CADENCE Management by design, not by luck.

PEOPLE SCIENCE · RESEARCH SYNTHESIS

Can AI Be Fair? The People Science of Algorithmic Talent Decisions

The prediction research is real, the fairness math is unforgiving, and the honest answer is architectural — not aspirational.

Read the live version at https://cadencehr.ai/resources/can-ai-be-fair
Research synthesis published by Cadence, July 2026. Citations are numbered in the text and listed in full under References; every source was verified to resolve as of 2026-07-27.

The vendor demo is going well until the employment lawyer in the back of the room asks her question: "When your algorithm ranks our people, and the ranking turns out to skew against a protected group — who validated it, who audits it, and who owns the decision it fed?" The sales engineer talks about accuracy. She didn't ask about accuracy. Every company evaluating AI for people decisions eventually meets some version of this question, and most answers are worse than the silence that follows.

This document is Cadence's attempt to answer it properly — with the actual research on algorithmic versus human judgment, the actual impossibility results, the actual law, and an honest account of what Cadence's AI does and refuses to do. It is part of the People Science research pillar, and it holds itself to that pillar's standard: no claim without evidence, no evidence without its boundary conditions.

Who this is for

- **HR and People Operations leaders** deciding how much decision authority any AI tool — Cadence included — should have over their people.
- **Legal and compliance teams** who need the research and regulatory landscape in one place, stated without vendor varnish.
- **Executives and investors** testing whether an AI-people-ops company can talk about fairness without either overclaiming or evading.
- **Skeptics of AI in people decisions** — this page concedes more of your case than most vendor content will, and then explains what a defensible architecture looks like anyway.

Key takeaways

1. Across 136 studies of human health and behavior, mechanical (rule-based) prediction was about 10% more accurate than clinical (holistic) judgment on average, and holistic judgment was clearly superior in only a small minority of comparisons. [1]

2. In employment selection and admissions specifically, meta-analytic evidence shows that mechanically combining the same information experts see outperforms holistic expert combination of it. [2]
3. Predictive accuracy does not imply fairness: algorithmic systems trained on historical data can inherit and formalize past discrimination — through biased labels, skewed samples, and proxy variables — without any discriminatory intent by their designers. [3]
4. It is mathematically impossible for a risk score to simultaneously satisfy calibration and balanced error rates across groups, except when base rates are equal or prediction is perfect — so every algorithmic talent tool embodies a fairness trade-off, whether or not its vendor acknowledges one. [4]
5. An audit of vendors selling algorithmic pre-employment assessment found wide variation in what "bias" and "validation" mean in practice, and concluded that vendor claims must be evaluated against both technical and legal standards, not taken at face value. [5]
6. U.S. employment law already reaches algorithms: under the EEOC's Uniform Guidelines framework, a selection procedure with adverse impact requires validation regardless of whether it is a paper test, an interview protocol, or a model. [6]
7. Regulation is operational, not hypothetical: New York City's Local Law 144 requires an independent bias audit, a published results summary, and candidate notice before an automated employment decision tool is used — and the NIST AI Risk Management Framework gives organizations a voluntary structure for governing AI risk, including bias. [7] [8]

Doesn't the research say algorithms beat human judgment?

On the narrow question of predictive accuracy, yes — and Cadence will not pretend otherwise, because this literature is one of the most replicated findings in applied psychology.

The tradition runs from Paul Meehl's 1954 monograph through Grove and colleagues' meta-analysis of 136 studies comparing clinical judgment with mechanical prediction across psychology and medicine: mechanical prediction was superior or equal in the great majority of comparisons and about 10% more accurate on average, while clinical judgment was clearly superior in only a handful of studies. [1] Kuncel and colleagues brought the question directly into employment territory: when experts and algorithms combine the *same* information to predict job performance or academic success, mechanical combination outperforms holistic judgment. [2]

The mechanism is mundane, which is why it generalizes. Humans weight evidence inconsistently — the same facts on a tired Friday get a different rating than on a fresh Monday. We anchor on vivid details, substitute confidence for validity, and cannot hold forty data points in working memory. A formula applies the same weights every time.

So the honest starting point is uncomfortable for both camps. To the AI skeptic: unstructured gut judgment is not the safe harbor it feels like — it is noisy, inconsistent, and carries its own biases with no audit trail at all. To the AI enthusiast: everything that follows in this document is why "more accurate on average" is nowhere near the end of the fairness conversation.

If algorithms predict better, why aren't they automatically fair?

Because accuracy and fairness are different properties, measured against different standards, and the second does not follow from the first.

Barocas and Selbst's analysis of data-driven discrimination lays out the mechanisms with legal precision. An algorithm learns from historical decisions and outcomes. If past performance ratings were biased, the model learns biased labels as ground truth. If certain groups are underrepresented in the data, the model is simply worse at predicting for them. If seemingly neutral variables — zip code, employment gaps, activity patterns — correlate with protected class membership, the model can reconstruct the protected attribute it was never given. None of this requires intent; it can emerge entirely from routine engineering choices, which is precisely what makes it hard to see and hard to litigate. [3]

Raghavan, Barocas, Kleinberg, and Levy then looked at what the algorithmic hiring market actually does about this: they examined vendors of algorithmic pre-employment assessments and found wide variation in what companies disclose about validation, what they mean by "bias," and how their debiasing practices map onto either technical fairness criteria or employment-discrimination law. Their conclusion was not that the tools are all bad — it was that vendor claims must be independently evaluated rather than accepted. [5]

The practical translation for a buyer: "our model is highly predictive" and "our model is fair" are separate claims requiring separate evidence, and a vendor offering the first as proof of the second is answering the wrong question.

Can't we just define fairness and test for it?

Here the news gets mathematically worse, and any vendor who does not tell you this is either unaware of it or hoping you are.

Kleinberg, Mullainathan, and Raghavan proved that three natural fairness conditions — that risk scores be calibrated within each group (a "7" means the same thing regardless of group), that groups have balanced false-positive experience, and that groups have balanced false-negative experience — cannot all be satisfied simultaneously, except in degenerate cases: when the groups have identical base rates, or when prediction is perfect. [4] Neither degenerate case describes any real workforce dataset.

This is not a solvable engineering deficiency. It is a theorem. Whenever base rates differ across groups — and in observed historical data they typically do, sometimes *because* of the past discrimination in that data — any scoring system must trade one fairness criterion against another. A tool that is calibrated will produce unequal error rates across groups; a tool that equalizes error rates will sacrifice calibration.

Two implications matter for talent decisions:

- **"Our algorithm passed the fairness test" is an underspecified claim.** Which criterion? At whose expense? Chosen by whom? The impossibility result means the choice was made, explicitly or by default.
- **Fairness in algorithmic talent decisions is governance, not certification.** The defensible posture is to choose criteria deliberately, document the trade-off, monitor outcomes, and keep humans accountable for the decisions — not to claim the math problem has been made to disappear.

What does employment law actually say about algorithmic talent tools?

More than most vendors mention, and less than a compliance checkbox can satisfy.

The Uniform Guidelines already apply. The EEOC's framework for employment tests and selection procedures does not care about the technology inside the procedure. Any measure used as a basis for an employment decision — hiring, promotion, referral, retention — that disproportionately screens out a protected group triggers the adverse-impact framework and requires validation showing the procedure is job-related and consistent with business necessity. [6] An algorithm that feeds promotion or performance decisions sits squarely inside that definition. The hub makes the same point about calibration processes generally: a talent review is not exempt from adverse-impact discipline just because it happens in a workshop instead of a hiring funnel — and it does not become exempt when the workshop gets a model. [6]

New York City made audits mandatory. Local Law 144 prohibits employers and employment agencies from using an automated employment decision tool for hiring or promotion in NYC unless the tool has undergone an independent bias audit within one year of use, a summary of the audit results is publicly available, and required notices go to candidates and employees. [7] Whatever one thinks of its scope, it establishes a precedent: the burden of demonstrating fairness sits with the party deploying the tool, on a clock, in public.

NIST provides the governance frame. The NIST AI Risk Management Framework is voluntary, but it gives organizations a structured vocabulary — govern, map, measure, manage — for identifying and mitigating AI risks, including harmful bias, across the system lifecycle. [8] For a People team, it is a useful standard to hold vendors against precisely because it assumes risk is managed continuously, not certified once.

The direction of travel is consistent: more jurisdictions, more audit obligations, more disclosure. A talent-decision architecture that depends on algorithmic classifications being unimpeachable is betting against both the math above and the regulatory trend.

What would make an AI talent tool unfair?

A frank checklist, applicable to any vendor including Cadence. An AI talent tool moves toward unfairness when:

1. **It learns from unexamined history.** Training on past ratings, promotions, or terminations without asking whether those records encode past bias imports the bias as ground truth. [3]
2. **It scores people against proxies.** Variables that track protected characteristics — location, schedule patterns, gaps, affinity signals — let a model reconstruct what the law says must not drive the decision. [3]
3. **Its errors are unevenly distributed.** A tool can be accurate on average while being systematically wrong about a subgroup, especially one underrepresented in training data. Average accuracy hides distributional harm. [4]
4. **Its fairness claim is unspecified.** "Debiased" without naming the criterion, the trade-off, and the validation evidence is marketing, not measurement. [5]

5. **Its outputs become decisions without human accountability.** When a score triggers consequences directly — filtered out, ranked down, flagged — a false positive costs the person, and nobody in particular is answerable. The severity of an unfair outcome scales with how automatic the pipeline is.
6. **It cannot be audited.** If the deploying employer cannot explain what the tool considers, cannot reproduce its behavior, and cannot monitor outcomes by group, it cannot meet its own legal burden — which the Uniform Guidelines and Local Law 144 place on the employer, not the vendor. [6] [7]
7. **It is unfair silently.** The most dangerous failure mode is not a biased tool; it is a biased tool nobody is measuring, wrapped in a claim that the algorithm handled fairness already.

Note what is *not* on the list: "it uses an algorithm at all." The mechanical-prediction literature is clear that unstructured human judgment is not a fairness refuge — it is inconsistent, undocumented, and biased in its own ways, with no audit trail. [1] [2] The choice is not between biased algorithms and fair humans. It is between decision systems designed for scrutiny and decision systems that hide from it.

How does Cadence answer the fairness question?

Architecturally. Not with a claim that its AI is fair, but with a design that bounds what the AI is allowed to decide — which is nothing.

AI surfaces patterns and prepares humans; humans make talent decisions. In Cadence, AI drafts, summarizes, and flags (AI summaries and coaching are in preview); it prepares a manager for a 1:1, surfaces convergence across signals — goals, recognition, engagement, ER history — and puts the pattern in front of a human with the context to interrogate it. It does not rank employees for termination, auto-score potential, or gate anyone's advancement. **AI develops managers, not replaces them** — and in the fairness context that refrain is load-bearing design, not slogan: the failure cost of the system is calibrated so that *a false positive costs a conversation, not a career.* If Cadence's AI wrongly flags a healthy situation, a manager spends thirty minutes confirming things are fine. That error budget is survivable in a way that a wrongly filtered candidate or an auto-downgraded rating is not.

Classifications are human-reviewed by construction. Cadence's talent-classification workflow — 9-box calibration — is roadmap (Coming Q3), and it is being designed as a structured *human* calibration process, per the 9-Box Talent Calibration Guide: two levels of raters scoring concurrently and blind, gaps between raters surfaced as discussion prompts and potential bias signals, prior ratings visible as trend rather than destiny, and every resulting action owned by a named human. That is the structured-evidence discipline the mechanical-prediction literature supports [2] — applied to *how humans combine evidence*, not as a license to let a model label people.

Fairness obligations stay visible. Because calibration outputs feed employment decisions, they sit inside adverse-impact scrutiny [6] — Cadence's position is that a talent tool should make that discipline easier to honor, not easier to forget: consistent criteria, documented process, and an auditable record of who decided what on which evidence. That documentation posture is also what privacy-respecting AI requires; see AI People-Ops Privacy for how the same architecture handles employee data, and AI That Develops Humans for the development-first design thesis.

What should you ask any vendor of an algorithmic talent tool?

The research and law above compress into a diligence script. These questions apply to every vendor in the category, and Cadence expects to be asked them too:

1. **What decisions does your system make versus inform?** The single most important architectural fact. Anything the system decides autonomously inherits the full weight of the adverse-impact framework with no human accountability layer. [6]
2. **What was the model trained on, and who examined that history for encoded bias?** "Historical performance data" is not a reassuring answer by itself — it is the primary mechanism by which past discrimination becomes future prediction. [3]
3. **Which fairness criterion do you optimize, and what did it cost?** The impossibility result guarantees a trade-off exists. [4] A vendor who says "all of them" has not done the math; a vendor who names the choice has at least made one.
4. **Are error rates reported by subgroup, or only in aggregate?** Average accuracy can conceal a subgroup the model is systematically wrong about. [4]
5. **What validation evidence exists, and would it survive Uniform Guidelines scrutiny?** Job-relatedness and business necessity are the employer's burden to demonstrate — the vendor's evidence is where that demonstration starts. [5] [6]
6. **Has the tool had an independent bias audit, and is the summary public?** Mandatory for automated employment decision tools used in NYC; a reasonable ask everywhere. [7]
7. **Can we audit and monitor it ourselves?** The NIST AI RMF assumes risk is managed continuously across the lifecycle; a tool that cannot be monitored by its deployer cannot be governed by its deployer. [8]
8. **What happens when the system is wrong about a person?** Trace the false positive to its consequence. If the answer is "a human double-checks in a conversation," the error budget is survivable. If the answer is "they're filtered out," it isn't.

Cadence's answers, for the record: its AI informs rather than decides; its classification workflow (9-box, roadmap Coming Q3) is structured human calibration rather than model scoring; and its false-positive cost is a conversation. Where Cadence's evidence is not yet independent — no third-party audit of a talent-classification feature that has not shipped — this page says so rather than implying otherwise.

The strongest objection: isn't "human review" just a fig leaf?

Steelman the skeptic, because this objection is serious:

> Automation bias is well documented in operator research: put a machine recommendation in front of a busy human and the human tends to agree with it. Your "humans decide" architecture may be a rubber stamp with extra steps — the algorithm effectively decides, while the human absorbs the legal accountability. Meanwhile the mechanical-prediction literature you yourself cite says humans add noise when they override formulas. You can't invoke Kuncel to structure judgment and then claim human override makes everything fair.

Three honest responses:

1. **The objection is right that review can be theater — which is why the design target is reviewability, not a checkbox.** A rubber stamp is a human asked to approve a conclusion. Cadence's pattern is to hand humans *evidence with disagreement built in*: blind concurrent ratings that can diverge, cross-signal patterns that can conflict, prompts framed as questions to investigate rather than conclusions to ratify. Structure that surfaces disagreement is harder to rubber-stamp than a single confident score.
2. **The stakes asymmetry is the real safeguard.** Automation bias is most dangerous when the automated output flows directly into a high-stakes outcome. Cadence's AI outputs flow into conversations and reviews, not into terminations, gates, or rankings. Where a human lazily accepts an AI prompt, the cost is a redundant check-in — the failure mode is bounded by architecture, not by hoping humans stay vigilant.
3. **The Kuncel point cuts differently than the skeptic thinks.** The mechanical-combination advantage is about *consistently combining evidence*, and Cadence applies it exactly there: structured criteria, consistent process, documented weights in human calibration. [2] It is not a finding that models should make employment decisions autonomously — and the same literature's authors, writing on algorithmic hiring, explicitly warn against taking vendor fairness claims on faith. [5]

What would change our mind: if deployment evidence showed managers systematically deferring to AI prompts rather than investigating them, or human-reviewed calibration reproducing the same disparities as unreviewed scoring, the architectural answer would be failing its own test — and this page would say so.

What Cadence does not claim

Do not claim, and this page does not:

- "Cadence's AI detects bias objectively." No system does; the People Science hub makes the same commitment, and this page inherits it.
- "Cadence's AI is fair" as a certified property. Fairness criteria trade off against each other [4]; Cadence's claim is about decision architecture and human accountability, not mathematical absolution.
- "Cadence's 9-box ratings are more accurate than human judgment." 9-box calibration is roadmap (Coming Q3), and it is designed as structured human judgment — not a model to out-predict humans.
- "Using Cadence satisfies EEOC, Local Law 144, or any legal obligation." Legal compliance belongs to the employer and its counsel; software can make the underlying discipline easier to practice and document, nothing more.

- Any Cadence customer outcome ("Cadence reduces bias by X%"). Cadence has no customer outcome data and says so proudly.

What Cadence does claim: its AI prepares and informs humans rather than classifying people autonomously; its talent workflows are designed so classifications are human-reviewed; and its architecture is built so that the cost of an AI error is a conversation, not a career.

The argument in one paragraph

Structured, mechanical combination of evidence beats unstructured holistic judgment on average — that finding is real, replicated, and it indicts gut-feel talent decisions more than it licenses algorithmic ones. Fairness does not come along for free: trained-in bias is a documented mechanism, the fairness criteria trade off by theorem, and the law already holds employers accountable for adverse impact regardless of the technology that produced it. So the honest answer to "can AI be fair?" is: no tool can certify its own fairness, and the trustworthy design is the one that never needs to — AI that surfaces patterns and prepares humans, humans who make and own the talent decisions, classifications that are reviewed rather than pronounced, and an error budget where a false positive costs a conversation, not a career. AI develops managers, not replaces them — in fairness terms, that is not a slogan; it is the load-bearing wall.

How to cite this document

Suggested citation: Cadence, "Can AI Be Fair? The People Science of Algorithmic Talent Decisions" (2026). https://cadencehr.ai/resources/can-ai-be-fair

Methodology and provenance. This synthesis was drafted in July 2026 by Cadence as part of the People Science research pillar. Sources were selected with a preference for peer-reviewed meta-analyses (clinical-versus-mechanical prediction), formal results (fairness impossibility theorems), peer-reviewed legal scholarship, and primary regulatory sources (EEOC, NYC DCWP, NIST) read directly rather than through secondary summaries. Every citation below was verified to resolve to the named source as of 2026-07-27. Claims about Cadence's product are labeled with current availability (live / preview / roadmap); claims that are Cadence's own design position rather than external research findings are identified as such in the text.

References

1. Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment*, 12(1), 19–30. doi:10.1037/1040-3590.12.1.19
2. Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). "Mechanical Versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis." *Journal of Applied Psychology*, 98(6), 1060–1072. doi:10.1037/a0034156
3. Barocas, S., & Selbst, A. D. (2016). "Big Data's Disparate Impact." *California Law Review*, 104(3), 671–732. doi:10.15779/Z38BG31
4. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (ITCS 2017)*. arXiv:1609.05807
5. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT| 2020)*, 469–481. doi:10.1145/3351095.3372828

6. U.S. Equal Employment Opportunity Commission, "Employment Tests and Selection Procedures" (fact sheet on the Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. Part 1607). eeoc.gov
7. New York City Department of Consumer and Worker Protection, "Automated Employment Decision Tools (AEDT)" (Local Law 144 of 2021). nyc.gov
8. National Institute of Standards and Technology, "AI Risk Management Framework" (AI RMF 1.0, 2023). nist.gov

---

This is a research synthesis and design-position statement, not a Cadence customer-outcome claim or legal advice. Module availability is labeled because this page covers both live and roadmap concepts.

This article is part of Cadence's People Science research pillar.

See how Cadence turns people science into operating rhythm at cadencehr.ai/product, or check plans at cadencehr.ai/pricing.
