Ethical AI Governance in HRM & Organisational Strategy: Designing trustworthy, inclusive, and sustainable digital workplaces

Evangelia Fragouli
Kingston University, London, UK

Keywords
HRM practices, AI, Digitisation, Organisational Strategy

Abstract

This paper explores how ethical AI governance within human resource management can strengthen organisational strategy in an era of rapid digital transformation. As organisations increasingly integrate AI into recruitment, performance management, and decision-making systems, concerns arise regarding fairness, transparency, and employee trust. The study employs a qualitative empirical study to examine how responsible AI practices can be designed and implemented within HRM. Key findings indicate that ethical AI governance is perceived as essential for enhancing organisational credibility, supporting fair talent processes, and reducing risks associated with bias and excessive monitoring. The analysis also shows that HRM plays a strategic role in aligning AI use with organisational values, shaping capability development, and fostering a culture that balances technological efficiency with human-centred principles. The paper offers practical guidance for HR practitioners seeking to implement responsible AI policies that contribute to sustainable organisational performance.

Introduction

Artificial intelligence (AI) has rapidly evolved into a transformative force reshaping organisational systems, workflows, and decision-making structures. Within human resource management (HRM), AI now underpins a wide range of functions, including recruitment, performance appraisal, workforce planning, and employee engagement. Recent scholarship highlights that AI adoption is no longer a purely technological decision but a strategic one, with implications for organisational legitimacy, transparency, and long-term sustainability. Studies show that AI integration can considerably enhance efficiency, standardisation, and predictive capability in HR processes, particularly in areas such as candidate screening, performance data analysis, and workforce planning (Venugopala et al., 2024). However, these potential benefits coexist with significant ethical concerns—most notably algorithmic bias, opacity in automated decision-making, and risks associated with pervasive monitoring—which threaten employee trust, well-being, and perceptions of fairness.

The literature underscores that ethical AI governance is increasingly viewed as essential for responsible adoption. Ethical frameworks that incorporate transparency, human oversight, fairness audits, and explainability mechanisms are commonly proposed as prerequisites for trustworthy AI systems in HRM (Aderla & Purnachander, 2024). Systematic reviews similarly suggest that although AI can promote greater objectivity and accessibility, its ability to reinforce existing inequalities through biased datasets and opaque algorithms remains a major challenge. Addressing these risks requires governance structures capable of aligning AI deployment with organisational values, legal requirements, and employee expectations (Naoum et al., 2026).

Recent research also highlights that HRM plays a central strategic role in shaping ethical AI adoption. Responsible integration is not merely a technical process but a socio-organisational one that involves developing new HR capabilities, fostering ethical cultures, and ensuring alignment between AI systems and organisational strategy (Mitra & Taherdoost, 2025). AI governance therefore becomes a strategic HR lever for shaping inclusive, transparent, and accountable work environments. Studies emphasise that organisations need hybrid governance models that balance algorithmic precision with human judgment to avoid negative consequences such as technostress, mistrust, exclusion, and reduced employee autonomy (Bhasin & Krishna, 2025). Ethical governance also enhances organisational integration and employee engagement, particularly when AI policies reflect human-centred values and fairness principles (Peethambaran et al., 2026).

Despite growing academic and practitioner interest, a gap persists in understanding how ethical AI governance within HRM can be effectively embedded into broader organisational strategy. While existing research provides insights into risks, opportunities, and governance frameworks, there is limited

examination of how HRM can strategically lead the development of trustworthy AI-enabled workplaces that prioritise inclusivity, sustainability, and human dignity. This paper addresses this gap by exploring the intersection between ethical AI governance and organisational strategy through an HRM lens.

Research Questions

- How can ethical AI governance within HRM strengthen organisational strategy in the context of digital transformation?
- What governance mechanisms and HR practices are most effective in mitigating ethical risks—such as bias, opacity, and privacy concerns—associated with AI-enabled HR systems?
- How can HRM strategically align AI adoption with organisational values to foster trustworthy, inclusive, and sustainable digital workplaces?

Research Objectives

- To examine the strategic role of HRM in designing and implementing ethical AI governance frameworks.
- To evaluate the organisational risks and ethical challenges arising from AI use in HRM and identify governance mechanisms that address them.
- To develop evidence-based guidance for organisations seeking to align AI-enabled HR practices with long-term sustainability, inclusivity, and organisational values.

Organisational strategy and digital transformation: opportunities, tensions, and impacts

A large and influential strand of research positions digital transformation as a fundamentally strategic, not merely technological, phenomenon. Early integrative work argued for the "fusion" of business and IT strategy into a single digital business strategy that shapes scope, scale, speed, and sources of value creation and capture, reframing how firms compete and perform (Bharadwaj, El Sawy, Pavlou, & Venkatraman, 2013). In parallel, process-oriented reviews conceptualise digital transformation as a multi-stage organisational journey in which digital technologies create disruptions that trigger strategic responses, structural realignments, and capability building—yielding both positive and negative outcomes depending on governance and context (Vial, 2019). Together, these perspectives emphasise that advantage under digitalisation emerges from managerial choices about where and how to compete, not from technologies per se.

Strategic lenses help clarify the capability logic underpinning digital transformation. Dynamic capabilities theory explains how firms sense technological opportunities and threats, seize them through investments and business model reconfiguration, and transform assets and routines to sustain advantage (Teece, 2007). Empirical and measurement work further operationalises these sensing–seizing–transforming capacities as predictors of innovation and business performance, reinforcing their centrality for digital change (Kump, Engelmann, Kessler, & Schweiger, 2019). Case-based strategy research shows that organisations able to orchestrate these capabilities—by clarifying digital vision, governing portfolios, and mobilising the organisation—are more likely to achieve "digital mastery" (Westerman, Bonnet, & McAfee, 2014) and to align ongoing transformation with enterprise strategy (Sia, Soh, & Weill, 2016).

A central finding across this literature is that strategy—rather than discrete technologies—drives digital maturity. Survey evidence from the MIT Sloan/Deloitte programme shows that clear, coherent digital strategy, coupled with leadership that cultivates risk-taking cultures and capability development, differentiates digitally maturing firms from laggards (Kane, Palmer, Phillips, Kiron, & Buckley, 2015). Complementary executive studies translate this into actionable design choices: where to invest, how to scale, and how to embed skills—arguing that digitally maturing firms develop talent pipelines and governance mechanisms that sustain transformation momentum (Hess, Matt, Benlian, & Wiesböck, 2016).

The strategy literature is explicit about the tensions and problems that accompany digitalisation. First are structural and competitive effects as products, services, and operations become "smart" and connected. Porter and Heppelmann show how data-rich, connected offerings reconfigure industry boundaries, alter bargaining power, and shift the basis of rivalry—forcing managers to revisit positioning, partnering, and data governance choices (Porter & Heppelmann, 2014; 2015). These changes reverberate inside the firm: product development, manufacturing, service, and IT functions are redefined, and new roles arise to manage data flows and cybersecurity, increasing coordination complexity and the risk of organisational overload (Porter & Heppelmann, 2015; Westerman et al., 2014).

Second are issues of alignment and execution. Comparative cases indicate that even when a digital vision exists, incumbents struggle to align legacy architectures, incentives, and cultural norms with new strategic logics, resulting in partial or stalled transformations (Sia et al., 2016; Hess et al., 2016). Reviews also caution that digital transformation requires rethinking strategy–structure–systems linkages: data abundance can decouple long-assumed, linear ties between strategy, structure, and information design, creating ambiguity about decision rights and accountability (Bhimani & Willcocks, 2014). This ambiguity heightens coordination costs and can undermine performance if not counterbalanced by governance that clarifies responsibilities and metrics (Westerman et al., 2014; Bharadwaj et al., 2013).

Third are capability and culture challenges. Evidence from bank transformation shows that becoming "digital to the core" requires simultaneous investments in platform modularity, agile ways of working, customer journey redesign, and talent systems that scale experimentation—essentially building a "26,000-person start-up" within an incumbent (Sia, Weill, & Xu, 2019; Sia et al., 2016). Across sectors, firms that nurture learning cultures and tolerate intelligent risk-taking are more likely to exploit digital technologies strategically; those that remain technology-centric or risk-averse often confine digital to isolated pilots, failing to reconfigure business models (Kane et al., 2015; Westerman et al., 2014).

Fourth are governance and portfolio questions. The strategy literature emphasises the need for mechanisms that integrate exploration and exploitation—balancing short-term operational wins with longer-horizon platform and data ecosystem bets (Hess et al., 2016). This is consistent with dynamic capabilities theory, which highlights managerial orchestration of assets and partners as a microfoundation of advantage under turbulence (Teece, 2007). In practice, organisations must decide whether to own platforms or participate in others' ecosystems, how to manage technical debt versus architectural renewal, and how to sequence transformation to avoid change fatigue (Westerman et al., 2014; Porter & Heppelmann, 2014).

Finally, the literature points to distributive and ethical implications that directly motivate an HRM-centred analysis. Reviews argue that ethical issues—bias, privacy, opacity, and surveillance risks—should be integrated into strategic information systems research on digital transformation rather than treated as downstream compliance concerns (Vial, 2019). Strategy reports likewise note that employees increasingly choose employers based on their digital ambition and values, making trust, transparency, and capability development strategic levers for attraction and retention (Kane et al., 2015). These themes connect directly to HR governance of AI and analytics, where choices about data use, algorithmic oversight, and worker autonomy shape organisational legitimacy and long-term performance.

In summary, the strategic literature converges on three ideas: (1) digital transformation is a strategic reconfiguration of the firm's logic of value creation that fuses business and technology choices (Bharadwaj et al., 2013; Vial, 2019); (2) competitive and organisational impacts hinge on dynamic capabilities—sensing, seizing, and transforming—and on disciplined governance that aligns culture, architecture, and portfolios (Teece, 2007; Hess et al., 2016; Westerman et al., 2014); and (3) ethical and human consequences are not peripheral but constitutive of digitally enabled strategy, with implications for trust, legitimacy, and the employment relationship (Kane et al., 2015; Vial, 2019). This synthesis sets the stage for the next subsection, where we integrate these strategy insights with HRM's evolving role in governing AI responsibly.

AI integration in recruitment: efficiency, bias, and organisational implications

Recruitment is one of the earliest HR functions where AI has been widely deployed, particularly through automated résumé screening, predictive analytics, and algorithmic decision support. Systematic reviews show that AI-enabled recruitment improves standardisation, consistency, and speed, allowing organisations to process large applicant pools more efficiently and to reduce human subjectivity in initial evaluations (Naoum, Szakadáti & Balogh, 2026). Evidence from topic-modelling research further demonstrates that automated candidate screening and interview analytics significantly enhance hiring efficiency and decision-making accuracy (Venugopala, Madhavana, Prasad & Raman, 2024). However, these benefits coexist with serious concerns about algorithmic bias. Multiple studies emphasise that AI recruitment systems may entrench historical inequalities when datasets reflect past discriminatory patterns, leading to unequal outcomes across demographic groups (Naoum et al., 2026). Ethical analyses in organisational contexts therefore recommend fairness audits, participatory design, and human oversight to mitigate bias and ensure transparent explainability of screening decisions (Aderla & Purnachander, 2024). Strategic literature also warns that organisations must align recruitment AI with broader digital

transformation goals, ensuring that tools support—not undermine—employer branding, workforce diversity, and long-term capability building (Bharadwaj, El Sawy, Pavlou & Venkatraman, 2013).

AI in performance management: data-driven evaluation, transparency challenges, and employee trust

AI-driven performance management tools enable real-time data capture, predictive performance scoring, and automated appraisal support. Recent analyses show that AI integration in performance evaluation enhances objectivity and reduces evaluator inconsistency by drawing on behavioural and productivity analytics (Venugopala et al., 2024). Yet, employees often perceive these systems as opaque and intrusive, which undermines trust. Studies highlight that the shift from human judgment to algorithmic assessment raises concerns about privacy, over-monitoring, and the legitimacy of automated decisions (Bhasin & Krishna, 2025). Organisational ethics scholarship similarly argues that transparency of AI logic, explainability of criteria, and clarity about data use are essential for maintaining employee confidence and fairness perceptions (Peethambaran, Srider & colleagues, 2026). From a strategic perspective, poorly governed AI appraisal systems can damage psychological contracts, erode employee engagement, and impede cultural change—thereby jeopardising broader organisational transformation efforts (Kane, Palmer, Phillips, Kiron & Buckley, 2015). The literature urges hybrid governance that balances algorithmic precision with managerial discretion, allowing humans to contextualise data-driven insights while ensuring accountability.

Algorithmic decision-making in HRM: governance, ethics, and organisational alignment

Beyond recruitment and appraisal, organisations increasingly deploy AI for broader decision support—such as workforce planning, talent allocation, and promotion decisions. A strategic information systems perspective conceptualises digital transformation as a process in which AI-driven disruptions necessitate clear governance structures and organisational alignment (Vial, 2019). Governance frameworks proposed in HRM emphasise transparency, fairness audits, explainability mechanisms, and employee-centric policies as prerequisites for trustworthy algorithmic decision-making (Aderla & Purnachander, 2024). Strategic capability frameworks, including dynamic capabilities theory, further suggest that organisations must develop sense–seize–transform competencies to integrate algorithmic decision systems effectively and to adapt HR processes to rapidly changing technological conditions (Teece, 2007). Ethical AI adoption requires alignment with organisational values, legal norms, and societal expectations. Studies warn that opaque decision systems may amplify accountability gaps and reduce organisational legitimacy if human oversight is weakened (Naoum et al., 2026). Conversely, ethically governed AI can enhance strategic decision quality and strengthen HRM's contribution to organisational performance.

Fairness and inclusion in AI-enabled HR practices

Fairness is a central concern across HR applications of AI. Systematic reviews confirm AI's dual potential: it can increase objectivity and accessibility in HR processes, but it can also reinforce systemic inequalities if inputs and algorithms replicate biased structures (Naoum et al., 2026). Research on sustainable HRM similarly shows that AI may support inclusion—for example, through paperless remote recruitment or accessible interfaces—while simultaneously generating new exclusion risks, such as technostress or reduced autonomy (Bhasin & Krishna, 2025). Conceptual work on human-centred HR practices emphasises that fairness must be embedded in both technological design and organisational culture. Studies grounded in social exchange and stakeholder theories demonstrate that when employees perceive AI systems as fair and transparent, commitment and engagement increase; when they perceive them as biased or obscure, distrust and resistance rise sharply (Peethambaran et al., 2026). This literature underscores that fairness is not solely a technical attribute but an organisational phenomenon shaped by governance, communication, and participatory processes.

Transparency, explainability, and the ethics of HR analytics

Transparency and explainability are widely identified as critical conditions for ethical AI in HRM. Reviews suggest that organisations often underestimate the cultural and psychological impact of opaque algorithmic systems, which can generate suspicion, fear, or perceived loss of agency among employees (Bhasin & Krishna, 2025). Strategic HRM research also indicates that transparency influences employee willingness to engage with AI-enabled systems, affecting data quality, adoption, and behavioural outcomes (Kane et al., 2015). Ethical frameworks propose clear communication strategies, audit trails, human review

procedures, and published criteria to demystify AI decision-making processes (Aderla & Purnachander, 2024). The literature repeatedly emphasises that explainability is essential for accountability—not only to correct errors but also to maintain legitimacy and compliance with regulatory expectations.

Employee trust, surveillance concerns, and organisational culture

The psychological and cultural consequences of AI-enabled HR systems are highlighted consistently across recent studies. Employee trust is shaped by perceptions of fairness, transparency, and the degree of human control over decisions (Peethambaran et al., 2026). Research links technostress, excessive monitoring, and data-driven surveillance to reduced well-being and lowered organisational identification (Bhasin & Krishna, 2025). Strategic digital transformation scholarship stresses that trust is a foundational cultural element that enables organisations to embed AI ethically and sustainably. Firms that foster cultures of learning, empowerment, and participatory governance are better able to integrate AI into HRM while maintaining psychological safety and engagement (Westerman, Bonnet & McAfee, 2014). Conversely, environments characterised by fear or low transparency tend to experience resistance, diminished data quality, and unsuccessful digital initiatives (Kane et al., 2015). This literature demonstrates that trust is not a by-product but a strategic outcome requiring deliberate design and ongoing stewardship.

Towards integrated frameworks: Strategic HRM, responsible AI, and sustainable digital workplaces

Across recruitment, performance management, and broader HR decision-making, the literature converges on the importance of integrating strategic HRM principles with responsible AI governance. Reviews urge organisations to embed AI systems within coherent digital strategies that balance technological efficiency with human-centred values (Bharadwaj et al., 2013; Vial, 2019). Observational and conceptual work further argues that ethical AI governance strengthens organisational legitimacy, enhances talent processes, promotes inclusion, and contributes to long-term sustainability when overseen by HRM as a strategic partner (Mitra & Taherdoost, 2025). In sum, the literature suggests that the future of digital workplaces will depend on hybrid human–algorithmic governance models that uphold fairness, transparency, accountability, and employee trust.

Methodology & Limitations

This study adopts a qualitative empirical research design to investigate how ethical AI governance in HRM is experienced and understood within organisational settings. A qualitative approach is appropriate because the research questions centre on employees', managers', and HR professionals' perceptions of fairness, transparency, trust, and strategic alignment — all of which are socially constructed, context-dependent, and best understood through rich, interpretive data rather than quantitative indicators.

Research design

The study employed semi-structured interviews with HR managers, line managers, and employees who work with or are affected by AI-enabled HR systems (e.g., automated screening tools, performance analytics dashboards, and algorithmic decision-support systems). The interviews explored participants' experiences with AI in recruitment, performance management, and decision-making processes, with a particular focus on perceived fairness, transparency, clarity of data use, opportunities and risks, and the organisational culture surrounding AI deployment.

Sample and recruitment

A purposive sampling strategy was used to recruit participants with direct, relevant experience of AI-enabled HR practices. A total of 18 participants were interviewed:

- 6 HR managers involved in implementing or overseeing AI tools;
- 6 line managers who use AI outputs in people management;
- 6 employees whose careers or daily work were affected by AI-based HR processes.

Participants were drawn from medium-sized and large organisations undergoing some degree of digital transformation. Recruitment occurred via organisational contacts and professional networks, ensuring diversity in roles, sectors, and exposure levels.

Data collection

Data were collected through interviews lasting 45–60 minutes, conducted either in person or online. A semi-structured protocol allowed researchers to explore core concepts while leaving space for participants to elaborate on experiences and concerns. All interviews were audio-recorded with consent and transcribed verbatim.

Data analysis

The study followed a thematic analysis approach. First, transcripts were read iteratively to become familiar with the data. Initial codes were then generated inductively, reflecting participants' lived experiences with AI tools. Codes were grouped into broader themes, such as perceived fairness of AI decisions, transparency and explainability, trust and surveillance concerns, and organisational alignment and culture. Themes were refined through constant comparison across interviews to identify convergences, divergences, and patterns.

Rationale and value of the empirical design

This empirical methodology adds value by capturing perspectives that cannot be obtained from secondary literature alone. It reveals how individuals interpret AI-enabled HR practices in everyday organisational life and how they make sense of fairness, transparency, and trust in relation to these technologies. The qualitative approach is particularly well-suited to uncovering latent fears, tacit knowledge, emotional responses, and unanticipated consequences that may not appear in formal documentation or policy statements.

By grounding the findings in real experiences, this methodology strengthens the paper's contribution to debates on ethical AI governance in HRM. It demonstrates how high-level organisational strategies translate — or fail to translate — into lived realities and offers insight into the conditions under which AI fosters or undermines legitimacy, inclusion, and employee well-being.

Limitations

This qualitative study has several limitations. The sample size, although appropriate for in-depth qualitative research, limits generalisability. Participants were selected purposively, meaning their experiences may not represent all organisational contexts or sectors. Data are based on self-reported perceptions, which may be subject to recall bias or influenced by individual attitudes toward technology. Additionally, the study draws on organisations already engaging with AI, so it may not reflect contexts where adoption is minimal or resisted. Despite these limitations, the methodology provides rich, contextualised insights into how AI-enabled HR practices are experienced, offering a foundation for future research with larger or mixed-methods designs.

Findings

The thematic analysis of the interview data revealed five core themes illustrating how participants experienced AI-enabled HRM practices in relation to fairness, transparency, trust, and organisational strategy. These findings are derived solely from the empirical material collected from HR managers, line managers, and employees.

Mixed perceptions of fairness in AI-enabled recruitment

Participants widely acknowledged that AI screening tools made recruitment faster and more consistent. HR managers described reductions in workload and improvements in standardisation when processing large applicant pools. However, employees and some line managers expressed concerns about fairness, particularly when they perceived the system as rigid or insensitive to context. Several participants questioned whether the AI could "see potential," describing fears that unconventional candidates might be filtered out without explanation. This tension between efficiency and fairness emerged across nearly all interviews.

Transparency gaps reduced trust in AI-based performance management

Across participant groups, transparency was raised as a major issue in AI-driven performance systems. Employees often reported uncertainty about which data were being collected, how performance scores were generated, and why certain indicators carried more weight than others. Line managers echoed these

sentiments, noting that they could not always explain AI-generated outputs to their teams. As a result, a sense of opacity led some employees to mistrust the system and doubt its legitimacy. HR managers acknowledged that communication around these tools remained underdeveloped, reinforcing the transparency gap.

Perceived surveillance and data intensity affected employee well-being

A recurring theme across employee interviews was the feeling of increased surveillance. Participants described discomfort with constant data capture—whether through productivity analytics, behavioural indicators, or passive monitoring systems. Some employees felt the systems were "watching too much," creating anxiety about how minor fluctuations in behaviour might be interpreted. Line managers also expressed concern that the emphasis on metrics could reduce discretion and relational judgement in performance dialogues. This perceived intensification of surveillance negatively affected employee well-being and contributed to scepticism toward AI.

Human oversight was viewed as essential for legitimacy and acceptance

Despite concerns about fairness and surveillance, participants did not reject AI outright. Instead, they emphasised the importance of maintaining human oversight. Employees stressed that AI recommendations were acceptable as long as a manager could contextualise, interpret, or override system outputs. Line managers similarly felt responsible for "adding the human layer," ensuring that algorithms did not become sole decision-makers. HR managers viewed hybrid models as necessary to maintain legal defensibility and organisational credibility. This shared preference for human–AI complementarity formed one of the strongest areas of agreement across participant groups.

Organisational strategy and culture shaped the acceptance of AI tools

The interviews revealed that employee perceptions of AI were strongly influenced by organisational communication, leadership signals, and cultural norms. In organisations where leaders articulated a clear strategic rationale for AI adoption and involved employees in early discussions, participants reported higher levels of trust and openness. Conversely, in cases where tools were implemented rapidly or communicated poorly, employees perceived AI as imposed and potentially threatening. This indicates that the broader organisational environment—not only the technology itself—plays a decisive role in shaping experiences of fairness, transparency, and trust.

Discussion

This discussion interprets the interview findings in light of the literature on digital transformation, strategic HRM, and responsible AI. It addresses the three research questions—strategic contribution, governance mechanisms, and values alignment—and, in doing so, fulfills the three research objectives.

RQ1: How ethical AI governance within HRM strengthens organisational strategy in digital transformation

The interviews showed that acceptance of AI hinged on human oversight, clear communication, and culture; when leaders articulated a rationale and invited involvement, trust increased, whereas rushed, opaque rollouts bred resistance. This pattern aligns with strategy research which frames digital transformation as a strategic (not purely technical) reconfiguration that fuses business and technology choices into a single digital business strategy (Bharadwaj, El Sawy, Pavlou, & Venkatraman, 2013). Reviews emphasise that transformation is a multi-stage journey of disruption, response, realignment, and capability development; outcomes depend on governance and context rather than tools alone (Vial, 2019).

Dynamic capabilities—sensing, seizing, and transforming—explain why firms that embed ethical AI governance into HR decision-making create strategic advantage: they scan for ethical risks and regulatory shifts (sensing), invest in explainable and auditable systems (seizing), and reconfigure roles, policies, and skills to sustain legitimacy and performance (transforming) (Teece, 2007). Empirically and in executive studies, digitally maturing firms succeed by coupling a clear digital strategy with leadership that cultivates learning and prudent risk-taking cultures—conditions the participants associated with greater trust and adoption (Kane, Palmer, Phillips, Kiron, & Buckley, 2015; Westerman, Bonnet, & McAfee, 2014). In banking, for instance, becoming "digital to the core" required governance, portfolio discipline, and enterprise-wide

mobilisation—echoing the theme that legitimacy rests on visible human stewardship and transparent rationale (Sia, Soh, & Weill, 2016).

Implications for strategy: HR-led ethical AI governance strengthens organisational strategy by (a) improving talent attraction and retention through trust and legitimacy, (b) reducing transformation risk via disciplined oversight and communication, and (c) enabling scalable, compliant data-driven HR practices that support enterprise goals (Bharadwaj et al., 2013; Vial, 2019; Kane et al., 2015). This addresses Objective 1 (examining HRM's strategic role) by showing HR's stewardship of ethics, culture, and capability building as a strategic lever.

RQ2: Which governance mechanisms and HR practices mitigate ethical risks (bias, opacity, privacy) in AI-enabled HR systems?

The data surfaced recurring concerns about fairness in recruitment, opacity in performance scores, and surveillance anxiety. The literature converges on concrete mitigations:

Fairness and bias mitigation. Systematic reviews in HRM confirm AI's dual potential: it can standardise and speed decisions, yet reproduce inequalities when training data and modelling choices reflect past discrimination; recommended mitigations include pre-deployment dataset audits, disparate-impact testing, ongoing model monitoring, and human-in-the-loop decision points (Naoum, Szakadáti, & Balogh, 2026; Aderla & Purnachander, 2024). The interviewees' desire for contextual human review is consistent with these hybrid models.

Transparency and explainability. Employee acceptance rises when criteria, data sources, and model rationales are communicated in accessible language and when managers can explain outputs; black-box systems erode trust and quality of engagement (Peethambaran, Srider, & colleagues, 2026; Kane et al., 2015). Participants explicitly called out the need to understand what data are collected and how scores are produced.

Privacy and proportionality. Performance analytics should follow purpose limitation and data minimisation; over-monitoring and opaque passive capture produce technostress and lower identification (Bhasin & Krishna, 2025).

Portfolio and accountability. Governance must clarify decision rights, roles (model owner, HR analytics lead, ethics reviewer), escalation channels, audit cadence, and criteria for pausing/rolling back models—treating algorithmic HR decisions as high-stakes portfolio assets requiring structured oversight (Hess, Matt, Benlian, & Wiesböck, 2016; Vial, 2019).

Data and ecosystem choices. Smart, connected, data-rich systems expand boundaries and heighten strategic choices about data governance, security, and partnerships; firms must define what data they collect, how it is secured, and who has access, or risk legitimacy loss (Porter & Heppelmann, 2014; 2015).

With recruitment specifically, evidence shows efficiency gains but persistent fairness risk, underscoring the case for participatory design, candidate-facing explanations, and recourse mechanisms alongside manager override controls (Naoum et al., 2026; Venugopala, Madhavana, Prasad, & Raman, 2024). With performance management, transparency of indicators and weightings, clear purpose boundaries, and human review of consequential outcomes speak directly to the legitimacy gaps the interviewees reported (Bhasin & Krishna, 2025; Peethambaran et al., 2026).

This answers RQ2 and meets Objective 2 by specifying mechanisms (audits, explainability, human oversight, privacy safeguards, and portfolio accountability) that target bias, opacity, and privacy risk.

RQ3: How HRM aligns AI adoption with organisational values to foster trustworthy, inclusive, and sustainable digital workplaces

The interviews revealed that perceived alignment—between stated values and AI practices—shaped trust. The literature similarly argues that success depends on embedding AI within coherent digital strategies and human-centred values: digital maturity arises when organisations link technology choices to purpose, people, and culture rather than isolated tools (Bharadwaj et al., 2013; Vial, 2019).

Inclusion and equity. Reviews on AI and DEI in HRM recommend treating fairness as an organisational property: design choices, communication, and participatory processes influence whether employees perceive AI as equitable; ongoing bias mitigation and employee voice are central to inclusion (Naoum et al., 2026; Peethambaran et al., 2026). Participants' concern that AI might miss "potential" reinforces the need to preserve contextual human judgement and provide candidate/employee recourse.

Trust and sustainability. Strategy research links trust-centric cultures to digital maturity and sustained performance; employees prefer employers committed to digital progress with clear values and upskilling pathways—conditions associated with retention and engagement (Kane et al., 2015; Westerman et al., 2014). Conversely, culture–technology misalignment increases resistance and stalls change, a dynamic the interviewees described when rollouts were rushed or communicated poorly.

Practical alignment model. To operationalise values alignment, HR can implement a five-step cycle: (1) articulate values-linked AI use-cases and red lines; (2) co-design with employees and managers to surface contextual risks; (3) establish transparency artefacts (plain-language summaries, feature rationales, recourse channels); (4) run fairness and privacy reviews pre- and post-deployment with defined human-in-the-loop checkpoints; (5) publish adoption and impact metrics tied to inclusion and well-being (Naoum et al., 2026; Bhasin & Krishna, 2025; Kane et al., 2015). This answers RQ3 and fulfills Objective 3 by translating values alignment into actionable guidance aimed at trustworthy, inclusive, and sustainable workplaces.

Integration back to empirical themes

Each discussion strand above ties directly to the five interview themes: mixed fairness perceptions (recruitment), transparency gaps (performance), surveillance concerns (well-being), the centrality of human oversight, and the decisive role of leadership communication and culture. The scholarly sources corroborate participants' experiences and convert them into concrete strategic and governance recommendations.

Actionable guidance mapped to the objectives

Objective 1 (strategic role of HRM). Position HR as the steward of ethical AI capabilities: lead enterprise sensing of ethical/regulatory shifts, chair cross-functional AI governance forums, and embed learning cultures and upskilling tied to digital strategy (Bharadwaj et al., 2013; Sia et al., 2016).

Objective 2 (evaluate risks and mechanisms). Institutionalise an HR AI control framework: bias audits and disparate-impact testing; explainability standards and manager-readiness to interpret outputs; privacy-by-design with data minimisation and purpose limitation; human-in-the-loop sign-offs for high-stakes decisions; post-deployment monitoring and an incident response path (Naoum et al., 2026; Aderla & Purnachander, 2024; Hess et al., 2016).

Objective 3 (evidence-based guidance for sustainable performance). Define values-to-metrics alignment: translate inclusion, transparency, and well-being into dashboard indicators (e.g., explanation coverage, recourse utilisation, monitored disparate impact, training completion, perceived fairness/trust pulse scores), and tie them to portfolio decisions about scaling or pausing models (Peethambaran et al. 2026;, Kane et al., 2015; Vial, 2019).

Limitations and future research

While the discussion integrates qualitative findings with established frameworks, transferability may vary across sectors and jurisdictions. Future work could compare sectors, test the proposed governance metrics quantitatively, and examine long-term effects of human-in-the-loop designs on inclusion and engagement.

Conclusion

This study examined how employees, line managers, and HR professionals experience AI-enabled HRM practices and how these perceptions relate to fairness, transparency, and trust. The empirical insights indicate that while AI can streamline recruitment and support performance evaluation, its ethical and strategic value depends on how it is implemented and communicated. Concerns about opacity, bias, and surveillance arise not solely from the technology but from organisational choices surrounding design, governance, and communication.

Human oversight, transparent processes, and alignment with organisational values emerged as core conditions for building trustworthy AI-enabled workplaces. These findings underscore that ethical AI governance is central to organisational strategy and to sustaining legitimacy, employee engagement, and well-being in digital transformations. Future research may deepen understanding through larger samples, comparative studies across sectors, or mixed-method approaches examining long-term impacts.

References

Aderla, N. & Purnachander, M. (2024) AI-powered human resource management: Innovation, ethics, and organisational sustainability. JETIR.

Bharadwaj, A., El Sawy, O.A., Pavlou, P.A. & Venkatraman, N. (2013) 'Digital business strategy: Toward a next **generation of insights', MIS Quarterly, 37(2), pp. 471–482.**

Bhasin, S. & Krishna, K. (2025) 'The integration of artificial intelligence in human resource management for sustainable workplaces', International Journal of Research in Human Resource Management, 7(2), pp. 1–7.

Bhimani, A. & Willcocks, L. (2014) 'Digitisation, "big data" and the transformation of accounting information', Accounting and Business Research, 44(4), pp. 469–490.

Hess, T., Matt, C., Benlian, A. & Wiesböck, F. (2016) 'Options for formulating a digital transformation strategy', MIS Quarterly Executive, 15(2), pp. 123–139.

Kane, G.C., Palmer, D., Phillips, A.N., Kiron, D. & Buckley, N. (2015) Strategy, not technology, drives digital transformation. MIT Sloan Management Review & Deloitte University Press.

Kump, B., Engelmann, A., Kessler, A. & Schweiger, C. (2019) 'Toward a dynamic capabilities scale: Measuring organisational sensing, seizing and transforming capacities', Industrial and Corporate Change, 28(5), pp. 1149–1172.

Mitra, M. & Taherdoost, H. (2025) 'Ethical theories, governance models, and strategic frameworks for responsible AI adoption and organizational success', Frontiers in Artificial Intelligence, 8.

Naoum, R., Szakadáti, T. & Balogh, G. (2026) 'Artificial Intelligence (AI) in human resource management (HRM): A systematic review of its dual impact on diversity, equity, and inclusion', Management Review Quarterly, Published 8 January.

Peethambaran, M., Srider, P. & colleagues (2026) 'The ethical frontier: Balancing AI innovation and human-centric HR practices', International Journal of Organizational Analysis.

Porter, M.E. & Heppelmann, J.E. (2014) 'How smart, connected products are transforming competition', Harvard Business Review, November.

Porter, M.E. & Heppelmann, J.E. (2015) 'How smart, connected products are transforming companies', Harvard Business Review, October.

Sia, S.K., Soh, C. & Weill, P. (2016) 'How DBS Bank pursued a digital business strategy', MIS Quarterly Executive, 15(2), pp. 105–121.

Teece, D.J. (2007) 'Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance', Strategic Management Journal, 28(13), pp. 1319–1350.

Venugopala, M., Madhavana, V., Prasad, R. & Raman, R. (2024) 'Transformative AI in human resource management: Enhancing workforce planning with topic modelling', Cogent Business & Management, 11(1).

Vial, G. (2019) 'Understanding digital transformation: A review and research agenda', Journal of Strategic Information Systems, 28(2), pp. 118–144.

Westerman, G., Bonnet, D. & McAfee, A. (2014) Leading digital: Turning technology into business transformation. Harvard Business Press.