



Graduated Autonomy for Artificial Intelligence Agents in Human Capital Management: A Systematic Review and a Consequence-Calibrated Governance Framework

Shivanand R. Koppalkar

Researcher, Walsh College, United States

Corresponding author: Shivanand R. Koppalkar, skoppalk@mail.walshcollege.edu, us@innovate-software.com. ORCID:
<https://orcid.org/0009-0009-7843-2595>.

Abstract

Human capital management platforms are starting to embed software agents built on large language models. Unlike earlier analytics tools, these agents can plan a sequence of steps, call system functions and change employee records. Artificial intelligence is therefore moving from advising managers to acting for them. Most governance guidance, however, was written for systems that recommend rather than act. This study asks how organisations should match the autonomy of such agents to the consequences of the people decisions they touch. A systematic review was conducted and reported using the PRISMA 2020 guideline. Eighteen structured searches of a multi-source scholarly index returned 180 records. After duplicate removal and eligibility screening, 55 peer-reviewed sources were retained, and four foundational and regulatory sources were added. Thematic synthesis produced seven findings. Algorithmic control of work is already extensive. Bias in hiring tools is persistent and partly organisational. Human review is weaker than assumed, because reviewers often defer to machine advice. Responsibility tends to diffuse as systems gain autonomy. Employee trust depends on fair process and visible human involvement. Language-model agents add new failure modes, such as fabricated content and hidden instructions planted in documents. Finally, current governance frameworks rarely specify controls for individual delegated actions. Building on these findings, the study proposes a consequence-calibrated governance framework. It sets an autonomy ceiling for each class of people decision and links every ceiling to five control layers. Six testable propositions and practical guidance for human resource and technology leaders are offered.

Keywords: *algorithmic management; automation bias; meaningful human control; large language models; people analytics*

1. Introduction

Human capital management (HCM) systems run some of the most consequential processes in any organisation. They record who is hired, how much each person is paid, how performance is rated and who leaves. Artificial intelligence (AI) has been part of this domain for more than a decade. Reviews show its use across recruitment, training, performance management and workforce planning [1]–[3]. They also show a steady gap between promise and practice. Tambe et al. [4] traced this gap to four problems: the complexity of human resource (HR) outcomes, small data sets, fairness and legal accountability, and negative employee reactions.

A new class of technology now raises the stakes. Large language models (LLMs) can be wrapped in software that perceives a goal, plans steps, remembers context and uses tools [5], [6]. Such systems are often called



agentic AI. They are marked by autonomy, adaptivity, proactiveness and a degree of decision agency [7]. A multi-expert analysis expects these agents to decentralise decision making and to shift tasks from people to software. The same analysis warns of unclear attribution and shared accountability [8]. Van Esch [9] goes further. He argues that the next phase of AI will be defined less by smarter agents and more by AI acting as an organising layer that reallocates decision rights.

In HCM, this shift from advice to action matters a great deal. An agent that drafts an offer letter, adjusts a pay element, reschedules a shift or closes a job requisition acts directly on a person's livelihood. Earlier research on algorithmic management studied how software directs, evaluates and disciplines workers [10], [11]. Research on AI-assisted HR decisions studied tools that score or rank people [4], [12]. In both streams, the dominant mental model was an algorithm that recommends and a human who decides. Agentic systems break that model. The agent may now take the action itself, or prepare it so fully that the human review becomes a formality.

Governance thinking has not yet caught up with this change. Organisational AI governance research focuses on principles, structures and lifecycle processes [13], [14]. Widely used standards, such as the risk management framework of the United States National Institute of Standards and Technology (NIST), are deliberately general [15]. The European Union (EU) Artificial Intelligence Act classifies many employments uses of AI as high-risk and requires human oversight [16]. Yet it does not tell an HR leader which agent actions may run without a person. Gluszak and Gluszak [17] recently argued that delegation, rather than technical architecture, is the missing variable in agentic AI governance. Their framework is cross-industry. To the author's knowledge, no review has yet linked the autonomy of AI agents to the specific consequences of HCM decisions for workers.

This study addresses that gap through a systematic review and a conceptual framework. It poses three research questions (RQs):

- RQ1. What capabilities and risks of agentic AI are relevant to human capital management?
- RQ2. Which governance mechanisms does the literature identify for algorithmic and AI-assisted people decisions, and how well do they extend to agents that act?
- RQ3. How can organisations calibrate the autonomy of AI agents to the consequences of HCM decisions?

The study makes three contributions. First, it synthesises evidence from HR management, information systems, human factors, ethics and computer science, fields that rarely cite each other on this topic. Second, it proposes the Consequence-Calibrated Agency (C2A) framework. The framework sets an autonomy ceiling for each class of HCM decision and ties each ceiling to five control layers. Third, it states six testable propositions for future empirical work. Section 2 sets out the conceptual background. Section 3 describes the method. Section 4 reports the results. Section 5 interprets them and presents the framework. Section 6 concludes.

2. Conceptual Background

2.1. From HR analytics to agentic HCM

Early AI in HRM centred on prediction. Systems screened résumés, forecast turnover and flagged performance risks [1], [2]. People analytics then broadened the data used about employees, from transactions to digital traces of daily work [18], [19]. Algorithmic management describes the next step, in

which software takes on managerial functions. Parent-Rochelleau and Parker [20] identify six such functions: monitoring, goal setting, performance management, scheduling, compensation and job termination. Kellogg et al. [10] describe how algorithms restrict, recommend, record, rate, replace and reward workers.

Generative AI changed the interface. HR staff could now ask for drafts, summaries and analyses in plain language [21], [22]. Agentic AI changes the locus of action. An LLM-based agent combines a model with memory, planning and tool use [5], [6]. In an HCM suite, the tools are the platform's own functions. They include reading records, creating transactions and sending messages. The agent can therefore move from producing text to changing the system of record.

2.2. Levels of automation and human control

Human factors research offers a useful vocabulary for this shift. Parasuraman et al. [23] separate four classes of function that can be automated: information acquisition, information analysis, decision and action selection, and action implementation. Each can be automated at levels from fully manual to fully automatic. Simmler and Frischknecht [24] add a second dimension. They distinguish automation, which follows fixed rules, from technical autonomy, which adapts its own behaviour. Both matter for accountability. Shneiderman [25] rejects the idea of a single slider between human and machine. He argues that systems can offer high automation and high human control at the same time. Raisch and Krakowski [26] make a related organisational point. Automation and augmentation are not alternatives, because over time each shapes the other.

2.3. Governance of AI in organisations

Birkstedt et al. [13] define AI governance as a system of rules, practices and processes that keeps an organisation's AI use aligned with its strategy, values and legal duties. Papagiannidis et al. [14] describe responsible AI governance through structural, relational and procedural practices. Janssen [27] treats generative AI as part of a complex adaptive system in which people, data, policies and models co-evolve. The NIST framework organises risk work into four functions: govern, map, measure and manage [15]. The EU AI Act lists AI used in employment, worker management and access to self-employment among high-risk systems in Annex III. It adds duties for human oversight and, for deployers, for informing affected workers [16]. A 2026 amending regulation, known as the Digital Omnibus on AI, moved the date from which these Annex III obligations apply to 2 December 2027 [28]. These sources give the scaffolding. The review below asks what the evidence says about filling it in for HCM agents.

3. Methodology

3.1. Design

The study used a systematic review with thematic synthesis. Reporting follows the PRISMA 2020 statement where its items apply to a qualitative synthesis [29]. A review design suits the research questions because the evidence on HCM agents is new, spread across disciplines and methodologically mixed. The protocol was defined before searching but was not registered.

3.2. Information source and search strategy

Searches ran on 20 October 2026 in the Consensus scholarly index. This index draws on Semantic Scholar, OpenAlex and PubMed and covers more than 400 million records. It ranks results by semantic relevance to a plain-language query. The filter that excludes preprints was applied to every search. Eighteen searches were run, grouped into four concept blocks (Table 1). For each search, the ten most relevant records were

exported. This gave 180 records in total. Fixing the number of records per search kept the screening load equal across concept blocks and made the process reproducible. The trade-off is discussed in Section 5.5.

In addition, four sources were added from outside the searches. Two are foundational works that the included reviews cite heavily [10], [26]. Two are governance instruments that any HCM deployment must consider [15], [16].

Table 1. Search strings by concept block.

Concept block	Search strings (S01–S18)
A. AI in HRM and people analytics	S01 artificial intelligence in human resource management systematic review; S09 generative AI ChatGPT human resource management opportunities risks; S13 artificial intelligence in human resources management challenges and a path forward; S17 people analytics employee data privacy ethics surveillance; S18 algorithm-based HR decision-making personal integrity ethics
B. Algorithmic and agentic decision making	S02 algorithmic management of workers control and autonomy; S03 large language model based autonomous agents survey; S07 agentic AI in organizations delegation of decision authority to AI agents; S14 model of types and levels of human interaction with automation
C. Fairness, trust and oversight	S04 algorithmic bias in AI-based hiring and recruitment; S06 employee trust and acceptance of AI decisions in human resources fairness perceptions; S08 automation bias human oversight of algorithmic decision support; S11 accountability gap responsibility for autonomous AI systems decisions
D. Governance, reliability and security	S05 responsible AI governance framework organizations; S10 explainable artificial intelligence concepts taxonomies responsible AI; S12 hallucination in large language models survey; S15 human-centered artificial intelligence reliable safe trustworthy governance structures; S16 indirect prompt injection attacks LLM-integrated applications

3.3. Eligibility criteria

Table 2 lists the criteria. In short, a record had to be a peer-reviewed journal article or proceedings paper in English. It also had to contribute evidence that transfers to HCM on at least one of three topics: AI in people management, the capabilities and risks of LLM-based or agentic systems, or organisational mechanisms for overseeing AI decisions.

Table 2. Eligibility criteria.

Criterion	Inclusion	Exclusion code
Publication type	Peer-reviewed journal article or conference proceedings paper	E1: unknown venue, no bibliographic metadata, or not peer-reviewed
Relevance	Addresses AI in people management, LLM-based or agentic AI capability or risk, or organisational oversight of AI decisions, with content that transfers to HCM	E2: domain-specific study (for example, medical imaging or autonomous driving) with no transferable HCM or governance content
Contribution type	Original analysis: empirical, review, technical or conceptual	E4: editorial, preface, comment or tutorial
Added value	Adds evidence not already covered by an included source	E5: redundant with a broader, more established or more HCM-specific included source
Language	English	Not applied (all records were in English)

3.4. Selection process

One reviewer screened titles, abstracts and bibliographic metadata for all records, supported by an AI assistant that helped organise records and draft decision notes. Every decision was checked by the author and logged with an exclusion code. Exact duplicates across searches were removed first. Records were then screened against criteria E1, E2 and E4. Records that passed were assessed for added value (E5). When several sources reported the same evidence, the review kept the source that was broader, more established or more specific to HCM. The full screening log, with one row per record, is available as supplementary material.

3.5. Data extraction and synthesis

For each included source, the following items were extracted: design, setting, HCM process covered, AI type, main finding, and any governance mechanism proposed or tested. Each source was then coded to one primary theme. Eight themes were defined. Five came from the research questions: agentic capability and delegation (T1), algorithmic management and HRM (T2), fairness and bias (T3), human oversight and automation bias (T4), and governance frameworks (T8). Three emerged during coding: accountability and responsibility (T5), employee justice and trust (T6), and reliability and security of LLM systems (T7). The coded themes were then compared across sources and condensed into seven findings. Each finding is supported by sources from more than one theme.

3.6. Quality appraisal

The evidence base mixes experiments, reviews, conceptual papers and technical studies. A single risk-of-bias score would not be meaningful across these designs. Quality was therefore appraised at the level of venue type, design transparency and, for experiments, sample size and manipulation clarity. These judgements informed the weight given to each source in the synthesis, but they were not used to exclude records.

4. Results

Study selection and characteristics

Figure 1 shows the selection flow. The 18 searches returned 180 records. Three were duplicates, leaving 177 for screening. At the first stage, 38 records were excluded: 5 under E1, 27 under E2 and 6 under E4. Of the 139 records assessed for added value, 84 were excluded as redundant under E5. This left 55 records from the searches. With the 4 sources added from other routes, the synthesis drew on 59 sources.

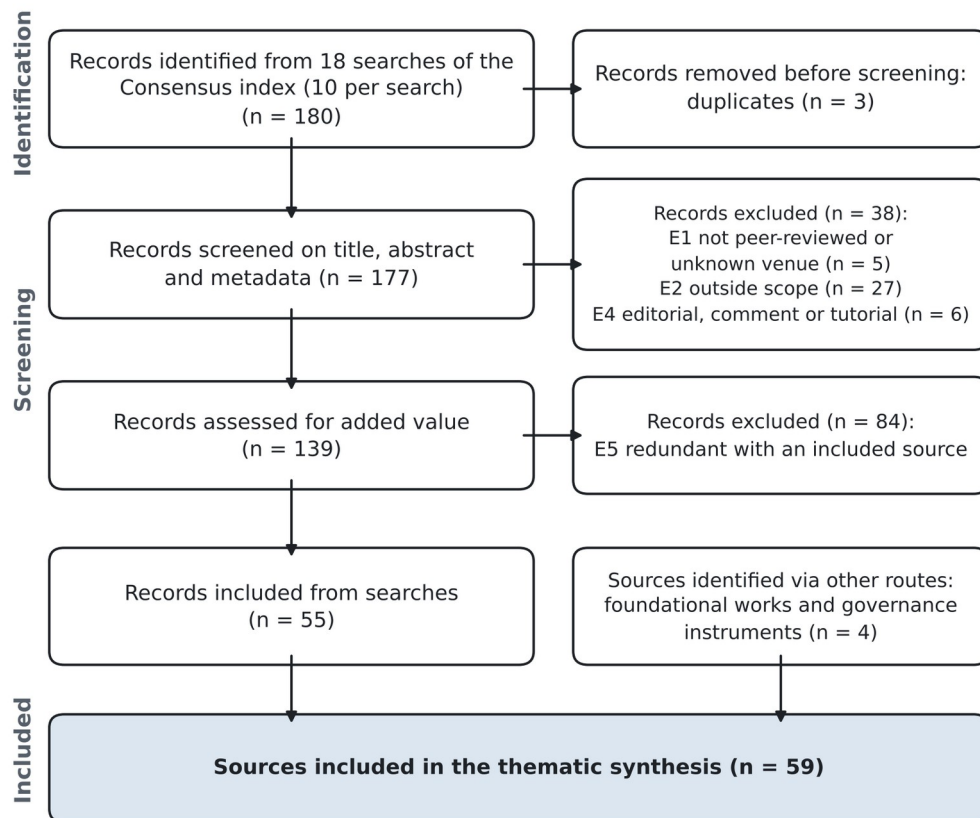


Fig. 1. Flow of records through the review, adapted from the PRISMA 2020 flow diagram.

The included evidence is recent. Thirty-eight of the 59 sources (64%) first appeared in 2023 or later. Reviews and surveys were the largest design group (22 sources), followed by conceptual or analytical papers (19) and experiments (10). Qualitative studies (3), technical studies (3) and governance instruments (2) made up the rest. Figure 2 shows the distribution of sources across the eight themes. Algorithmic management and HRM (11 sources) and fairness and bias (9) were the largest themes. Employee justice and trust (4) and reliability and security (5) were the smallest.

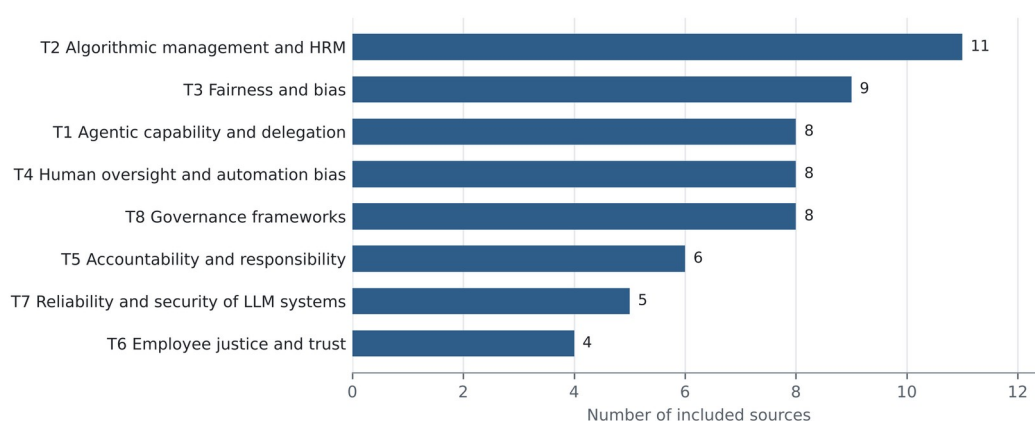


Fig. 2. Number of included sources by primary theme (n = 59).

The seven findings below report what the sources say. Interpretation is left to Section 5. Table 3 summarises the findings.

Finding 1: Algorithmic control of work is extensive, and agents extend it from deciding to doing

The literature shows that software already performs core managerial work. Parent-Rochelleau and Parker [20] map algorithms onto six management functions, from monitoring to termination. Noponen et al. [30] reviewed 172 articles and found that organisations have mostly used algorithmic systems to control workers rather than to enable them. Meijerink and Bondarouk [11] argue that the same systems can both restrict and enable autonomy, depending on design and on how workers respond. Zhang et al. [31] identify five perspectives on algorithmic management. One of them treats the algorithm as an "agentic boss".

Agentic AI deepens this pattern. Surveys of LLM-based agents describe systems that combine a model with memory, planning and tool use [5], [6]. Islam et al. [7] find four research streams in agentic AI. One of them is the balance between AI autonomy and human control. Candrian and Scherer [32] found that people were more willing to delegate decisions to AI than to another person, especially when losses were possible. Raisch and Krakowski [26] warn that managers who treat automation and augmentation as separate choices will mismanage both. Together, these sources indicate that the move from recommending to executing is already under way in management practice.

Finding 2: Bias in HR algorithms is persistent and partly organisational

Fairness is the most studied risk of AI in HR. A multidisciplinary survey of algorithmic hiring found the field split between two narratives [12]. One sees algorithms replacing biased recruiters. The other sees them automating discrimination. Chen [33] traces bias to limited training data and to the choices of system designers. The study recommends both technical fixes and internal and external governance. Raghavan et al. [34] examined what hiring-assessment vendors disclosed about validation. They found that de-biasing methods interact in complex ways with anti-discrimination law.

Bias is not only a data problem. Soleimani et al. [35] interviewed 39 HR professionals and AI developers. They found that HR staff need both technical and people skills to challenge developers effectively. Leicht-Deobald et al. [36] argue that the efficiency logic of algorithmic HR decisions pushes employees from personal integrity towards mere compliance. Work on people analytics adds privacy and surveillance concerns. These include profiling, coercion and the use of employee data for purposes employees never agreed to [18], [19], [37]. Andrieux et al. [38] show that generative AI widens these concerns, because each new capability creates new ethical questions for each HR function.

Finding 3: Human oversight is weaker than assumed

Many governance proposals rely on a human in the loop. The experimental evidence suggests that this safeguard is fragile. Agudo et al. [39] found that incorrect AI advice lowered decision accuracy, especially when people saw the advice before forming their own view. Sele and Chugunova [40] found that letting people adjust algorithmic recommendations increased uptake. However, participants were least likely to intervene when the recommendations were least accurate. In their study, the human-in-the-loop design increased uptake but decreased accuracy.

Two experiments come directly from personnel selection. Kupfer et al. [41] found that warning decision makers about possible system errors increased how carefully they checked candidates. Showing less aggregated data had a similar effect. Telling them they were responsible for the decision did not. Rosenthal-von der Pütten and Sach [42] presented biased hiring recommendations to 260 participants. About 60% of those in the bias condition did not notice the bias, even when asked about it directly. Alon-Barkat and Busuioc [43] add a nuance. In their public-sector experiments, people did not over-rely on algorithms in general. They did, however, adopt advice more readily when it matched group stereotypes.

A systematic review of 35 studies confirms that automation bias depends on AI literacy, expertise, task difficulty and the design of explanations [44]. That review found that explanations often raise acceptance without improving accuracy. Active verification effort was the most promising point of intervention. Trust research points the same way. Hoff and Bashir [45] show that trust in automation has dispositional, situational and learned layers, and that well-calibrated trust is a design goal rather than a default. Gonzalez et al. [46] argue that human-AI teams outperform either party alone only under specific conditions, including clear role partitioning and deliberate interrogation of AI outputs.

Finding 4: Responsibility diffuses as autonomy grows

As systems act with less human input, it becomes harder to say who is responsible for outcomes. Santoni de Sio and Mecacci [47] separate four responsibility gaps: culpability, moral accountability, public accountability and active responsibility. They argue that designing for meaningful human control is the most complete response. Vallor and Vierkant [48] suggest that the real gap lies elsewhere. In their view, it is an asymmetry of vulnerability between the people affected and the sociotechnical systems that affect them. Simmler and Frischknecht [24] show that higher automation and higher technical autonomy each complicate accountability in different ways.

Recent work names specific organisational mechanisms. Issa et al. [49] describe "agentic loafing", in which professionals hand decisions to AI without examining that choice. They identify three drivers: pressure to conform to performance cultures, the structural isolation of responsibility, and legitimacy gained through quantification. The result is what they call risk evaporation, where no one is accountable when an AI-assisted decision fails. Leszczyński et al. [50] report a contrasting effect in hiring. Managers who shared tasks with an algorithm behaved more ethically, partly because they could not shift blame to other people. Capasso et al. [51] link these concerns to workers' rights to equality, privacy and work. They argue for explicit ethical safeguards in algorithmic HRM.

Finding 5: Employee trust depends on fair process and visible human involvement

Employees judge AI decisions through the lens of organisational justice. In a study spanning six HR functions, people generally felt treated with more respect when a human made the decision [52]. Positive outcomes improved perceptions regardless of who decided. Köchling et al. [53] found that less human involvement in career development decisions lowered perceived fairness and raised perceived privacy intrusion. These perceptions in turn predicted turnover intention and employer attractiveness.

The picture is not one-sided. In a field experiment on structured, data-heavy performance evaluation, employees rated AI as fairer and more accurate than the average human manager [54]. This suggests that trust depends on the task, the data and how the process is explained. Budhwar et al. [21] note that generative AI heightens existing concerns about well-being, bias, privacy and context-insensitive outputs. These are the concerns that most affect employee trust. Together, these studies indicate that the presence of a human matters, but so does the quality of the process that the human and the system follow.

Finding 6: Language-model agents add new reliability and security failure modes

LLMs produce fluent text that can be wrong. Huang et al. [55] classify these hallucinations and review methods to detect and reduce them. They note that retrieval-based grounding reduces the problem but does not remove it. Farquhar et al. [56] show that one class of hallucination, arbitrary confabulation, can be detected by measuring uncertainty at the level of meaning. Practitioner guidance for HR therefore stresses structured prompts and a separate verification step before any output is used [22].

Agents add security risks that ordinary HR software does not face. Greshake et al. [57] showed that LLM-integrated applications blur the line between data and instructions. An attacker can plant instructions in content the system later retrieves. Milani et al. [58] demonstrated this risk in a résumé comparison scenario. Hidden prompts in a curriculum vitae steered the model's output and misled the HR user. Su et al. [59] survey risks that grow with agent autonomy. These include memory poisoning, tool misuse and irreversible chains of tool calls. They find that most current defences work in isolation and lack integration across the agent's perception, memory and action modules.

Finding 7: Governance frameworks rarely reach the level of a delegated action

Organisational AI governance research is still young. Birkstedt et al. [13] found few explicit definitions and limited understanding of how governance is implemented. They recommend a dedicated AI oversight unit. Papagiannidis et al. [14] found that principles are common but operational practice is not. Janssen [27] argues that no single party takes responsibility when the whole sociotechnical system is poorly understood, so joint accountability is needed. Shneiderman [25] proposes governance at three levels: reliable engineering in teams, a safety culture in organisations and independent oversight across industry. Barredo Arrieta et al. [60] tie responsible AI to fairness, explainability and accountability, and stress that explanations must suit their audience.

Formal instruments share this level of abstraction. The NIST framework is designed to apply across sectors and use cases [15]. The EU AI Act sets duties for high-risk employment AI but leaves the design of oversight to providers and deployers [16]. Gluszek and Gluszek [17] analysed 795 publications bibliometrically and found three research communities developing in parallel: AI ethics governance, machine learning operations and agentic enterprise integration. In their analysis, the combination of high autonomy and high accountability is the least theorised. Srivastava [61] offers formal support for a middle path. In an analytical model of pricing and inventory decisions, partial delegation to an AI agent could outperform both full autonomy and full human control.

Table 3. Summary of findings.

Finding	Main sources	Implication for HCM agents
F1. Algorithmic control is extensive; agents extend it to action	[5]–[7], [11], [20], [30]–[32]	The question is no longer whether software manages work, but which actions it may take alone
F2. Bias is persistent and partly organisational	[12], [18], [33]–[37]	Technical de-biasing is necessary but not sufficient; HR capability and data limits matter
F3. Human oversight is fragile	[39]–[44]	A reviewer in the loop does not by itself ensure good decisions; review must be designed
F4. Responsibility diffuses with autonomy	[24], [47]–[51]	Every delegated action needs a named human owner and a record
F5. Trust depends on process and human presence	[21], [52]–[54]	Employees need to know when an agent is involved and how to challenge the outcome
F6. LLM agents add new failure modes	[55]–[59]	Agents need bounded tools, separation of data from instructions and output verification
F7. Governance rarely reaches the action level	[13]–[17], [27], [61]	Organisations must translate principles into rules for specific delegated actions

5. Discussion

5.1. Interpreting the evidence

Read together, the seven findings point to one central argument. In HCM, the autonomy of an AI agent and the consequence of the decision it touches must be governed together. Neither can be set safely on its own.

Three tensions in the evidence support this view. The first concerns human oversight. The policy default is to keep a human in the loop. Yet the experiments show that people often defer to machine advice and miss its errors [39], [40], [42]. Keeping a human in the loop is therefore necessary for consequential decisions but not sufficient. The form of the review matters as much as its presence [41], [44].

The second tension concerns trust. Employees generally prefer human involvement [52], [53]. Yet in structured, data-rich tasks they may rate AI as fairer than managers [54]. This suggests that the right level of autonomy depends on the task and on how much the decision affects the person, not on a fixed preference for humans or machines.

The third tension concerns accountability. As autonomy rises, responsibility tends to evaporate [47], [49]. Yet clear human ownership can make managers behave more ethically when they work with AI [50]. This implies that accountability is a design choice rather than a fixed property of the technology.

These tensions explain why HCM needs graduated autonomy. Full autonomy is unsafe for decisions that determine employment. Full human control wastes the value of agents for routine service work. The analytical result that partial delegation can dominate both extremes [61] fits this pattern. The framework below makes the middle ground explicit.

5.2. The Consequence-Calibrated Agency framework

The C2A framework has three parts: a scale of agent autonomy, a classification of decision consequence and a set of control layers. Figure 3 shows how they fit together.

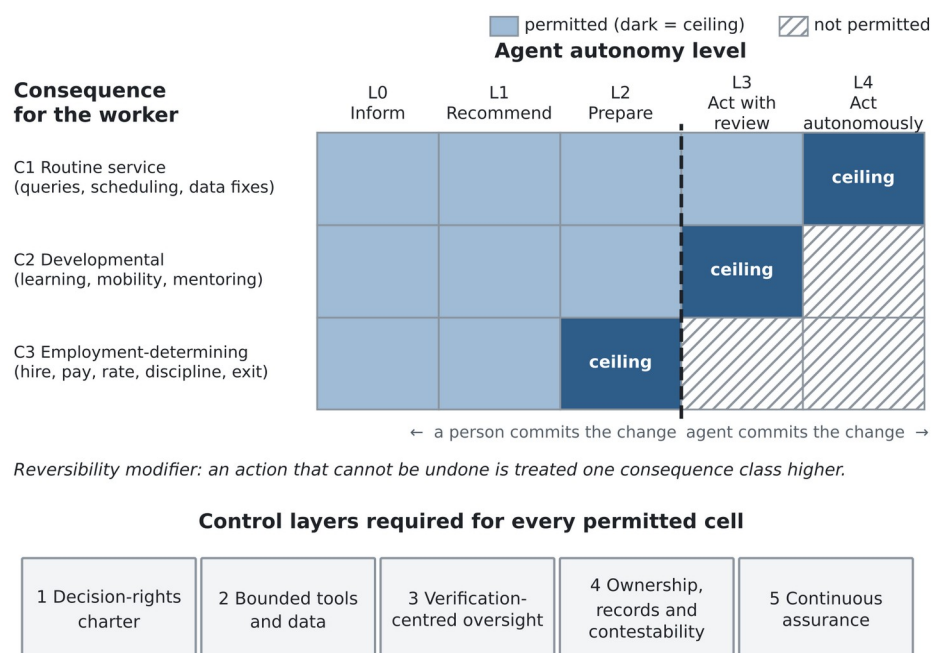


Fig. 3. The Consequence-Calibrated Agency (C2A) framework. Shaded cells show the permitted autonomy for each consequence class; the five control layers apply to every permitted cell.

5.2.1. *Autonomy levels*

Table 4 defines five levels of agent autonomy. The levels adapt the function classes of Parasuraman et al. [23] to HCM transactions. The key line falls between Level 2 and Level 3. Up to Level 2, a person performs the action that changes the system of record. From Level 3, the agent performs it.

Table 4. Autonomy levels for HCM agents.

Level	Name	What the agent does	Who commits the change	HCM example
L0	Inform	Retrieves and summarises information	No change is made	Answers a policy question with a link to the source policy
L1	Recommend	Analyses and ranks options	A person decides and acts	Suggests learning courses or shortlists internal candidates for review
L2	Prepare	Builds a complete transaction for approval	A person approves each transaction	Drafts an offer letter or a pay adjustment for a manager to approve
L3	Act with review	Executes, then notifies; change can be reversed within a set window	The agent, under human review	Books interview slots or updates an address change
L4	Act autonomously	Executes and logs without prior notice	The agent, under periodic audit	Corrects formatting errors in records or routes a ticket

5.2.2. *Consequence classes*

The framework classifies each HCM decision by its consequence for the worker, not by its cost to the organisation. Three classes are defined. Class C1 covers routine service, such as answering questions, scheduling and correcting data. Class C2 covers developmental decisions, such as learning recommendations, mentoring matches and internal job suggestions. These shape opportunities but do not determine them. Class C3 covers employment-determining decisions: hiring, pay, promotion, performance ratings, discipline and termination. Class C3 broadly matches the employment uses that the EU AI Act lists as high-risk [16].

A reversibility modifier adjusts the class. An action that cannot be fully undone, such as sending a rejection to a candidate, is treated one class higher than its content alone would suggest. This reflects the warning by Su et al. [59] about irreversible tool chains.

Each class receives an autonomy ceiling. Class C1 actions may run up to L4. Class C2 actions may run up to L3, or L2 when sensitive personal data are involved. Class C3 actions may not exceed L2. For Class C3, the agent may prepare but never commit. Table 5 applies these ceilings to common HCM processes.

Table 5. Example autonomy ceilings for HCM processes.

HCM process	Consequence class	Ceiling	Rationale from the review
Policy and benefits questions	C1	L4 (L0 for guidance with legal effect)	Low consequence, but hallucination risk requires cited sources (F6)
Interview scheduling	C1	L3	Reversible, high volume; review window catches errors (F1)
Learning recommendations	C2	L3	Shapes development but does not determine employment (F5)
Internal mobility matching	C2	L2	Uses sensitive performance data; bias risk (F2)

HCM process	Consequence class	Ceiling	Rationale from the review
Candidate screening and shortlisting	C3	L1	High-risk under the EU AI Act; bias and prompt-injection exposure (F2, F6)
Offer and compensation changes	C3	L2	Agent may prepare; a named manager must approve (F4)
Performance ratings	C3	L1	Trust depends on fair process and human presence (F5)
Discipline and termination	C3, irreversible	L0 to L1	Highest consequence; responsibility must stay with people (F4)

5.2.3. Control layers

Autonomy ceilings alone are not enough. Each permitted cell in Figure 3 must also be supported by five control layers (Table 6). Each layer answers one finding from the review.

Table 6. The five control layers of the C2A framework.

Layer	Purpose	Example controls	Findings addressed
1. Decision-rights charter	States which actions each agent may take, at which level, for which population	Written autonomy ceiling per process; approval by HR, legal and employee representatives; change control for any increase in autonomy	F1, F7
2. Bounded tools and data	Limits what the agent can see and do	Least-privilege tool scopes; separation of retrieved content from instructions; screening of uploaded documents for embedded instructions; no direct write access for C3 actions	F2, F6
3. Verification-centred oversight	Makes human review effective rather than ceremonial	Reviewer forms a view before seeing the agent's recommendation; disaggregated evidence rather than a single score; explicit display of error rates; sampling of approved items for second review	F3
4. Ownership, records and contestability	Keeps responsibility with people and gives employees a voice	A named human owner for every delegated action; an action log with inputs, tools called and outputs; notice to employees when an agent was involved; a simple route to challenge and appeal	F4, F5
5. Continuous assurance	Detects drift, bias and incidents over time	Adverse-impact testing by demographic group; monitoring of override and reversal rates; incident reporting; periodic re-certification of each ceiling	F2, F6, F7

Layer 3 draws directly on the automation bias evidence. Kupfer et al. [41] found that warnings about system errors and less aggregated data improved checking. Agudo et al. [39] found that advice shown before a person's own judgement did the most harm. The framework therefore asks reviewers to record their own view first for Class C3 decisions. Layer 4 draws on the responsibility literature. It gives every delegated action a named owner, which addresses risk evaporation [49] and supports meaningful human control [47]. Layer 2 responds to the security evidence. Résumés, emails and documents are exactly the kind of retrieved content through which indirect prompt injection works [57], [58].

5.2.4. Relation to existing frameworks

The C2A framework builds on, but differs from, three existing approaches. The levels of automation of Parasuraman et al. [23] describe what is automated, not when automation is acceptable. C2A adds the consequence axis that makes that judgement possible. The Delegated Agentic Governance Model of Gluszk

and Gluszak [17] also links governance to delegation tiers. C2A differs in three ways. It defines consequence from the worker's point of view. It adds a reversibility modifier. It specifies oversight design and employee contestability as explicit layers. Finally, C2A can be read as an HCM-specific way to carry out the govern, map, measure and manage functions of the NIST framework [15] and the oversight duties of the EU AI Act [16].

5.3. Propositions for future research

The framework yields six propositions (Table 7). Each can be tested with field data from HCM deployments, with experiments or with simulation.

Table 7. Testable propositions.

No.	Proposition	Suggested test
P1	Deploying an HCM agent above the autonomy ceiling for its consequence class increases the rate of adverse employment-decision incidents.	Incident and reversal data across deployments with different ceilings
P2	Verification-centred review designs (own judgement first, disaggregated evidence, error salience) reduce acceptance of incorrect agent recommendations compared with standard approval screens.	Between-subjects experiment with HR professionals
P3	A named human owner and a visible appeal route increase employees' perceived procedural justice and trust in agent-assisted decisions.	Vignette experiment or field survey with employees
P4	Separating retrieved content from instructions and limiting tool scopes reduce successful prompt-injection attempts in HCM workflows without a material loss of task utility.	Red-team testing on a benchmark of HR documents
P5	For developmental (C2) processes, partial delegation (L2 to L3) yields higher net value than either L0 or L4.	Field comparison or simulation of service quality, error cost and staff time
P6	Governance maturity, measured across the five control layers, mediates the relationship between agent adoption and sustained employee trust.	Multi-wave organisational survey

5.4. Practical implications

For HR leaders, the first step is an inventory. Each agent use case should be listed with its consequence class and its current level of autonomy. Any case where the level exceeds the ceiling needs action before scale-up. HR leaders also need to invest in the joint technical and people skills that Soleimani et al. [35] found lacking. Without them, HR staff cannot challenge vendors or developers effectively.

For technology and platform leaders, the framework translates into design rules. Agent permissions should follow the decision-rights charter, not the full permissions of the user who invokes the agent. Retrieved documents should be treated as untrusted data. Approval screens for Class C3 actions should be redesigned so that review is substantive. Action logs should be complete enough to reconstruct any decision on appeal.

For employee representatives and regulators, the framework offers a common language. A ceiling per decision class is easier to negotiate and audit than a general principle. Under the EU AI Act, deployers of high-risk employment AI must inform workers' representatives and affected workers before such a system is used at the workplace [16]. These obligations now apply from December 2027 [28], which gives organisations a short window to prepare. A published decision-rights charter would help meet that duty in a way that employees can understand.

5.5. Limitations

This study has several limitations. First, it used a single scholarly index and exported the ten most relevant records per search. Relevance ranking favours well-cited and closely matching work. Some relevant studies, especially recent ones with few citations, may have been missed. Second, screening was done by one reviewer, supported by an AI assistant, and was based on titles, abstracts and metadata. A second independent reviewer and full-text screening would strengthen reliability. Third, the redundancy criterion (E5) involves judgement. The screening log is published so that readers can examine each decision. Fourth, much of the agentic AI evidence comes from outside HCM, and many 2025–2026 sources are new and not yet widely tested. Fifth, the C2A framework is conceptual. Its ceilings are reasoned from the evidence but have not yet been validated in practice. Finally, regulation is changing quickly. The application date of the EU AI Act's high-risk obligations for employment systems has already been moved once, to December 2027 [28]. Practitioners should check the current position before relying on any date.

6. Conclusion

AI agents are moving into the systems that hire, pay, rate and dismiss people. This review shows why that move needs a different kind of governance. Algorithmic control of work is already extensive. Bias persists. Human review is weaker than often assumed. Responsibility tends to diffuse, and trust depends on fair process. LLM agents add new failure modes. Meanwhile, most governance frameworks stop short of saying which specific actions an agent may take alone.

The Consequence-Calibrated Agency framework addresses this gap. It links the autonomy of each HCM agent action to the consequence of that action for the worker. It then requires five control layers: a decision-rights charter, bounded tools and data, verification-centred oversight, ownership with contestability, and continuous assurance. The central rule is simple. For decisions that determine employment, an agent may prepare but must not commit.

Future research should test the six propositions with field data, experiments and red-team studies. Studies in different legal and cultural settings would also show how far the consequence classes travel. As agents become more capable, the case for graduated, consequence-calibrated autonomy is likely to grow stronger.

Funding

This research received no external funding.

Conflict of Interest

The author is the founder of Innovate Software Consulting Inc Ltd, which provides consulting services on enterprise HCM and analytics systems. The author declares no other conflict of interest.

Data Availability Statement

This review analysed published literature only. No new primary data were collected. The complete screening log, listing every retrieved record with its search string, decision code and assigned theme, is provided as a supplementary file (IJSTM_C2A_screening_log.xlsx). The author will deposit it in a public repository, such as Zenodo, on acceptance.

References

- [1] D. Vrontis, M. Christofi, V. Pereira, S. Tarba, A. Makrides, and E. Trichina, "Artificial intelligence, robotics, advanced technologies and human resource management: A systematic review," *Int. J. Hum. Resour. Manag.*, vol. 33, no. 6, pp. 1237–1266, 2022, doi: 10.1080/09585192.2020.1871398.
- [2] A. M. Votto, R. Valecha, P. Najafirad, and H. R. Rao, "Artificial intelligence in tactical human resource management: A systematic literature review," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, Art. no. 100047, 2021, doi: 10.1016/j.jjimei.2021.100047.
- [3] P. Budhwar, A. Malik, M. T. T. De Silva, and P. Thevisuthan, "Artificial intelligence – challenges and opportunities for international HRM: A review and research agenda," *Int. J. Hum. Resour. Manag.*, vol. 33, no. 6, pp. 1065–1097, 2022, doi: 10.1080/09585192.2022.2035161.
- [4] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *Calif. Manage. Rev.*, vol. 61, no. 4, pp. 15–42, 2019, doi: 10.1177/0008125619867910.
- [5] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, et al., "A survey on large language model based autonomous agents," *Front. Comput. Sci.*, vol. 18, no. 6, Art. no. 186345, 2024, doi: 10.1007/s11704-024-40231-1.
- [6] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, et al., "The rise and potential of large language model based agents: A survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, Art. no. 121101, 2025, doi: 10.1007/s11432-024-4222-0.
- [7] M. A. Islam, S. Somu, and F. M. F. Aldaihani, "The rise of agentic AI: Synthesis of current knowledge and future research agenda," *Glob. Bus. Organ. Excell.*, vol. 45, no. 3, pp. 402–416, 2026, doi: 10.1002/joe.70019.
- [8] L. Hughes, Y. K. Dwivedi, T. Malik, M. Shawosh, M. A. Albashrawi, I. Jeon, et al., "AI agents and agentic systems: A multi-expert analysis," *J. Comput. Inf. Syst.*, vol. 65, no. 4, pp. 489–517, 2025, doi: 10.1080/08874417.2025.2483832.
- [9] P. van Esch, "From agentic AI to AI-orchestrated organizations: Understanding the next surge in artificial intelligence," *Bus. Horiz.*, early access, 2026, doi: 10.1016/j.bushor.2026.03.003.
- [10] K. C. Kellogg, M. A. Valentine, and A. Christin, "Algorithms at work: The new contested terrain of control," *Acad. Manag. Ann.*, vol. 14, no. 1, pp. 366–410, 2020, doi: 10.5465/annals.2018.0174.
- [11] J. Meijerink and T. Bondarouk, "The duality of algorithmic management: Toward a research agenda on HRM algorithms, autonomy and value creation," *Hum. Resour. Manag. Rev.*, vol. 33, no. 1, Art. no. 100876, 2023, doi: 10.1016/j.hrmr.2021.100876.
- [12] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 1, Art. no. 16, 2025, doi: 10.1145/3696457.
- [13] T. Birkstedt, M. Minkinen, A. Tandon, and M. Mäntymäki, "AI governance: Themes, knowledge gaps and future agendas," *Internet Res.*, vol. 33, no. 7, pp. 133–167, 2023, doi: 10.1108/INTR-01-2022-0042.
- [14] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *J. Strateg. Inf. Syst.*, vol. 34, no. 2, Art. no. 101885, 2025, doi: 10.1016/j.jsis.2024.101885.
- [15] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," U.S. Dept. of Commerce, Gaithersburg, MD, USA, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.
- [16] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, Jul. 12, 2024. [Online]. Available: <http://data.europa.eu/eli/reg/2024/1689/oj>
- [17] L. Gluszak and F. Gluszak, "Delegated agentic governance: A delegation-centred framework for managing autonomous AI in organisations," *J. Inf. Knowl. Manag.*, early access, Jul. 2026, doi: 10.1142/S0219649226500486.
- [18] L. M. Giermindl, F. Strich, O. Christ, U. Leicht-Deobald, and A. Redzepi, "The dark sides of people analytics: Reviewing the perils for organisations and employees," *Eur. J. Inf. Syst.*, vol. 31, no. 3, pp. 410–435, 2022, doi: 10.1080/0960085X.2021.1927213.
- [19] A. Tursunbayeva, C. Pagliari, S. Di Lauro, and G. Antonelli, "The ethics of people analytics: Risks, opportunities and recommendations," *Pers. Rev.*, vol. 51, no. 3, pp. 900–921, 2022, doi: 10.1108/PR-12-2019-0680.

- [20] X. Parent-Rochelleau and S. K. Parker, "Algorithms as work designers: How algorithmic management influences the design of jobs," *Hum. Resour. Manag. Rev.*, vol. 32, no. 3, Art. no. 100838, 2022, doi: 10.1016/j.hrmr.2021.100838.
- [21] P. Budhwar, S. Chowdhury, G. Wood, H. Aguinis, G. J. Bamber, J. R. Beltran, et al., "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT," *Hum. Resour. Manag. J.*, vol. 33, no. 3, pp. 606–659, 2023, doi: 10.1111/1748-8583.12524.
- [22] H. Aguinis, J. R. Beltran, and A. Cope, "How to use generative AI as a human resource management assistant," *Organ. Dyn.*, vol. 53, Art. no. 101029, 2024, doi: 10.1016/j.orgdyn.2024.101029.
- [23] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 3, pp. 286–297, 2000, doi: 10.1109/3468.844354.
- [24] M. Simmler and R. Frischknecht, "A taxonomy of human-machine collaboration: Capturing automation and technical autonomy," *AI Soc.*, vol. 36, no. 1, pp. 239–250, 2021, doi: 10.1007/s00146-020-01004-z.
- [25] B. Shneiderman, "Human-centered artificial intelligence: Three fresh ideas," *AIS Trans. Hum.-Comput. Interact.*, vol. 12, no. 3, pp. 109–124, 2020, doi: 10.17705/1thci.00131.
- [26] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," *Acad. Manage. Rev.*, vol. 46, no. 1, pp. 192–210, 2021, doi: 10.5465/amr.2018.0072.
- [27] M. Janssen, "Responsible governance of generative AI: Conceptualizing GenAI as complex adaptive systems," *Policy Soc.*, vol. 44, no. 1, pp. 38–51, 2025, doi: 10.1093/polsoc/puae040.
- [28] European Parliament and Council of the European Union, "Regulation (EU) 2026/1744 of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)," *Off. J. Eur. Union*, L series, 2026. [Online]. Available: <http://data.europa.eu/eli/reg/2026/1744/oj>
- [29] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.
- [30] N. Noponen, P. Feshchenko, T. Auvinen, V. Luoma-aho, and P. Abrahamsson, "Taylorism on steroids or enabling autonomy? A systematic review of algorithmic management," *Manag. Rev. Q.*, vol. 74, no. 3, pp. 1695–1721, 2024, doi: 10.1007/s11301-023-00345-5.
- [31] M. M. Zhang, F. L. Cooke, D. Ahlstrom, and N. McNeil, "The rise of algorithmic management and implications for work and organisations," *New Technol. Work Employ.*, vol. 40, no. 3, pp. 659–671, 2025, doi: 10.1111/ntwe.12343.
- [32] C. Candrian and A. Scherer, "Rise of the machines: Delegating decisions to autonomous AI," *Comput. Hum. Behav.*, vol. 134, Art. no. 107308, 2022, doi: 10.1016/j.chb.2022.107308.
- [33] Z. Chen, "Ethics and discrimination in artificial intelligence-enabled recruitment practices," *Humanit. Soc. Sci. Commun.*, vol. 10, Art. no. 567, 2023, doi: 10.1057/s41599-023-02079-x.
- [34] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proc. 2020 Conf. Fairness, Accountability, and Transparency (FAT '20)**, Barcelona, Spain, 2020, pp. 469–481, doi: 10.1145/3351095.3372828.
- [35] M. Soleimani, A. Intezari, J. Arrowsmith, D. J. Pauleen, and N. Taskin, "Reducing AI bias in recruitment and selection: An integrative grounded approach," *Int. J. Hum. Resour. Manag.*, early access, Mar. 2025, doi: 10.1080/09585192.2025.2480617.
- [36] U. Leicht-Deobald, T. Busch, C. Schank, A. Weibel, S. Schafheitle, I. Wildhaber, and G. Kasper, "The challenges of algorithm-based HR decision-making for personal integrity," *J. Bus. Ethics*, vol. 160, no. 2, pp. 377–392, 2019, doi: 10.1007/s10551-019-04204-w.
- [37] L. Manroop, A. Malik, and M. Milner, "The ethical implications of big data in human resource management," *Hum. Resour. Manag. Rev.*, vol. 34, no. 2, Art. no. 101012, 2024, doi: 10.1016/j.hrmr.2024.101012.
- [38] P. Andrieux, R. D. Johnson, J. Sarabadani, and C. Van Slyke, "Ethical considerations of generative AI-enabled human resource management," *Organ. Dyn.*, vol. 53, no. 1, Art. no. 101032, 2024, doi: 10.1016/j.orgdyn.2024.101032.

- [39] U. Agudo, K. G. Liberal, M. Arrese, and H. Matute, "The impact of AI errors in a human-in-the-loop process," *Cogn. Res. Princ. Implic.*, vol. 9, Art. no. 1, 2024, doi: 10.1186/s41235-023-00529-3.
- [40] D. Sele and M. Chugunova, "Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making," *PLOS ONE*, vol. 19, no. 2, Art. no. e0298037, 2024, doi: 10.1371/journal.pone.0298037.
- [41] C. Kupfer, R. Prassl, J. Fleiß, C. Malin, S. Thalmann, and B. Kubicek, "Check the box! How to deal with automation bias in AI-based personnel selection," *Front. Psychol.*, vol. 14, Art. no. 1118723, 2023, doi: 10.3389/fpsyg.2023.1118723.
- [42] A. M. Rosenthal-von der Pütten and A. Sach, "Michael is better than Mehmet: Exploring the perils of algorithmic biases and selective adherence to advice from automated decision support systems in hiring," *Front. Psychol.*, vol. 15, Art. no. 1416504, 2024, doi: 10.3389/fpsyg.2024.1416504.
- [43] S. Alon-Barkat and M. Busuioc, "Human-AI interactions in public sector decision making: 'Automation bias' and 'selective adherence' to algorithmic advice," *J. Public Adm. Res. Theory*, vol. 33, no. 1, pp. 153–169, 2023, doi: 10.1093/jopart/muac007.
- [44] G. Romeo and D. Conti, "Exploring automation bias in human-AI collaboration: A review and implications for explainable AI," *AI Soc.*, vol. 41, no. 1, pp. 259–278, 2026, doi: 10.1007/s00146-025-02422-7.
- [45] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.
- [46] C. Gonzalez, K. Donahue, D. G. Goldstein, H. Heidari, M. S. Jalali, B. Schelble, et al., "Toward a science of human-AI teaming for decision making: A complementarity framework," *PNAS Nexus*, vol. 5, no. 3, Art. no. pgag030, 2026, doi: 10.1093/pnasnexus/pgag030.
- [47] F. Santoni de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philos. Technol.*, vol. 34, no. 4, pp. 1057–1084, 2021, doi: 10.1007/s13347-021-00450-x.
- [48] S. Vallor and T. Vierkant, "Find the gap: AI, responsible agency and vulnerability," *Minds Mach.*, vol. 34, Art. no. 20, 2024, doi: 10.1007/s11023-024-09674-0.
- [49] H. Issa, F. J. Petani, and D. Glavas, "Agentic loafing: An AI decision delegation risk," *Risk Anal.*, vol. 46, no. 8, 2026, doi: 10.1111/risa.70306.
- [50] G. Leszczyński, P. Gaczek, and M. Ławrynowicz, "The bright side of AI in hiring: Collaborating with algorithms supports ethical decision-making," *J. Bus. Res.*, vol. 203, Art. no. 115812, 2026, doi: 10.1016/j.jbusres.2025.115812.
- [51] M. Capasso, P. Arora, D. Sharma, and C. Tacconi, "On the right to work in the age of artificial intelligence: Ethical safeguards in algorithmic human resource management," *Bus. Hum. Rights J.*, vol. 9, no. 3, pp. 346–360, 2024, doi: 10.1017/bhj.2024.26.
- [52] S. Banks, P. Formosa, Y. Griep, and D. Richards, "AI decision making with dignity? Contrasting workers' justice perceptions of human and AI decision making in a human resource management context," *Inf. Syst. Front.*, vol. 24, no. 3, pp. 857–875, 2022, doi: 10.1007/s10796-021-10223-8.
- [53] A. Köchling, M. C. Wehner, and S. A. Ruhle, "This (AI)n't fair? Employee reactions to artificial intelligence (AI) in career development systems," *Rev. Manag. Sci.*, vol. 19, no. 4, pp. 1195–1228, 2025, doi: 10.1007/s11846-024-00789-3.
- [54] S. Qin, N. Jia, X. Luo, C. Liao, and Z. Huang, "Perceived fairness of human managers compared with artificial intelligence in employee performance evaluation," *J. Manag. Inf. Syst.*, vol. 40, no. 4, pp. 1039–1070, 2023, doi: 10.1080/07421222.2023.2267316.
- [55] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, 2025, doi: 10.1145/3703155.
- [56] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.
- [57] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in *Proc. 16th ACM*

- Workshop Artif. Intell. Secur. (AISec '23)*, Copenhagen, Denmark, 2023, pp. 79–90, doi: 10.1145/3605764.3623985.
- [58] A. Milani, V. Franzoni, and E. Florindi, "Indirect prompt injection in large language models," *Neural Comput. Appl.*, vol. 38, no. 13, Art. no. 530, 2026, doi: 10.1007/s00521-026-12266-x.
- [59] H. Su, J. Luo, C. Liu, X. Yang, Y. Zhang, Y. Dong, et al., "A survey on autonomy-induced security risks in large model-based agents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 10, pp. 11382–11401, 2026, doi: 10.1109/TPAMI.2026.3688650.
- [60] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
- [61] A. Srivastava, "Governing AI-enabled decision making: Delegation, autonomy, and control at the operations–marketing interface," *Prod. Oper. Manag.*, early access, Jul. 2026, doi: 10.1177/10591478261473004.