

From Prediction to Practice: Artificial Intelligence in Performance Management and Employee Retention

A conceptual review and an integrative framework linking evaluative technology, justice perceptions, and turnover behaviour

CT. Vasantha¹, Dr. S. Bharathi²

¹Research Scholar, School of Management, Dhanalakshmi Srinivasan University, Samayapuram, Trichy

²Research Supervisor and HoD, School of Management, Dhanalakshmi Srinivasan University, Samayapuram, Trichy

Abstract—

Artificial intelligence (AI) has entered two of the most consequential domains of human resource management: the evaluation of employee performance and the management of voluntary turnover. The dominant scholarly and practitioner narrative treats these as engineering problems — better measurement, better prediction, better retention. This paper argues that framing is incomplete in a way that matters. Drawing on a conceptual review of research spanning strategic HRM, organizational justice, work design, people analytics, and the sociology of algorithmic control, it advances three claims. First, the technical maturity of AI in performance management and retention is substantially lower than headline accuracy figures imply, because the field's most common evaluation practices reward prediction on convenience datasets rather than decision quality in organizational settings. Second, and more importantly, performance management systems do not influence retention primarily through the accuracy of their outputs; they do so through the meanings employees assign to being measured, ranked, and explained by a machine. Third, because these meaning-making processes are shaped by design choices rather than by algorithmic performance, the pivotal research questions are socio-technical rather than computational. The paper synthesises these arguments into the Measurement–Meaning–Mobility (3M) framework, which specifies how AI-enabled evaluation produces retention outcomes through three sequential mechanisms: what the system captures and omits, how employees interpret being subject to it, and how those interpretations translate into turnover cognitions and behaviour. Eight propositions are developed, together with implications for governance under emerging regulatory regimes including the EU AI Act. The paper concludes with a research agenda emphasising causal identification, employee-side outcomes, and the study of AI-enabled HRM as a system of organizational communication rather than a set of predictive tools.

Keywords—*artificial intelligence; performance management; employee retention; turnover; algorithmic management; organizational justice; people analytics; HR technology governance*

1. INTRODUCTION

Two long-standing problems in human resource management have proved stubbornly resistant to solution. The first is that organizations are poor at evaluating how well their people are doing their jobs. Decades of research on performance appraisal have documented rater bias, leniency, central tendency, halo effects, political distortion, and the near-total absence of evidence that annual review rituals improve anything. The second is that organizations are poor at keeping the people they most want to keep. Turnover research spanning a century has produced a rich catalogue of antecedents without producing much practical ability to forecast, at the level of an individual employee, who will leave and why (Hom, Lee, Shaw, & Hausknecht, 2017).

Artificial intelligence arrives as a proposed answer to both. Continuous data capture promises evaluation that is frequent rather than annual, behavioural rather than impressionistic, and consistent rather than dependent on a particular manager's mood in a particular quarter. Supervised learning on historical HR records promises early warning of departure, converting a

reactive exit-interview function into a proactive retention function. Generative models promise to write the feedback that managers avoid writing. Vendor materials routinely describe attrition models achieving accuracies above ninety percent, and the academic literature contains many such figures.

There is reason for caution. Tambe, Cappelli, and Yakubovich (2019) identified a substantial gap between the promise and the reality of AI in HR, grounding it in four structural constraints: the causal complexity of HR phenomena, the small size of the datasets most employers actually possess, accountability demands arising from fairness and legal obligations, and the possibility that employees will react badly to being managed by algorithm. Seven years on, the technology has advanced considerably while three of those four constraints remain essentially intact. Large language models have not made it easier to know what a good employee is. They have not enlarged the sample size available to a firm with 800 employees and fourteen annual departures. They have not resolved the question of who is accountable when a model

recommends that a person be placed on a performance improvement plan.

The fourth constraint — employee reaction — has received the least systematic theoretical treatment and is, this paper argues, the most consequential for the specific pairing of performance management and retention. The reason is structural rather than incidental. When AI is applied to recruitment, those affected are mostly outsiders to the organization; a rejected applicant's reaction has limited organizational consequences. When AI is applied to performance evaluation, those affected are insiders whose continued presence is the outcome of interest. The evaluative system and the retention outcome are not independent. A performance management system that employees experience as reductive, opaque, or surveillant is not a neutral instrument for measuring retention risk; it is itself a source of retention risk. Any framework that treats AI-enabled performance management and AI-enabled retention prediction as separate applications misses the loop that connects them.

1.1 Purpose and contribution

This paper conducts a conceptual review with the aim of reorganising, rather than merely summarising, what is known. It addresses four questions: (1) What can AI systems actually do in performance management and retention, and how does the empirical evidence compare with the claims made for it? (2) Through what mechanisms does AI-enabled evaluation influence employees' intention to stay or leave? (3) How should the relationship between evaluative technology and retention outcomes be theorised? (4) What follows for governance, practice, and future research?

The intended contributions are three. Substantively, the paper offers a critical reading of the attrition-prediction literature, arguing that its reported performance is inflated by dataset and evaluation conventions that would not survive scrutiny in an applied decision-making context. Theoretically, it develops the Measurement–Meaning–Mobility framework, which integrates organizational justice theory, work design theory, and the sociology of algorithmic control into a single causal chain and specifies eight testable propositions. Practically, it derives a set of design principles for AI-enabled performance management that are consistent both with the employee-reaction evidence and with the obligations emerging from regulatory instruments such as the EU AI Act.

1.2 Structure

Section 2 describes the review approach and its boundaries. Section 3 sets out conceptual foundations: definitions, and the four theoretical lenses the paper draws on. Section 4 reviews AI applications in performance management. Section 5 reviews AI applications in retention, with particular attention to the methodological status of attrition prediction. Section 6 develops the linking mechanisms. Section 7 presents the 3M framework and its propositions. Section 8 addresses governance and regulation. Section 9 draws out practical implications. Section 10 sets out limitations and a research agenda, and Section 11 concludes.

2. REVIEW APPROACH AND BOUNDARIES

This is a conceptual review rather than a systematic one. The distinction matters for what the reader should expect. A systematic review aims at exhaustive, reproducible coverage of a defined literature and reports coverage as a finding in itself. A conceptual review aims at theoretical reorganisation; its

contribution is the argument, and the literature is evidence marshalled in service of it. Several capable systematic reviews of AI in HRM already exist, and there is little to be gained from producing another catalogue of applications. What is missing is an account of how the pieces relate.

The literature drawn on falls into five streams. The first is strategic HRM and people analytics, including critical assessments of the analytics movement's evidentiary base (Angrave, Charlwood, Kirkpatrick, Lawrence, & Stuart, 2016; Marler & Boudreau, 2017). The second is the technical literature on machine learning for employee attrition, drawn on selectively and read as an object of critique rather than a source of accepted findings. The third is organizational behaviour research on reactions to algorithmic decision-making, including experimental work on fairness perceptions (Lee, 2018; Newman, Fast, & Harmon, 2020; Bankins, Formosa, Griep, & Richards, 2022). The fourth is work design and occupational health research, particularly the meta-analytic evidence on electronic performance monitoring (Ravid, White, Tomczak, Miles, & Behrend, 2023; Siegel, König, & Lazar, 2022). The fifth is the sociological and legal literature on algorithmic control and its regulation (Kellogg, Valentine, & Christin, 2020; Ajunwa, 2020).

Three boundaries are worth stating plainly. First, the paper concerns employees in ongoing employment relationships, not applicants and not platform-mediated gig workers. Algorithmic management in gig work is a substantial literature in its own right and its findings do not transfer cleanly, because gig workers typically lack the alternative channels — managers, HR partners, colleagues, grievance procedures — through which employees interpret and contest evaluation. Second, the paper treats retention as the outcome and performance management as the intervention, which reverses the emphasis of most analytics writing. Third, the paper deliberately does not attempt to predict where the technology will be in five years. Technological forecasting in this field has a poor record, and the argument advanced here concerns mechanisms that are relatively robust to changes in model architecture.

3. CONCEPTUAL FOUNDATIONS

3.1 What “AI in HRM” refers to

The term covers a range of technologies that differ in ways relevant to the argument. Four are worth distinguishing.

Descriptive analytics summarise what has happened: dashboards of headcount, tenure distribution, engagement scores, regrettable versus non-regrettable attrition. These are not AI in any meaningful sense, but they are frequently marketed as such, and they form the substrate on which the other three depend.

Predictive models estimate the probability of a future event from historical patterns. Employee attrition prediction is the canonical case. These systems are typically supervised classifiers — logistic regression, random forests, gradient-boosted trees — trained on labelled records of past leavers and stayers.

Measurement and inference systems convert raw behavioural traces into constructs. Communication-pattern analysis that infers collaboration quality, sentiment models that infer engagement from survey free-text or internal messaging, keystroke and application telemetry that infer productivity. These systems are doing something epistemically distinct from

prediction: they are constructing the variables, and the validity of that construction is rarely interrogated.

Generative systems produce natural-language artefacts: draft performance reviews, development plans, coaching prompts, summaries of 360-degree feedback. Their spread into performance management is recent and their consequences largely unstudied.

These categories differ in their failure modes. A predictive model fails by being wrong about the future. A measurement system fails by being wrong about the present — by producing a number that does not correspond to the construct it names. A generative system fails by producing fluent text that is confidently detached from the underlying record. The last two failure modes are harder to detect and, in performance management, more damaging, because there is no ground truth against which to check them.

3.2 Performance management as a socio-technical practice

It is useful to distinguish performance management from performance appraisal. Appraisal is the act of rating. Performance management is the broader system through which organizations set expectations, observe behaviour, provide feedback, allocate rewards, and make decisions about development, advancement, and exit. Appraisal is one component; the system also includes goal-setting, calibration meetings, manager–employee conversations, documentation practices, and the informal understandings that surround all of these.

The distinction matters because AI is typically introduced at the component level while its effects propagate at the system level. A firm may adopt an AI tool solely to help managers draft review text. But the existence of that tool changes what managers do: it lowers the cost of producing documentation, which changes how much documentation exists, which changes what evidence is available in a dispute, which changes how employees behave when they suspect they are being documented. The tool was adopted as a writing aid; it functions as a modification to the organization's disciplinary architecture.

Kellogg, Valentine, and Christin (2020) provide the most useful vocabulary for this. Reviewing the interdisciplinary literature on algorithms at work, they identify six mechanisms of algorithmic control, which they group into direction (restricting and recommending), evaluation (recording and rating), and discipline (replacing and rewarding). Their central observation is that algorithmic control constitutes a form of rational control distinct from the technical and bureaucratic control that characterised the preceding century, in that it is comprehensive, instantaneous, interactive, and opaque. Performance management systems that incorporate AI typically engage all three groups of mechanisms at once, which is why treating them as measurement instruments understates what they do.

3.3 Retention: from turnover prediction to retention capability

Turnover theory offers two broad accounts of why people leave. The traditional attitudinal account, running from March and Simon through the meta-analytic syntheses, treats turnover as the endpoint of a process in which dissatisfaction produces withdrawal cognitions, which produce search behaviour, which produces departure. The unfolding model (Lee & Mitchell, 1994) complicates this by introducing shocks — discrete, jarring events that trigger reconsideration of the employment relationship without necessarily being preceded by declining

satisfaction. Job embeddedness theory (Mitchell, Holtom, Lee, Sablinski, & Erez, 2001) adds a further dimension by locating retention in the web of on- and off-the-job connections that make leaving costly.

Each of these has implications for what an AI system can and cannot do. The attitudinal account is the most tractable: it posits a gradual process with observable intermediate states, some of which leave traces in engagement surveys, communication patterns, or system usage. The shock account is the least tractable, because shocks are by definition unanticipated and frequently exogenous — a partner's job relocation, an unsolicited approach from a recruiter, a public disagreement with a manager. Embeddedness sits in between: some components are observable from HR records (tenure, internal mobility, team stability) and some are not (community ties, family circumstances).

Rubenstein, Eberly, Lee, and Mitchell's (2018) meta-analytic review of turnover antecedents is instructive here. The strongest predictors are attitudinal and relational rather than demographic or administrative. But the variables that dominate published attrition models are overwhelmingly demographic and administrative, because those are what sit in an HRIS. The field has, in effect, optimised prediction over the variables that theory identifies as weakest, because those are the variables that are cheap.

A useful reframing distinguishes *turnover prediction* from *retention capability*. Prediction is the ability to assign a departure probability to an individual. Retention capability is the organizational ability to change that probability. These are related but not identical, and the relationship is weaker than practitioners assume. Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015) make the general version of this point in distinguishing prediction policy problems from causal ones: knowing who will leave is valuable only insofar as one can act on it, and the actions available are determined by causal knowledge that predictive models do not supply.

3.4 Four theoretical lenses

Organizational justice

Justice theory distinguishes distributive justice (are outcomes fair?), procedural justice (is the process that produced them fair?), and interactional justice, itself often split into interpersonal (was I treated with dignity?) and informational (was the process explained to me?) components. Leventhal's procedural criteria — consistency, accuracy, bias suppression, correctability, representativeness, ethicality — map with unusual directness onto arguments made for algorithmic evaluation. Algorithms are consistent by construction. They can be made accurate in a narrow statistical sense. They do not hold personal grudges.

The difficulty is that satisfying some procedural criteria while violating others does not produce a net gain, and it may produce a net loss. Newman, Fast, and Harmon (2020) demonstrate this in what is arguably the most important empirical finding for the present argument. Across a series of studies, they show that algorithmic decisions are perceived as less fair than identical decisions made by humans, that this effect is mediated by perceptions of reductionism — the belief that the algorithm has quantified and decontextualised the person — and that employees evaluated algorithmically report lower organizational commitment, with perceived unfairness mediating that effect. The mechanism is not that people dislike algorithms in the abstract. It is that being reduced to a vector is

experienced as a failure of accuracy, even when the algorithm is more accurate by conventional measures.

This produces a genuine paradox rather than a resolvable tension. Bias suppression and representativeness can pull against one another. A system that eliminates a manager's idiosyncratic preferences also eliminates the manager's knowledge that this quarter's numbers were depressed by a family bereavement.

Job demands–resources and work design

The JD-R tradition treats job characteristics as either demands, which deplete, or resources, which sustain and motivate. AI-enabled performance management alters both. On the demand side, continuous monitoring introduces evaluation apprehension as a chronic rather than episodic condition. On the resource side, timely and specific feedback is a genuine resource, as is the reduction of administrative burden.

Parent-Rocheleau and Parker (2022) formalise this by treating algorithms as work designers, arguing that algorithmic management systematically reshapes job characteristics — autonomy, feedback, task variety, social support — often without anyone having made a deliberate decision about job design. This is the most practically important insight in the work design literature for present purposes. Performance management technology is procured by HR and IT as an administrative system, and its job design consequences are unintended side effects that nobody owns.

Labour process and algorithmic control

The control perspective, developed from Edwards's (1979) account of the workplace as contested terrain and extended by Kellogg and colleagues (2020), treats evaluative technology as a managerial instrument whose deployment provokes worker response. Its analytical value is that it predicts resistance rather than treating it as anomalous. Workers gaming metrics, withholding information from systems, developing folk theories about how the algorithm works, and coordinating informally to manage their visible outputs are not implementation failures; they are the expected behaviour of people with interests in a system that measures them.

HR attributions and signalling

Nishii, Lepak, and Schneider (2008) showed that the effect of an HR practice on employee attitudes depends on the attribution employees make about why the practice exists. A practice attributed to concern for employee wellbeing produces different outcomes than the identical practice attributed to cost reduction or control. This is the lens that explains why the same AI system can succeed in one organization and fail in another. The system does not carry its own meaning. It acquires meaning from the context of its introduction, the stated rationale, the trustworthiness of the party introducing it, and what employees observe about how its outputs are used.

4. ARTIFICIAL INTELLIGENCE IN PERFORMANCE MANAGEMENT

4.1 Measurement and data capture

The foundational application is the automated capture of performance-relevant data. This ranges from the relatively benign — pulling completed-ticket counts from a service desk system — to the highly invasive — keystroke logging, screenshot capture, webcam-based attention inference, sentiment analysis of internal communications.

The evidence on electronic performance monitoring is the clearest and most consistently ignored body of findings in this

entire domain. Ravid, White, Tomczak, Miles, and Behrend (2023) conducted a meta-analysis across 94 independent samples and more than 23,000 participants, and found no evidence that monitoring improves worker performance. Monitoring was associated with increased worker stress regardless of the characteristics of the monitoring. Transparency and lower invasiveness predicted more positive worker attitudes, but they did not convert monitoring into a performance gain. Siegel, König, and Lazar (2022), analysing 70 samples and 233 effect sizes, reached compatible conclusions: small negative associations with job satisfaction, small positive associations with stress and with counterproductive work behaviour, and essentially no association with performance. Their moderator analyses further suggested that attaching performance targets and feedback to monitoring exacerbated rather than mitigated the negative effects, and that laboratory studies understate the size of the effects observed in the field.

Two implications follow. The first is that the intensity of data capture and the quality of performance management are not positively related, and may be negatively related beyond a threshold. The second is that the marginal case for invasive capture is weak on its own terms: if the performance gain is approximately zero, then the stress, attitude, and counterproductive-behaviour costs are being incurred for an informational benefit that has to justify itself entirely on other grounds.

A further, less discussed problem is construct validity. Behavioural telemetry measures activity, and activity is a proxy for performance whose quality varies enormously across roles. In highly routinised work with standardised outputs, activity and performance are reasonably well aligned. In knowledge work they are not, and the gap is not random: it systematically disadvantages work that is reflective, preparatory, mentoring-oriented, or otherwise invisible to instrumentation. Obermeyer, Powers, Vogeli, and Mullainathan (2019), studying a widely deployed healthcare algorithm, demonstrated how a plausible-seeming proxy label can encode systematic bias at scale — in their case, using healthcare costs as a proxy for healthcare need, which understated the needs of Black patients because less had historically been spent on them. The structural lesson transfers directly. An attrition or performance model trained on a proxy inherits whatever inequities produced the proxy.

4.2 Rating, calibration, and the correction of human bias

A second application uses models to adjust or audit human ratings: detecting leniency patterns, flagging demographic disparities in rating distributions, normalising across managers with different standards. This is among the more defensible uses, because it targets a well-documented problem and because the algorithm operates on the evaluation rather than on the person.

The claim that automation eliminates bias requires care. Ajunwa (2020) characterises this as the paradox of automation as an anti-bias intervention: systems introduced to reduce discrimination can launder it instead, giving biased outcomes the appearance of neutrality while removing the human decision-maker who might have been held accountable. Barocas and Selbst (2016) set out the mechanisms in detail, showing how discriminatory outcomes can arise from target-variable definition, training-label bias, feature selection, and proxy variables, without anyone intending discrimination and often without any protected attribute appearing in the model.

Where models correct human bias, they do so relative to a defined statistical criterion. Where the bias is encoded in the criterion itself — in what counts as good performance — the model reproduces it with greater consistency than the humans it replaced. Consistency is not neutrality. A consistently biased system is, from the perspective of an affected employee, worse than an inconsistently biased one, because inconsistency at least offers the possibility of a different outcome next time.

4.3 Generative AI and the production of feedback

The newest and least understood application is the use of large language models to draft performance reviews, development plans, and coaching prompts. The appeal is straightforward: writing good feedback is difficult, time-consuming, and unpleasant, and managers avoid it.

Three concerns deserve theoretical attention, none of which has been adequately studied.

The first is the decoupling of judgement from articulation. The effort of writing feedback is not purely transaction cost; it is partly the cognitive process by which a manager works out what they actually think. A manager who accepts a generated draft may never have formed the judgement the draft expresses. The resulting document is fluent, well-structured, and epistemically hollow — and hollowness in feedback is detectable by its recipient in ways that matter.

The second is interactional justice. Research on justice perceptions consistently finds that respectful treatment and individualised consideration are distinct contributors to fairness judgements, separable from procedural and distributive concerns. Feedback that the recipient believes was machine-generated may be read as a signal that the manager did not consider them worth the time, which damages precisely the relational dimension that generative tools were not designed to touch.

The third is homogenisation. Language models converge on characteristic registers. At scale, generated feedback across an organization will exhibit reduced variance in phrasing, structure, and emphasis. Since performance documentation is an input to promotion and termination decisions, compression of the signal reduces the discriminative value of the very records those decisions rely on.

The broader evidence on generative AI at work is more encouraging on productivity than on evaluation. Studies of generative assistance in customer service and in professional writing have documented meaningful gains in output and quality (Brynjolfsson, Li, & Raymond, 2025; Noy & Zhang, 2023). Those findings concern AI as a tool used *by* workers. They say nothing about AI used *on* workers, and the conflation of the two is one of the more consequential errors in current practitioner discourse.

4.4 What the evidence supports

Assembling the above, a defensible reading is as follows. AI can reduce administrative burden in performance management, and this is a real benefit. It can surface patterns in rating distributions that human reviewers miss, and this is a real benefit. It can make feedback more frequent, though frequency and quality are not the same thing. It has not been shown to improve the accuracy of performance evaluation against any external criterion, largely because no such criterion exists. It has not been shown to improve performance itself. And the monitoring infrastructure on which the more ambitious applications depend has a meta-analytic evidence base pointing

towards stress and attitudinal costs without compensating performance gains.

5. ARTIFICIAL INTELLIGENCE IN EMPLOYEE RETENTION

5.1 The state of attrition prediction

Machine learning applied to employee attrition is now a large literature. Systematic reviews report broad application of logistic regression, decision trees, random forests, support vector machines, gradient boosting, and neural networks, with ensemble methods and class-rebalancing techniques such as SMOTE frequently improving reported metrics (Talebi et al., 2025; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Reported accuracies commonly exceed 90%, and figures above 98% appear regularly.

These figures should be treated with considerable scepticism, for five reasons that compound.

Dataset provenance

A substantial proportion of published attrition studies use the IBM HR Analytics Employee Attrition dataset, which comprises 1,470 records and is synthetic — it was constructed by data scientists for demonstration purposes, not collected from an operating organization. Relationships in it are artefacts of its generation process. A model achieving high accuracy on this dataset has demonstrated that it can recover a synthetic data-generating process. It has demonstrated nothing about real employees. Systematic reviews of the area note the difficulty of obtaining reliable and relevant real-world data as a central limitation, which is precisely why the synthetic dataset has become a *de facto* benchmark.

Class imbalance and the accuracy metric

Annual voluntary turnover in most organizations is between 5% and 20%. A classifier that predicts nobody will leave is 85% accurate at a 15% turnover rate. Reported accuracy is therefore close to uninformative unless the base rate is stated and a naive baseline provided, which published work frequently omits. Precision, recall, and calibration on the minority class are the relevant quantities, and it is in those terms that real-world attrition models are typically mediocre.

Resampling contamination

Synthetic oversampling techniques are applied widely. Where oversampling is applied before the train–test split rather than within training folds, synthetic points derived from test-set observations appear in training, and the reported test performance is inflated by leakage. This error is common enough in the applied literature that reported figures cannot be taken at face value without examining the pipeline.

Temporal structure

Attrition is a time-to-event problem being forced into a binary classification frame. Whether an employee will leave is less useful than when, and random train–test splits on a cross-sectional snapshot ignore the temporal ordering entirely, permitting models to learn from futures that would not be available at prediction time. Survival analysis and its machine-learning extensions are the appropriate tool and are conspicuously underused.

Sample size at the level of the deploying firm

This is the constraint Tambe and colleagues (2019) identified and it remains decisive. A firm of 1,000 employees with 12% annual turnover generates roughly 120 positive cases per year. After excluding involuntary separations, retirements, and non-regrettable departures, the number of cases of actual

interest may be forty. Learning a stable, non-overfitted model of voluntary departure among high performers from forty examples, across dozens of candidate features, is not a problem that better algorithms solve. It is a problem of information.

5.2 From prediction to intervention

Suppose a model is well-calibrated and identifies a genuinely elevated risk for a particular employee. What follows?

The gap between prediction and action is where most retention analytics programmes fail, and the failure is conceptual rather than technical. A predictive model ranks employees by probability. It does not indicate what would change that probability, because it was not trained to answer that question. A feature that is highly predictive — time since last promotion, say — may be predictive because it causes dissatisfaction, or because it is a marker of some third factor, or because employees who have already decided to leave stop pursuing promotion. These three cases call for entirely different responses, and the model cannot distinguish them. Explainability methods such as SHAP and LIME (Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016) are frequently offered as the solution, but they decompose a model's predictions, not the world's causal structure. A feature attribution says what the model used, not what an intervention would do.

Rudin (2019) argues that in high-stakes decision settings, post-hoc explanation of opaque models is the wrong response and interpretable models should be used instead. Retention decisions affecting compensation, development, and advancement are high-stakes in exactly the sense intended. The argument has additional force in HR, where the legal and regulatory environment increasingly requires explanations that are accurate rather than merely plausible.

The constructive alternative is the one Tambe and colleagues (2019) propose: causal reasoning, randomisation and experimentation, and employee contribution. Organizations serious about retention should be running experiments — randomised or quasi-randomised evaluations of retention interventions — rather than refining predictions. A 60%-accurate model attached to an intervention known to work is worth more than a 90%-accurate model attached to an intervention of unknown effect.

5.3 The ethics of pre-emptive labelling

Attrition prediction carries a distinctive ethical hazard that has no close analogue elsewhere in HR analytics: it generates information about an employee's presumed future intentions, and that information is available to the organization before, and possibly instead of, the employee's own disclosure.

Several problems follow. Self-fulfilling dynamics are the most direct. An employee flagged as a flight risk may be excluded from long-horizon projects, passed over for development investment, or treated with the guardedness accorded to someone presumed to be leaving. The prediction then contributes to the outcome it forecast. Because these decisions are made informally and are rarely recorded, the effect is largely invisible to the analytics function that generated the flag.

Differential treatment raises a distributional question that is usually left implicit. Retention resources are finite, and a model that ranks by risk will, in combination with a value ranking, direct counter-offers and development opportunities towards

high-risk high-value employees. This is rational from the firm's perspective and constitutes, from the employee's perspective, a system in which credible threat of departure is rewarded and loyalty is not. Over time, this teaches employees a lesson about how the organization allocates attention.

Consent and reasonable expectation are also at issue. Employees supply data to their employer for identifiable purposes: payroll, benefits administration, access management, collaboration. Repurposing that data to infer intentions they have not expressed exceeds any reasonable construction of the original purpose, and it is on exactly this point that data protection law is most likely to bite.

Finally, there is an asymmetry worth naming. An employee who is considering leaving typically has strong incentives not to reveal it. A system that infers the intention anyway converts a domain the employee controlled into one the employer controls. Whatever its efficiency merits, this is a reallocation of informational power within the employment relationship, and it should be discussed as such rather than as a technical enhancement.

6. THE LINKING MECHANISMS

The preceding sections treated performance management and retention separately, in line with how the literature is organised. This section argues that the separation is untenable and specifies four mechanisms through which AI-enabled evaluation produces retention outcomes.

6.1 Justice perceptions

The first and best-evidenced mechanism runs through perceived fairness. Newman, Fast, and Harmon's (2020) demonstration that algorithmic evaluation lowers perceived procedural fairness via perceived reductionism, and that this in turn lowers organizational commitment, supplies the core of the pathway. Related work has found that workers' justice perceptions differ systematically between human and AI decision-makers in HRM contexts (Bankins, Formosa, Griep, & Richards, 2022), and that the transparency of algorithmic management relates to justice perceptions and, in turn, to intention to quit (Bujold, Parent-Rochelleau, & Gaudet, 2022). Qin, Jia, Luo, Liao, and Huang (2023) examined perceived fairness of human versus AI evaluation specifically in performance appraisal, finding that the comparison is more conditional than a simple aversion account would predict.

The conditionality is the interesting part. The evidence does not support a blanket claim that employees reject algorithmic evaluation. It supports the narrower claim that algorithmic evaluation is penalised when it is experienced as decontextualising, and that this penalty can be attenuated by transparency, by voice, and by the presence of meaningful human review.

6.2 Work design and strain

The second mechanism runs through job characteristics. AI-enabled performance management alters autonomy, feedback, and social support, frequently without deliberate design (Parent-Rochelleau & Parker, 2022). Where monitoring intensifies, the meta-analytic evidence points to elevated stress independent of how the monitoring is configured (Ravid et al., 2023). Strain, in turn, is among the more robust predictors of withdrawal.

The pathway from surveillance to turnover therefore does not require the employee to hold any particular view about

algorithms. It operates through accumulated strain and reduced job resources, and it would operate the same way if the monitoring were done by a human supervisor with unlimited attention.

6.3 Psychological contract and attribution

The third mechanism runs through interpretation. Employees hold beliefs about the reciprocal obligations of the employment relationship, and the introduction of an evaluative technology is read as information about the employer's side of that bargain. Following Nishii and colleagues (2008), the attribution made about why the system was introduced determines the direction of its effect. A system introduced with a stated rationale of improving feedback quality, accompanied by consultation and by observable restraint in how outputs are used, supports a commitment attribution. The same system introduced without explanation, discovered by employees rather than announced to them, and followed by visible disciplinary action supports a control attribution.

This mechanism explains a pattern that otherwise looks like noise in the literature: the same technology producing opposite outcomes across organizations. The technology is not the operative variable. The attribution is.

6.4 Resistance and metric displacement

The fourth mechanism is behavioural rather than attitudinal, and it affects the validity of the system rather than the employee's intention directly. Where evaluation is algorithmic and its criteria are inferable, employees optimise against the criteria. Kellogg and colleagues (2020) document the range of responses, from individual gaming to collective action.

Metric displacement has a specific consequence for the loop this paper describes. As employees optimise for measured indicators, those indicators decouple from the underlying performance construct. The data feeding the attrition model degrades. Model performance declines, is attributed to model specification, and prompts the collection of more data, which intensifies monitoring, which increases strain and accelerates gaming. The system's own operation undermines its informational basis.

7. THE MEASUREMENT–MEANING–MOBILITY FRAMEWORK

The four mechanisms can be organised into a sequential framework with a feedback loop. The 3M framework holds that AI-enabled performance management influences retention through three stages: measurement (what the system captures, omits, and constructs), meaning (how employees interpret being subject to it), and mobility (how those interpretations translate into turnover cognitions and behaviour). Outcomes at the third stage feed back into the first by altering the data the system receives.

MEASUREMENT	MEANING	MOBILITY
Construct coverage	Justice perceptions	Turnover intention
Capture intensity	HR attributions	Search behaviour
Decision authority	Strain and resources	Voluntary exit
<i>Feedback loop: gaming and metric displacement degrade the data on which measurement depends</i>		

Fig. 1. Figure 1. The Measurement–Meaning–Mobility framework. The three stages form the primary causal sequence; mobility-stage behaviour feeds back to degrade the measurement stage. The measurement-to-meaning transition is moderated by transparency, voice, and human review; the meaning-to-mobility transition is moderated by job embeddedness and perceived labour market alternatives.

7.1 Stage 1: Measurement

Three properties of the measurement stage carry the framework's predictions.

Construct coverage is the proportion of performance-relevant contribution that the system captures. Coverage is never complete, and incompleteness is systematic rather than random: it disadvantages contributions that are relational, preparatory, or otherwise unlogged.

Capture intensity is the granularity and continuity of observation, ranging from periodic outcome measures to continuous behavioural telemetry.

Decision authority is the degree to which model outputs determine consequential decisions, ranging from advisory input to automated determination. Shrestha, Ben-Menahem, and von Krogh (2019) offer a useful typology of decision-making structures that combine human and algorithmic input, distinguishing full delegation, hybrid sequential arrangements, and aggregated human–AI decision-making.

PROPOSITION 1

Lower construct coverage increases perceived reductionism, which mediates a negative relationship between AI-enabled evaluation and perceived procedural justice. The effect is strongest for employees whose contributions are least amenable to instrumentation.

PROPOSITION 2

Capture intensity relates to strain monotonically and to perceived informational justice non-monotonically: moderate intensity improves feedback sufficiency, while high intensity is experienced as surveillance and degrades it.

PROPOSITION 3

The negative effect of AI-enabled evaluation on justice perceptions increases with decision authority and is attenuated where human review is both present and demonstrably capable of overriding the model.

7.2 Stage 2: Meaning

The meaning stage comprises the interpretive and affective states through which measurement produces behavioural consequences: justice perceptions across all four dimensions, the attribution made about the system's purpose, and the net balance of demands and resources.

The critical theoretical claim is that this stage is not a transmission mechanism but a transformation. The same measurement configuration yields different meaning-stage states depending on the moderators shown in Figure 1. This is why implementation process, rather than system capability, is the principal lever available to practice.

PROPOSITION 4

The relationship between AI-enabled evaluation and justice perceptions is moderated by attribution: where employees attribute adoption to commitment-oriented motives, the relationship is attenuated or reversed; where they attribute it to control- or cost-oriented motives, it is amplified.

PROPOSITION 5

Procedural voice — the opportunity to contest, contextualise, or supplement algorithmic evaluation — attenuates the negative effect of reductionism on justice perceptions, and does so more effectively than explanation alone.

Proposition 5 separates two interventions that are usually bundled under “transparency”. Explanation tells the employee why the system produced an output. Voice gives the employee a channel to change it. The reductionism mechanism concerns information the system did not consider; explanation does not address that, while voice does. This suggests that the field's heavy investment in explainability may be targeting the less effective of the two available remedies.

7.3 Stage 3: Mobility

The mobility stage covers turnover cognitions, search behaviour, and voluntary exit. Its relationship to the meaning stage is well established in turnover theory and requires no novel argument, but two moderators deserve specification.

PROPOSITION 6

The effect of meaning-stage states on voluntary exit is moderated by job embeddedness and by perceived labour market alternatives, such that identical deteriorations in justice perceptions produce larger turnover effects among less embedded employees and in tighter external labour markets.

PROPOSITION 7

Because the employees most able to act on dissatisfaction are those with the strongest external options, the turnover produced through this pathway is disproportionately regrettable. AI-enabled performance management that degrades justice perceptions will therefore worsen the composition of turnover even where it does not raise its overall rate.

Proposition 7 has a practical edge worth stating explicitly. An organization evaluating an AI performance management deployment by monitoring aggregate turnover may see no change, or even a decline, while the quality of who stays deteriorates. Aggregate turnover is the wrong metric for detecting this failure mode.

7.4 The feedback path

PROPOSITION 8

Where decision authority is high and construct coverage is low, employees engage in metric-directed behaviour that reduces the criterion validity of captured measures over time, degrading model performance and creating pressure for further capture intensity. This constitutes a self-reinforcing loop rather than an equilibrium.

Proposition 8 identifies what may be the framework's most consequential implication for practice: the failure mode of AI-enabled performance management is not a single bad decision but a gradual, self-concealing decline in the informational quality of the system, accompanied by intensifying data collection justified by that very decline.

TABLE 1

Stage	Core constructs	Principal design levers	Observable failure signature
Measurement	Construct coverage; capture intensity; decision authority	What is instrumented; what is excluded; where human judgement is binding	Systematic under-recognition of relational and preparatory work
Meaning	Procedural, informational, interpersonal justice; HR attributions; demands-resources balance	Rationale and consultation; explanation; contestation channels	Divergence between engagement scores and system-derived indicators
Mobility	Turnover intention; search behaviour; voluntary exit	Embeddedness-building practice; internal mobility	Stable aggregate turnover with deteriorating quality

Feedback	Metric displacement; criterion contamination	Periodic revalidation; criterion rotation; audit against unmeasured outcomes	of leavers
			Declining performance attributed to specification rather than gaming model

Fig. 2. Table 1. Stages of the 3M framework, with associated constructs, design levers, and the observable signatures by which failure at each stage can be detected.

8. GOVERNANCE AND REGULATION

The regulatory environment for AI in employment has moved faster than the academic literature and now imposes obligations that intersect directly with the mechanisms above.

8.1 The EU AI Act

Regulation (EU) 2024/1689 adopts a risk-based structure, and Annex III designates AI systems used in employment and workers' management as high-risk. The designation covers recruitment and selection; decisions affecting the terms of the work relationship, including promotion and termination; the allocation of tasks based on individual behaviour or personal traits; and the monitoring and evaluation of performance and behaviour. AI-enabled performance management is squarely within scope, and so, in most configurations, is attrition prediction used to inform decisions about individuals.

High-risk classification triggers substantive obligations on providers: risk management across the system lifecycle, data governance ensuring that training, validation, and test data are relevant and sufficiently representative, technical documentation, record-keeping, transparency, human oversight, accuracy, robustness, and cybersecurity. Deployers carry their own duties, and one is of particular significance for this paper's argument: employers must inform workers' representatives and affected workers before putting a high-risk system into service at the workplace.

That requirement converts a matter that the framework identifies as the principal moderator of employee response — the communicative context of adoption — into a legal obligation. Organizations required to announce and explain will find themselves doing, under compulsion, the thing the evidence suggests they should do voluntarily. The application date for Annex III high-risk obligations has been subject to revision during the Act's implementation, and organizations should verify the current timetable rather than rely on figures reported at any given moment.

8.2 Other instruments

New York City's Local Law 144 requires annual independent bias audits of automated employment decision tools and notice to candidates and employees, with published audit summaries. Illinois regulates the use of AI analysis in video interviews. Colorado's 2024 legislation addressing algorithmic discrimination in consequential decisions includes employment within scope. Beyond AI-specific instruments, the GDPR's provisions on automated individual decision-making and on purpose limitation apply to European employment data, and purpose limitation is the provision most directly in tension with the repurposing of operational data for attrition inference.

Cappelli and Rogovsky (2023), writing for the International Labour Organization, situate these developments within a broader question about whether AI-enabled HRM is compatible with a human-centred agenda for work — a framing that treats the regulatory question as one about the organization of work rather than about technology compliance.

8.3 What good governance requires

Compliance and good governance are not the same, and the framework suggests where they diverge. Five elements follow from the analysis:

- Purpose specification before capability. The question “what decision will this improve, and how will we know?” should precede procurement. Systems adopted because they were available accumulate uses that nobody authorised.
- Construct validity review. Someone should be accountable for the question of whether measured indicators correspond to the performance construct, and for documenting what the system systematically fails to see.
- Contestation as a first-class feature. Following Proposition 5, employees should be able to supplement the record with context the system could not capture, and that supplement should demonstrably affect outcomes.
- Human review with genuine authority. Review that never overturns is not oversight. Override rates should be monitored; a rate near zero indicates either a perfect model or an ineffective review process, and the second is far more likely.
- Evaluation against employee-side outcomes. Deployments should be assessed against justice perceptions, strain, and the composition of turnover, not only against model metrics and administrative efficiency.

9. IMPLICATIONS FOR PRACTICE

Several practical conclusions follow from the argument, some of which run against current practitioner consensus.

- Invest in intervention knowledge rather than prediction accuracy. The marginal value of moving an attrition model from 78% to 84% AUC is small compared with the value of knowing, from a properly designed evaluation, whether a particular retention intervention works. Organizations should run experiments.
- Treat monitoring intensity as a cost, not a capability. The meta-analytic evidence gives no support for the proposition that more observation produces better performance, and consistent support for the proposition that it produces more stress. Capture should be justified against a specific decision it improves.
- Announce before deploying, and state the rationale honestly. The attribution mechanism means that the introduction is not a communications exercise appended to the implementation; it is a substantial part of the implementation. A rationale that employees find implausible is worse than no rationale, because it introduces a credibility problem on top of the original concern.
- Build contestation channels, not just explanations. Explanation addresses informational justice. Contestation addresses reductionism, which is the mechanism the evidence identifies as primary.
- Keep attrition risk scores away from line managers. Individual-level risk scores in the hands of managers

create the conditions for self-fulfilling treatment while offering little that aggregate analysis does not. Risk modelling is most defensible at the level of segments, teams, and policies, where it informs systemic intervention without labelling individuals.

- Monitor the composition of turnover, not only its rate. Proposition 7 implies that the characteristic failure of these systems is invisible in aggregate turnover statistics.

10. LIMITATIONS AND FUTURE RESEARCH

10.1 Limitations

This is a conceptual paper and inherits the limitations of the form. Its propositions are argued rather than tested. Its selection of literature, while systematic in intent, reflects judgements about relevance that another author might make differently. The framework privileges employee-side mechanisms, and a reader more persuaded by efficiency arguments would weight the evidence differently — though it should be noted that the efficiency case rests substantially on vendor claims and on studies whose methodological limitations Section 5 documents.

The evidence base itself constrains what can be claimed. Much of the fairness-perception research is experimental, conducted with vignettes and non-employee samples, and its external validity in ongoing employment relationships is not established. The monitoring meta-analyses aggregate studies conducted largely before contemporary AI capabilities existed, and while the psychological mechanisms should be relatively stable, this is an assumption rather than a finding. Research on generative AI in performance management barely exists, and Section 4.3's claims are correspondingly speculative.

The framework is also, by design, silent on cross-cultural variation. Justice perceptions, expectations of managerial discretion, and attitudes towards workplace monitoring differ substantially across national and institutional contexts, and a framework developed principally from North American and Western European research should not be assumed to transfer.

10.2 A research agenda

- Field studies with employee-side outcomes. The most serious gap is the near-absence of field research measuring what happens to justice perceptions, strain, and turnover when organizations deploy these systems. Staged rollouts offer natural quasi-experimental opportunities that are currently being wasted.
- Causal identification in retention analytics. The prediction-to-intervention gap identified in Section 5.2 is a research problem before it is a practice problem. Work applying causal inference methods to retention — instrumental variables, regression discontinuity around promotion and compensation thresholds, and randomised evaluation of interventions — would be more valuable than further refinement of classifiers.
- Direct tests of Proposition 5. Whether voice outperforms explanation in attenuating reductionism effects is an empirically tractable question with immediate design implications, and it is currently unanswered.
- The generative AI question. Whether employees can detect AI-generated feedback, whether detection affects its reception, and whether the availability of generation

tools changes the quality of managerial judgement are all open and all consequential.

- Longitudinal study of the feedback loop. Proposition 8 describes a dynamic process that cross-sectional research cannot observe. Panel designs tracking measurement validity, gaming behaviour, and capture intensity over multiple years would test whether the loop operates as theorised.
- Comparative institutional analysis. The divergence between jurisdictions with strong worker-representation requirements and those without offers a natural comparison for testing whether the consultation obligations emerging from regulation actually produce the attribution effects the framework predicts.

11. CONCLUSION

The literature on AI in human resource management has been organised around a question about capability: what can these systems do, and how well? This paper has argued that for the specific domains of performance management and retention, capability is the less important question, and that organising around it has produced a body of work that is technically sophisticated and practically thin.

The more important question concerns meaning. Performance management systems influence retention because employees interpret them, and those interpretations are shaped by design and implementation choices that have almost nothing to do with model architecture. An organization that deploys a mediocre model with genuine consultation, meaningful contestation rights, and restraint in how outputs are used will likely see better retention outcomes than one deploying an excellent model as an instrument of unannounced surveillance. The 3M framework attempts to make this claim precise enough to test.

There is a broader point underneath. The appeal of AI in this domain rests partly on a hope that the difficulty of managing people is a measurement problem, and that sufficiently good instrumentation will dissolve it. The evidence reviewed here does not support that hope. Evaluating contribution requires judgements about value that no dataset contains. Retaining people requires understanding what they want, which they will disclose only to an organization they have reason to trust. Better instruments do not substitute for either, and instruments deployed as though they might can make both harder. The most useful contribution AI can make to performance management and retention may turn out to be a modest one: removing administrative work that consumes the time managers need for the parts of the job that remain irreducibly human.

REFERENCES

- [1] Ajunwa, I. (2020). The paradox of automation as anti-bias intervention. *Cardozo Law Review*, 41(5), 1671–1742.
- [2] Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal*, 26(1), 1–11. <https://doi.org/10.1111/1748-8583.12090>
- [3] Bakker, A. B., & Demerouti, E. (2017). Job demands–resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology*, 22(3), 273–285. <https://doi.org/10.1037/ocp0000056>
- [4] Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Information Systems Frontiers*, 24(3), 857–875. <https://doi.org/10.1007/s10796-021-10223-8>
- [5] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732.
- [6] Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.
- [7] Bujold, A., Parent-Rochelleau, X., & Gaudet, M. C. (2022). Opacity behind the wheel: The relationship between transparency of algorithmic management, justice perception, and intention to quit among truck drivers. *Computers in Human Behavior Reports*, 8, 100245. <https://doi.org/10.1016/j.chbr.2022.100245>
- [8] Cappelli, P., & Rogovsky, N. (2023). *Artificial intelligence in human resource management: A challenge for the human-centred agenda?* (ILO Working Paper No. 95). International Labour Organization.
- [9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- [10] Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, 86(3), 425–445. <https://doi.org/10.1037/0021-9010.86.3.425>
- [11] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- [12] Edwards, R. (1979). *Contested terrain: The transformation of the workplace in the twentieth century*. Basic Books.
- [13] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- [14] Gal, U., Jensen, T. B., & Stein, M.-K. (2020). Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. *Information and Organization*, 30(2), 100301. <https://doi.org/10.1016/j.infoandorg.2020.100301>
- [15] Giermnd, L. M., Strich, F., Christ, O., Leicht-Deobald, U., & Redzepi, A. (2022). The dark sides of people analytics: Reviewing the perils for organisations and employees. *European Journal of Information Systems*, 31(3), 410–435. <https://doi.org/10.1080/0960085X.2021.1927213>
- [16] Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102(3), 530–545. <https://doi.org/10.1037/apl0000103>
- [17] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410. <https://doi.org/10.5465/annals.2018.0174>
- [18] Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5), 491–495. <https://doi.org/10.1257/aer.p20151023>
- [19] Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, 123, 106878. <https://doi.org/10.1016/j.chb.2021.106878>
- [20] Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1). <https://doi.org/10.1177/2053951718756684>
- [21] Lee, T. W., & Mitchell, T. R. (1994). An alternative approach: The unfolding model of voluntary employee turnover. *Academy of Management Review*, 19(1), 51–89. <https://doi.org/10.2307/258835>
- [22] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
- [23] Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of HR analytics. *International Journal of Human Resource Management*, 28(1), 3–26. <https://doi.org/10.1080/09585192.2016.1244699>
- [24] Meijerink, J., Boons, M., Keegan, A., & Marler, J. (2021). Algorithmic human resource management: Synthesizing developments and cross-disciplinary insights on digital HRM. *International Journal of Human Resource Management*, 32(12), 2545–2562. <https://doi.org/10.1080/09585192.2021.1925326>
- [25] Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal*, 44(6), 1102–1121. <https://doi.org/10.2307/3069391>

- [26] Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160, 149–167. <https://doi.org/10.1016/j.obhdp.2020.02.005>
- [27] Nishii, L. H., Lepak, D. P., & Schneider, B. (2008). Employee attributions of the “why” of HR practices: Their effects on employee attitudes and behaviors, and customer satisfaction. *Personnel Psychology*, 61(3), 503–545. <https://doi.org/10.1111/j.1744-6570.2008.00121.x>
- [28] Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192. <https://doi.org/10.1126/science.adh2586>
- [29] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- [30] Parent-Rochelleau, X., & Parker, S. K. (2022). Algorithms as work designers: How algorithmic management influences the design of jobs. *Human Resource Management Review*, 32(3), 100838. <https://doi.org/10.1016/j.hrmr.2021.100838>
- [31] Qin, S., Jia, N., Luo, X., Liao, C., & Huang, Z. (2023). Perceived fairness of human managers compared with artificial intelligence in employee performance evaluation. *Journal of Management Information Systems*, 40(4), 1039–1070. <https://doi.org/10.1080/07421222.2023.2267319>
- [32] Ravid, D. M., Tomczak, D. L., White, J. C., & Behrend, T. S. (2020). EPM 20/20: A review, framework, and research agenda for electronic performance monitoring. *Journal of Management*, 46(1), 100–126. <https://doi.org/10.1177/0149206319869435>
- [33] Ravid, D. M., White, J. C., Tomczak, D. L., Miles, A. F., & Behrend, T. S. (2023). A meta-analysis of the effects of electronic performance monitoring on work outcomes. *Personnel Psychology*, 76(1), 5–40. <https://doi.org/10.1111/peps.12514>
- [34] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- [35] Rubenstein, A. L., Eberly, M. B., Lee, T. W., & Mitchell, T. R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, 71(1), 23–65. <https://doi.org/10.1111/peps.12226>
- [36] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [37] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). <https://doi.org/10.1145/3287560.3287598>
- [38] Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83. <https://doi.org/10.1177/0008125619862257>
- [39] Siegel, R., König, C. J., & Lazar, V. (2022). The impact of electronic monitoring on employees' job satisfaction, stress, performance, and counterproductive work behavior: A meta-analysis. *Computers in Human Behavior Reports*, 8, 100227. <https://doi.org/10.1016/j.chbr.2022.100227>
- [40] Talebi, S., et al. (2025). Machine learning approaches for predicting employee turnover: A systematic review. *Engineering Reports*, 7(8), e70298. <https://doi.org/10.1002/eng.2.70298>
- [41] Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4), 15–42. <https://doi.org/10.1177/0008125619867910>
- [42] Tursunbayeva, A., Pagliari, C., Di Lauro, S., & Antonelli, G. (2022). The ethics of people analytics: Risks, opportunities and recommendations. *Personnel Review*, 51(3), 900–921. <https://doi.org/10.1108/PR-12-2019-0680>
- [43] Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review. *International Journal of Human Resource Management*, 33(6), 1237–1266. <https://doi.org/10.1080/09585192.2020.1871398>

Note on sources. Every reference above corresponds to a real publication. Before submission, verify each entry against the original record — page numbers and volume details for very recent items in particular can shift between online-first and final publication. Any claim you are not prepared to defend from the source should be removed rather than softened.