Unpriced Risk Is Not Return: Why Most Artificial Intelligence Business Cases in Human Resources Fail as Business Cases

Heather Grizzle · April 4, 2025

The savings that are measurable in HR AI are small. The savings that are large cannot be measured. What a defensible artificial intelligence business case includes.

Contents

1. The claims and the measurement that would be needed to support them
2. What is actually measurable, and what it is worth
3. Quality of hire is not a metric
4. Predictive attrition and the accuracy gap
5. Automated compliance does not discharge a compliance obligation
6. The costs the model omits
7. A defensible structure
8. Conclusion
9. References

The claims and the measurement that would be needed to support them

The standard business case for artificial intelligence in human resources has stabilized around five claims: faster and better hiring, lower administrative cost, predictive workforce planning, improved retention, and automated compliance. The claims are repeated with enough consistency across vendor materials, conference programming, and professional commentary that they now function as background assumption rather than as propositions requiring evidence.

They require evidence, and the obstacle to producing it is not the technology. Human resources entered this period with the weakest measurement infrastructure of any major corporate function. Finance has audited statements. Operations has throughput and defect rates. Human resources has a set of metrics that are variably defined, rarely validated, seldom baselined before an intervention, and frequently produced by the vendor whose product is being evaluated. An organization that could not measure the performance of its hiring process before deploying an automated tool cannot measure the improvement afterward, and what it reports as return is generally the difference between two numbers that were never comparable.

Novara Consulting Group's position is that the central problem with artificial intelligence business cases in human resources is not that the returns are absent. Some are real, and they are described below. It is that returns are asserted at a level of precision the underlying measurement cannot support, while the costs that would offset them are excluded from the model entirely. A calculation constructed that way is not an investment analysis. It is an advocacy document with a currency symbol in it.

What is actually measurable, and what it is worth

It is an advocacy document with a currency symbol in it.

The five standard claims separate cleanly into two classes, and the separation is more instructive than the claims themselves.

The first class contains effects that are genuinely measurable and comparatively modest. Cycle time in recruitment is measurable, has an established baseline in most applicant tracking systems, and does respond to automation of scheduling, screening, and correspondence. Administrative cost per transaction in payroll and benefits administration is measurable, and so are error rates in those processes. Organizations that automate these functions generally do realize savings, and those savings can be substantiated.

Two qualifications matter. The first is that most documented savings in payroll and benefits administration derive from workflow automation and rules-based processing, technologies that predate the current generation of artificial intelligence by decades and that are frequently relabeled as AI in vendor positioning and in the business cases built on it. Attributing long-established process automation savings to a recent AI purchase inflates the measured return of that purchase and, more consequentially, misattributes the causal mechanism, which means the organization learns the wrong lesson about what produced the benefit. The second qualification is that these savings are, in absolute terms, unspectacular.

They do not on their own justify the enterprise commitments now being made.

The second class contains effects that would be substantial if they were real and that cannot presently be measured: improved quality of hire, improved retention, and better workforce planning. These are the claims that carry the business case, and they carry it precisely because their magnitude is unconstrained by measurement.

Quality of hire is not a metric

Quality of hire is the load-bearing claim in nearly every recruitment technology business case, and it does not have a validated common definition. Depending on the organization it is operationalized as first-year performance rating, ninety-day retention, twelve-month retention, hiring manager satisfaction, time to productivity, or a composite of several of these, each with different measurement properties and none with established construct validity as a measure of hiring effectiveness.

The consequence is that any claimed improvement in quality of hire is substantially an artifact of definitional choice. An organization that defines quality of hire as ninety-day retention and deploys a tool that screens for candidates resembling prior long-tenured employees will observe an improvement. It has not learned that hiring improved. It has learned that a system optimized for a proxy moved the proxy. Where the vendor supplies both the tool and the measurement framework, which is common, the finding is closer to a tautology than to evidence.

This matters well beyond methodological hygiene. Under the Uniform Guidelines on Employee Selection Procedures at 29 C.F.R. Part 1607, an employer whose selection process shows adverse impact is expected to produce evidence that the procedure is valid for the job. Criterion-related validation requires a defensible measure of job performance. An organization that cannot say precisely what quality of hire means for its own roles, and cannot show that its criterion measure is itself unbiased, does not have the evidence the framework contemplates. The measurement weakness that inflates the business case is the same weakness that undermines the legal defense, which is an unusual instance of a methodological problem and a liability problem being the same problem.

Predictive attrition and the accuracy gap

Predictive workforce planning deserves separate treatment, because the distance between published performance and deployed performance is widest here and the legal exposure is least understood.

The academic literature on employee attrition prediction routinely reports accuracy figures in the high nineties. Those figures are generated overwhelmingly on a small number of benchmark datasets, principally the IBM HR Analytics set of roughly 1,470 records and a widely circulated Kaggle human resources set of roughly 15,000 records, and typically after synthetic class balancing through oversampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique). These are legitimate methodological choices for comparing algorithms against one another. They are not estimates of field performance on an employer's own workforce, and they should never appear in a business case as though they were. Practitioner assessment of realistic field accuracy for flight risk models is considerably lower, in the range of seventy to eighty percent, and analysts in the field have expressed sustained skepticism toward vendor claims above ninety percent.

The gap between eighty percent and the high nineties is not a rounding difference in this application, because the base rate is low. Voluntary attrition in most organizations runs well under twenty percent annually, which means a model with strong headline accuracy can still produce false positives substantially outnumbering true positives. In an attrition model, a false positive is an employee incorrectly identified as likely to leave.

The governance question follows immediately, and the return-on-investment framing obscures it. What does the organization do with that name. If the answer is nothing, the model has produced no return. If the answer is anything at all, whether differential retention offers, altered development investment, exclusion from long-horizon assignments, or changed succession consideration, then the organization has made a differential employment decision about an individual on the basis of a prediction. That is an employment action, and its lawfulness depends on what drove the score. Guidance on building these models routinely recommends including employee demographics, with age and tenure named among the inputs that give the model context. An employer that feeds age into a model, uses the output to allocate development investment, and cannot reconstruct why a given employee was scored as they were has constructed an Age Discrimination in Employment Act exposure and entered it in the ledger as a benefit.

A COMPOUNDING OPERATIONAL RISK

Commentators have also noted a compounding operational risk that has no line in any business case: flight risk scores communicated to managers who have not been trained to act on them can themselves produce the departure the model predicted.

Automated compliance does not discharge a compliance obligation

The claim that artificial intelligence ensures adherence to labor law and internal policy should be removed from business cases entirely, and not because the tools are useless. Monitoring systems that flag missed breaks, overtime thresholds, classification anomalies, or leave administration errors can be genuinely valuable, and the operational case for them is reasonable.

The word that fails is ensures. A monitoring system does not discharge a legal obligation; it produces outputs the employer then owns. If the system flags a potential violation and no one acts, the organization has created a dated internal record establishing notice, which is materially worse than not having monitored. If the system fails to flag a violation that occurred, the employer has acquired no defense, because reliance on a vendor tool is not a recognized excuse for non-compliance under the Fair Labor Standards Act or the statutes the Equal Employment Opportunity Commission (EEOC) enforces. If the system's determinations are wrong in a patterned way, the employer has automated the error and scaled it.

The events of January 2025 make this worse rather than better. When the EEOC removed its artificial intelligence guidance from its website, including the May 2022 technical assistance on the Americans with Disabilities Act and the May 2023 technical assistance on Title VII, employers lost the interpretive material that told them which practices the agency considered problematic. The statutes did not change. A compliance tool configured against guidance that has since been withdrawn is now operating without a stated reference standard, and the employer, not the vendor, holds that gap.

Compliance automation therefore adds a documentation and response obligation rather than removing one. That obligation has a cost, the cost is real, and it belongs on the cost side of the model.

The costs the model omits

Most human resources AI business cases contain a cost side consisting of license fees, implementation services, and some allowance for training. That is the acquisition cost of the software. It is not the cost of operating an automated employment decision system, which additionally includes at minimum the following.

Validation work, meaning the analysis establishing that what each system scores corresponds to a documented requirement of the job, together with the revalidation required when the system or the role changes. Accommodation pathway operation, meaning the standing capacity to provide an alternative assessment on request, with named decision-makers, defined turnaround, and a functioning interactive process rather than a form. Record retention and reproduction capability, meaning the ability to produce the score, the inputs, and the system configuration in force for any individual decision throughout the preservation period contemplated by 29 C.F.R. Part 1602. Monitoring, meaning periodic testing of outcomes rather than reliance on a vendor summary. Vendor cooperation in discovery and administrative investigation, which is expensive, is absent from standard commercial terms unless negotiated, and is charged to the employer when needed. And the internal governance function itself, meaning the identified person accountable for the above, which in most organizations does not currently exist and must be created or contracted.

Alongside these operating costs sits a contingent liability that virtually no business case prices, and it is no longer hypothetical. In August 2023 the EEOC settled with iTutorGroup for $365,000 over software configured to reject applicants above specified ages, which is a concrete figure for a simple configuration decision. In July 2024 the Northern District of California allowed claims against a screening software vendor to proceed on the theory that the vendor acted as an agent of the employers using its tools, which unsettles the common assumption that liability sits entirely with either the employer or the vendor and can be allocated cleanly by contract. In March 2025 the American Civil Liberties Union (ACLU), the ACLU of Colorado, Public Justice, and Eisenberg and Baum filed a complaint with the Colorado Civil Rights Division and the EEOC on behalf of a Deaf and Indigenous employee concerning an automated video interview and a denied captioning accommodation, a matter both companies deny. That claim is individualized, requires no comparative statistics, and is inexpensive to bring.

The correct treatment is not to guess at a settlement figure. It is to state the exposure explicitly, note that systemic-practice claims are frequently outside the coverage of standard employment practices liability policies, and require that the governance controls reducing the exposure be funded as part of the deployment rather than deferred.

An organization that adds these lines will frequently find the investment still makes sense. That is a fine outcome and it is not the point. The point is that the resulting number can be defended to a board, an auditor, and eventually a court, which the current number cannot.

A defensible structure

The following produces a business case that survives scrutiny and can be built by an existing human resources and finance function.

1. Establish the baseline before deployment, not after: cycle time, cost per hire, administrative cost per transaction, error rates, and attrition by segment, each measured for at least four quarters prior.
2. Separate automation savings from AI savings. State which claimed benefits derive from rules-based workflow processing that would be available without the AI purchase, and price those separately.

3. Define quality of hire explicitly for your own roles before measuring any change in it, state the criterion measure, and state its limitations. If it cannot be defined, do not claim improvement in it.
4. Refuse benchmark-derived accuracy figures for any predictive model. Require field performance on the organization's own data, with the base rate stated and false positive volume expressed in headcount rather than percentage.
5. For any model producing individual-level scores, document what action follows from the score before deployment, and have that action reviewed as an employment decision rather than as an analytics output.
6. Price the operating obligations named above as recurring cost, not as one-time implementation.
7. State the contingent liability as a disclosed line item with the controls that reduce it, whether or not a dollar figure is assigned.
8. Require validation evidence, field performance data, and reproduction capability as contract

terms at acquisition. Retain any refusal, which is itself information about what the organization knows.

Conclusion

There is a real business case for artificial intelligence in human resources, and it is more modest and more defensible than the one currently in circulation. It rests on cycle time, transaction cost, and error reduction in administrative processes, all of which can be substantiated. It does not rest on quality of hire, retention, or predictive accuracy claims that the function's measurement practice cannot support and that carry liability the model does not price.

The organizations that will report durable return from these systems are not the ones that adopted earliest or bought most. They are the ones that measured before they deployed, that can say what each system scores and why it matters to the job, and that treated the governance apparatus as part of the cost of ownership rather than as an expense to be avoided. Return that depends on unmeasured benefit and unpriced risk is not return. It is a forecast, and the difference becomes apparent at the point where someone asks for the working.

References

- American Civil Liberties Union, ACLU of Colorado, Public Justice, and Eisenberg & Baum LLP. Complaint filed with the Colorado Civil Rights Division and the U.S. Equal Employment Opportunity Commission, March 19, 2025.
- EEOC v. iTutorGroup, Inc., No. 1:22-cv-02565 (E.D.N.Y.), consent decree August 2023.
- Mobley v. Workday, Inc., No. 3:23-cv-00770 (N.D. Cal.), order on motion to dismiss, July 2024.
- Society for Human Resource Management. "The Dangers of Using Predictive Analytics to Gauge Employee Flight Risk." 2023.
- U.S. Equal Employment Opportunity Commission. *The Americans with Disabilities Act and the Use of*

Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees, EEOC-NVTA-2022-2, May 12, 2022, and *Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII*, May 18, 2023. Both removed from the agency website January 27, 2025.

- Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. Part 1607; recordkeeping requirements, 29 C.F.R. Part 1602.