flaws in the statistical design, such as improper data aggregation.[4] Therefore, a careful statistical design that appropriately accounts for all relevant recruiting factors is essential to ensure the robustness of the audit results.

Conclusion

Mobley signals a potential shift in the governance of AI-based hiring tools, emphasizing that both employers and vendors must take a more proactive role in evaluating and defending the neutrality of their recruiting systems. Employers cannot outsource compliance, and vendors cannot rely on disclaimers of neutrality; both must actively ensure fairness.

Rigorous auditing — through randomized experiments or matched-pair testing — offers a practical way to assess whether AI-driven hiring systems operate impartially across groups and to determine if any observed differences reflect legitimate, job-related factors. This case underscores that algorithmic efficiency must not come at the expense of equity, and proactive auditing is essential to manage legal risk and promote inclusive hiring.

Hossein Borhani is a vice president at Charles River Associates.

[1] Gerber, A. S., & Green, D. P. (2012). Field Experiments: Design, Analysis, and Interpretation. W. W. Norton.

[2] Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." American Economic Review 94 (4): 991–1013.

[3] Havens Realty Corp. v. Coleman, 455 U.S. 363 (1982).

[4] Bickel PJ, Hammel EA, O'Connell JW. Sex Bias in Graduate Admissions: Data from Berkeley. Science. 1975. 187(4175): 398-404.