

Can Governance, Ethics, and System Design Make AI in HR Truly Trustworthy?

Sapheya Aftimos and Dunya Hassoun

Australian University, West Mishref, Safat 13015, Kuwait

ABSTRACT

AI is rapidly transforming HRM, with organizations and universities adopting AI systems to recruit, evaluate, and analyze their workforces. While these technologies promise efficiency, flaws or biases in algorithms can affect individuals' careers and organizational performance, turning technological improvement into a matter of trust. Without proper governance, AI risks perpetuating bias and undermining accountability. Yet current research focuses largely on efficiency and cost savings, and the lack of integration between HRM and IT perspectives means organizations often implement AI without considering how governance structures and design decisions affect people. This study addresses that gap by investigating the governance of AI systems in HR through four questions: how governance structures shape trust, what role ethical frameworks play in mitigating bias, how system design influences transparency and accountability, and whether demographic and professional differences shape trust. Applying a socio-technical systems analysis to survey data from organizational and higher education participants, the findings show near-universal support for governance frameworks, with trust significantly associated with professional role, AI familiarity, and governance orientation.

Keywords: Artificial intelligence, Human resource management, Trustworthiness, Explainability, Bias and fairness, Governance, System design

INTRODUCTION

Artificial intelligence (AI) is rapidly reshaping human resource management (HRM), with organizations increasingly using AI for recruitment, evaluation, and workforce analytics. While these technologies offer efficiency and decision-support benefits, they also raise concerns regarding fairness, transparency, accountability, and bias. As AI influences decisions affecting employees, trust becomes a critical consideration. AI does not operate in isolation but within a socio-technical system involving human judgment, organizational policies, and institutional priorities. However, existing research has largely emphasized efficiency and performance outcomes, with comparatively less attention given to governance, ethics, and trust. Furthermore, HRM and IT scholarship often examine AI from different perspectives, creating a need for more integrated approaches. This study addresses this gap by examining how governance, ethical frameworks, and system design contribute to trustworthy AI in HR. Drawing on survey data and a socio-technical perspective, it explores the conditions under which AI can be trusted in

decisions affecting people. The study contributes by integrating HRM and IT perspectives, positioning governance as a key driver of trust, and offering practical insights for organizations seeking to adopt AI responsibly while maintaining transparency, accountability, and fairness.

RESEARCH QUESTIONS

This study examines the factors that influence trust in AI-supported HRM. Specifically, it investigates the role of governance, ethics, and system design in shaping perceptions of trustworthy AI. To examine how organizational and socio-technical factors shape trust in AI-supported HRM, this study addresses the following research questions:

1. How do professionals perceive the role of governance in building trust in AI-driven HR systems?
2. What ethical concerns are most associated with AI use in HR?
3. Which system design features are considered most important for trustworthy AI in HR?
4. Do demographic and professional differences influence trust in AI-supported HR systems?

LITERATURE REVIEW

AI has expanded rapidly into HR processes, with one recent estimate suggesting that as many as 86% of employers have adopted technology-driven interviewing and screening tools (Mujtaba & Mahapatra, 2025). This adoption has generated debate around governance, ethics, and transparency, and scholarship has shifted from emphasizing efficiency gains toward examining the conditions under which AI systems can be considered trustworthy when their outputs directly affect people (Bujold et al., 2023).

Governance and Trust in AI-Driven HR

Governance frameworks have emerged as a foundational determinant of trust in AI-supported HR decisions. Berg and Johnston (2025) caution that flawed algorithms, incomplete datasets, and opaque design can undermine the fairness and legitimacy of HR processes, calling for governance built on worker participation, transparency, and accountability. Yet governance tends to be reactive: organizations adopting AI without clear oversight face discrimination and reputational risks (Sony et al., 2025), prompting calls for governance that extends beyond compliance to proactive monitoring, ethical audits, and stakeholder engagement (Hunkenschroer & Luetge, 2022). External governance is also gaining prominence, as regulatory instruments such as the EU AI Act, which classifies AI in employment as high-risk and imposes transparency and human-oversight requirements, signal a shift from voluntary policies toward enforceable controls operating at both organizational and societal levels (Laux, 2024; Malin et al., 2025).

Ethical Concerns: Bias, Fairness, and Accountability

Ethical considerations are increasingly viewed as central to the trustworthiness of AI in HR. Varma et al. (2023) argue that AI adoption in people management must be evaluated not only for efficiency but also for fairness, inclusivity, and employee well-being, while Bar-Gil et al. (2024) show that without safeguards embedded throughout the AI lifecycle, AI systems risk reinforcing existing inequalities. Bias remains among the most significant concerns: systems trained on historical organizational data may reproduce existing disparities, so responsible implementation requires representative training data, explainable algorithms, and mechanisms to mitigate discriminatory outcomes (Bujold et al., 2023; Sony et al., 2025). Because AI-related bias can evolve as systems interact with new data, continuous monitoring and regular bias assessments are also necessary to ensure decisions remain fair throughout the system lifecycle (Malin et al., 2025).

System Design Features for Trustworthy AI

Beyond governance structures, research highlights four key design features that influence the trustworthiness of AI in HR. First, transparency and explainability enable employees, applicants, and HR professionals to understand and challenge AI-supported decisions. Trust depends not only on explainable algorithms and auditable decision processes but also on clear organizational communication about how AI is used and what safeguards are in place (Malin et al., 2025; Sony et al., 2025). Second, fairness-oriented design requires representative datasets, fairness-aware algorithms, and ongoing bias auditing to reduce discriminatory outcomes (Mujtaba & Mahapatra, 2025). Third, human oversight can be embedded through mechanisms ranging from human review of decisions to human control over system use (Malin et al., 2025; Laux, 2024). Finally, data security and lawful data handling are widely regarded as essential foundations for trust (Hunkenschroer & Luetge, 2022; Malin et al., 2025). Although the literature consistently identifies these features, limited evidence exists regarding their relative importance, and stakeholder priorities may vary across HR functions and organizational contexts (Hunkenschroer & Luetge, 2022).

Individual Differences in Trust in AI-Supported HR

The literature suggests that trust in AI is not uniform across individuals. Research indicates that familiarity with AI and frequency of technology use are important predictors of trust, while older individuals tend to express greater skepticism toward AI-supported decision-making (Alasmari et al., 2025). Professional position also matters, as stakeholder groups within the same organization, such as HR professionals, applicants, and workers' representatives, often articulate distinct requirements before accepting AI in HR processes (Malin et al., 2025). However, this evidence derives largely from general-population or single-stakeholder samples rather than HR professionals evaluating AI in their own decision-making contexts. Taken together, the literature establishes governance, ethical safeguards, and system design as determinants of trust in AI-driven HR, but these strands have developed largely in isolation, with studies rarely asking professionals to

weigh these factors against one another. This study addresses that gap by examining how professionals perceive governance, which ethical concerns and design features they prioritize, and whether demographic and professional characteristics shape these perceptions.

CONCEPTUAL FRAMEWORK AND HYPOTHESES

Conceptual Framework

The conceptual framework is grounded in literature showing that trust in AI-driven HR systems is shaped by governance structures, ethical AI practices, and system design features such as transparency, accountability, fairness, and human-centered design (Rai, 2020; Budhwar et al., 2022). Accordingly, this study positions these three dimensions, governance, ethical practices, and system design, as the primary factors influencing trust in AI-driven HR.

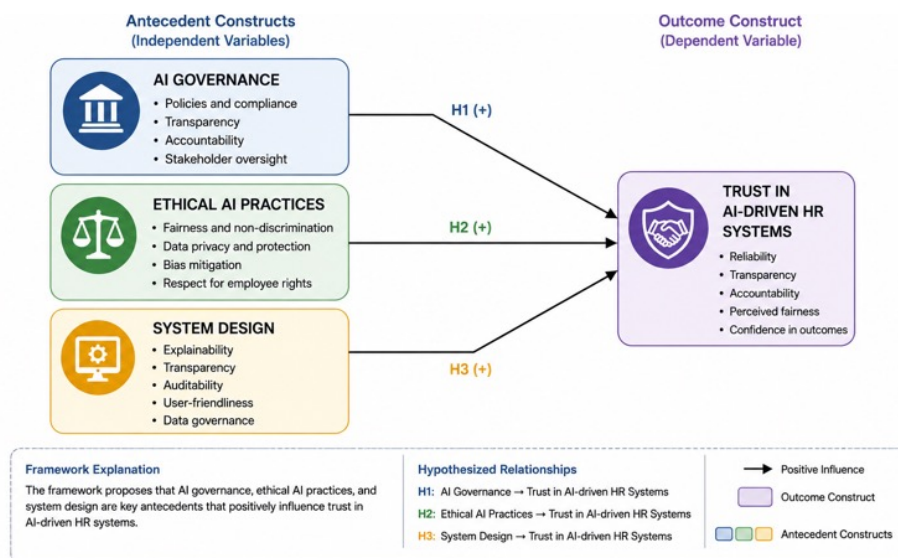


Figure 1: Conceptual framework: determinants of trust in AI-driven HR systems.

The framework in Figure 1 illustrates the proposed relationships between governance, ethics, system design, and trust in AI-driven HR systems. The model suggests that stronger governance mechanisms, ethical AI practices, and transparent system design positively influence perceptions of trust in AI-supported HR decision-making.

Hypotheses Development

Existing literature identifies transparency, accountability, and regulatory oversight as essential components of trustworthy AI implementation

(Rai, 2020), and prior studies emphasize the role of governance in AI-supported workforce strategies (Aftimos & Diab-Bahman, 2025). In HR contexts, governance frameworks may help ensure that AI-driven decisions remain fair, explainable, and aligned with organizational policies. Accordingly, the following hypothesis is proposed:

H1: Strong AI governance frameworks positively influence trust in AI-driven HR systems.

Ethical concerns surrounding AI in HR, particularly bias, discrimination, and fairness, have gained prominence in academic and professional discussions. Ethical AI practices can strengthen confidence in AI-supported decision-making by promoting fairness, inclusivity, and responsible data use (Budhwar et al., 2022), and ethical safeguards have been linked to sustainable digital transformation in organizational settings (Aftimos & Khafash, 2026). Therefore, the following hypothesis is proposed:

H2: Ethical AI practices positively influence trust in AI-driven HR systems.

System design is also considered a critical factor influencing AI trustworthiness. Features such as explainability, transparency, auditability, and user-friendly interfaces can improve user understanding and acceptance of AI-supported HR systems (Kellogg et al., 2020). Human-centered and transparent system design may reduce uncertainty and increase confidence in AI-generated outcomes. Therefore, the following hypothesis is proposed:

H3: Transparent and user-friendly system design positively influences trust in AI-driven HR systems.

METHODOLOGY

Research Design

This study employs a quantitative, exploratory design to examine trust in AI-driven HRM, focusing on the role of governance, ethics, and system design in shaping trustworthy AI practices. A survey-based approach was used to capture participant perceptions and attitudes (Creswell, 2014), while the exploratory design reflects the evolving nature of AI governance and increasing reliance on AI-supported decision-making. The study also adopts a socio-technical perspective, recognizing that trust in AI depends not only on technical capability but also on organizational structures, governance, ethics, and human interaction (Trist, 1981). The empirical analysis is complemented by recent literature on responsible AI in HRM.

Sample and Data Collection

Data were collected over a one-month period via an online Microsoft Forms questionnaire targeting HR professionals, IT professionals, academics, and managers across international contexts. A non-probability convenience

sample was used given the study's exploratory nature and access through professional and academic networks; participation was voluntary and informed. The final dataset comprised 152 valid responses.

Survey Instrument

The survey instrument examined perceptions of AI governance, ethics, system design, and trust in AI-supported HR systems. It included demographic questions, multiple-choice and Likert-scale questions measuring participant perceptions regarding fairness, transparency, accountability, bias, and trust in AI-driven HR practices. One open-ended question was included to seek recommendations for enhancing trust in AI for HR.

Data Analysis Techniques

Data were analyzed in SPSS Statistics version 22. Descriptive statistics summarized response trends, and cross-tabulation with chi-square tests examined relationships between demographics, AI familiarity, governance perceptions, ethical concerns, and trust. Open-ended responses were thematically reviewed for recurring patterns in transparency, accountability, and human oversight to contextualize the quantitative findings.

Ethical Considerations

Ethical considerations were incorporated throughout the research process. Participation in the study was voluntary, and respondents were informed about the purpose of the research before completing the survey. The questionnaire included an informed consent statement, and participants were assured that their responses would remain anonymous and confidential. No personally identifiable information was collected, and the data was used solely for academic research purposes.

RESULTS

Respondent Demographics

The sample skewed toward mid-career professionals (44.7% aged 31–40; 23.7% aged 41–50), with 63.8% reporting more than ten years of professional experience. Gender distribution was balanced (50.7% male, 49.3% female). HR professionals (30.3%), academic faculty (24.3%), and IT professionals (21.1%) represented the largest occupational groups, while the remaining respondents came from research, banking, marketing, administration, and management. Most participants reported prior exposure to AI applications in HR, with 80.9% indicating they were somewhat or very familiar with such tools.

Table 1: Respondent profile.

Variable	Category	Frequency	Percentage (%)
Age	20–30	19	12.5
	31–40	68	44.7
	41–50	36	23.7
	Above 50	29	19.1
Gender	Male	77	50.7
	Female	75	49.3
Professional Role	Human Resources	46	30.3
	Academic Faculty	37	24.3
	Information Technology	32	21.1
	Researcher	15	9.9
	Other	22	14.5
Years of Experience	Less than 5 years	21	13.8
	5–10 years	34	22.4
	10–15 years	41	27.0
	More than 15 years	56	36.8

Familiarity, Governance, and Trust Perceptions of AI in HR

Governance frameworks drew strong consensus: 66.4% ($n = 101$) called them essential and 28.9% conditionally agreed, a combined 95.3%. The most critical mechanisms were transparent reporting of AI decisions (30.9%) and independent audits (27.6%), followed by clear regulatory standards (21.1%) and employee consent and data protection (20.4%). Concern about algorithmic bias was near-universal (94.7%): 40.1% moderately, 35.5% somewhat, and 19.1% very concerned. Privacy and confidentiality ranked as the leading ethical priority (40.1%), followed by fairness and non-discrimination (25.7%), accountability and explainability (17.8%), and human oversight (16.4%). Overall, 65.8% agreed or strongly agreed that AI can improve fairness in HR decisions.

Table 2: Key perceptions of AI adoption and trustworthiness in HR.

Variable	Most Frequent Response	Frequency	Percentage (%)
Familiarity with AI applications in HR	<i>Somewhat familiar</i>	81	53.3
AI can improve fairness in HR decisions	<i>Agree</i>	62	40.8
Governance frameworks are essential for trustworthy AI	<i>Yes</i>	101	66.4
AI will become more trustworthy within five years	<i>Yes</i>	95	62.5

Ethics and System Design Perceptions of AI in HR

Respondents favored continuous monitoring and evaluation as the key trust-enabling design feature (36.8%), followed by human-in-the-loop oversight (29.6%) and explainable AI (26.3%), while 82.9% rated user-friendly design as important or extremely important, underscoring usability as a critical adoption factor. Looking ahead, 62.5% believed AI in HR will become more trustworthy within five years given proper governance.

Table 3: Perceptions of bias and user-friendly design in AI-supported HR systems.

Variable	Category	Frequency	Percentage (%)
Concern about bias in AI-driven HR systems	<i>Very concerned</i>	29	19.1
	<i>Moderately concerned</i>	61	40.1
	<i>Somewhat concerned</i>	54	35.5
	<i>Not concerned</i>	8	5.3
Importance of user-friendly design	<i>Extremely important</i>	48	31.6
	<i>Important</i>	78	51.3
	<i>Neutral</i>	24	15.8
	<i>Not important</i>	2	1.3

HR Functions Benefiting Most From AI Integration

Training and development (50.7%), performance management (48.0%), recruitment and selection (40.1%), and employee engagement (36.2%) were identified as the functions benefiting most from AI integration (Table 4), suggesting respondents view AI's value as concentrated in routine, data-rich HR processes.

Table 4: Perceived benefits of AI across HR functions.

Variable	Category	Frequency	Percentage (%)
HR functions benefiting most from AI	<i>Training & Development</i>	77	50.7
	<i>Performance Management</i>	73	48.0
	<i>Recruitment & Selection</i>	61	40.1
	<i>Employee Engagement</i>	55	36.2

Cross-Tabulation and Chi-Square Findings

Chi-square analyses revealed four significant associations. Professional role was associated with both future trust in AI ($\chi^2 = 43.51$, $df = 26$, $p = .017$) and fairness beliefs ($\chi^2 = 102.52$, $df = 52$, $p < .001$), with IT professionals reporting more positive perceptions than HR and administrative respondents. AI familiarity was also significantly related to future trust ($\chi^2 = 30.93$, $df = 4$, $p < .001$) and fairness beliefs ($\chi^2 = 79.26$, $df = 8$, $p < .001$), indicating that greater familiarity with AI is associated with stronger confidence in its trustworthiness and fairness. In addition, governance orientation was

significantly associated with future trust ($\chi^2 = 72.11$, $df = 4$, $p < .001$), with respondents who viewed governance as essential reporting higher levels of trust in AI. No significant associations were found for gender, experience level, or the relationship between ethical principles and design feature preferences, suggesting that these perceptions are relatively consistent across demographic groups.

Respondent Recommendations

Of 152 respondents, 144 provided open-ended recommendations on how to strengthen trust in AI for HR. The most frequent calls were for clearer regulation, standards, and compliance frameworks ($n = 26$), independent audits and ongoing monitoring ($n = 25$), human-in-the-loop oversight with final human review ($n = 24$), and explicit safeguards against bias and discrimination ($n = 23$). Transparency and explainability of AI outputs ($n = 18$) and AI training and literacy for HR practitioners ($n = 14$) were also raised repeatedly, alongside narrower references to accountability ($n = 11$) and data privacy ($n = 8$). These qualitative findings reinforce the quantitative results, indicating that respondents view trustworthy AI as dependent on a combination of safeguards rather than any single one.

DISCUSSION

The findings highlight the role of governance, ethics, and system design in shaping trust in AI-driven HRM. Although AI is widely valued for efficiency and decision support, respondents placed greater emphasis on transparency, accountability, fairness, and human oversight when assessing trustworthy AI in HR. Respondents also viewed AI as enhancing both operational and strategic HR activities, particularly training, performance management, and recruitment. At the same time, participants expressed generally positive attitudes toward adoption of AI and its future potential when supported by appropriate safeguards. The findings support literature linking governance to trust in AI-supported systems (Rai, 2020). Respondents viewed transparency, compliance, accountability, and regulatory oversight as essential, suggesting that organizations are more likely to adopt AI in HR when governance mechanisms clearly define responsibility, monitoring, and ethical boundaries. This aligns with prior work on governance in AI-supported workforce strategies and digital transformation (Zabalawi et al., 2024). The significant association between governance orientation and future trust provides direct support for H1, while the prominence respondents assigned to ethical safeguards and to design features such as explainability, continuous monitoring, and user-friendliness offers exploratory support for H2 and H3. Ethical concerns also remained central to perceptions of AI trustworthiness. Many respondents expressed concerns regarding bias, fairness, discrimination, and transparency in AI-driven HR systems. Interestingly, participants

simultaneously viewed AI as a mechanism that can improve fairness in HR decision-making while acknowledging the risks of algorithmic bias. This finding suggests that respondents recognize both the opportunities and challenges associated with AI, reinforcing the need for governance structures that promote fairness while mitigating unintended discriminatory outcomes. These findings echo recent work emphasizing ethical governance, stakeholder trust, accountability, and sustainability-oriented decision-making (Bar-Gil et al., 2024). Respondents further emphasized user-friendly interfaces, explainability, continuous monitoring, and human-in-the-loop oversight. This supports literature suggesting that transparent and human-centered design improves acceptance of AI-supported technologies (Kellogg et al., 2020) and reinforces socio-technical perspectives that trust is shaped by the interaction between technology, organizational processes, and human users (Trist, 1981).

CONCLUSION

This study examined how governance frameworks, ethical AI practices, and system design influence trust in AI-driven human resource management. The findings highlight that trust in AI-supported HR decisions is shaped not only by technological capabilities but by transparency, accountability, fairness, and human-centered design. The study contributes to ongoing discussions on responsible and trustworthy AI by shifting attention from efficiency-centred narratives toward governance and ethical considerations in HR decision-making. The findings may support policymakers, HR professionals, and system designers in developing AI-supported HR systems that promote fairness, transparency, and trust. These findings should nevertheless be interpreted in light of the study's limitations: the non-probability convenience sample limits generalizability, and the exploratory analyses relied on categorical, single-item measures. Future research may explore these relationships across different industries, organizational contexts, and longitudinal settings, using probability samples and validated multi-item scales.

ACKNOWLEDGMENT

The authors would like to acknowledge the participants for their valuable contributions to this study.

REFERENCES

- Aftimos, S. and Diab-Bahman, R., 2025. AI-driven talent management: Transforming workforce strategies in the digital age. In: M. Lowe and Y. Luximon, eds. *Creativity, Innovation and Entrepreneurship*. AHFE International Conference. AHFE Open Access, vol. 166. USA: AHFE International. <https://doi.org/10.54941/ahfe1006110>
- Aftimos, S. and Khafash, M., 2026. Embedding talent management in the digital transformation of universities in Kuwait. In: *Organizational Change, Innovation, and Consumer Behavior in the Digital Economy*. Hershey, PA: IGI Global. <https://doi.org/10.4018/979-8-2600-1453-0.ch002>

- Alasmari, A. A., Alruwaili, R. F., Alotaibi, R. F., Youssef, I. K. and Asklany, S. A., 2025. Demographic influences on trust in artificial intelligence across cognitive domains: A statistical perspective. *PLOS ONE*, 20(11), Article e0331003. <https://doi.org/10.1371/journal.pone.0331003>
- Bar-Gil, O., Ron, T. and Czerniak, O., 2024. AI for the people? Embedding AI ethics in HR and people analytics projects. *Technology in Society*, 77, 102527. <https://doi.org/10.1016/j.techsoc.2024.102527>
- Berg, J. and Johnston, H., 2025. AI in human resource management: the limits of empiricism. ILO Working Paper 154. Geneva: International Labour Office. <https://doi.org/10.54394/NMSH7611>
- Budhwar, P., Malik, A., De Silva, M.T. and Nyfoudi, M., 2022. AI and the changing nature of human resource management: Implications for organizations and employees. *Human Resource Management Review*, 32(4), p. 100876.
- Bujold, A., Roberge-Maltais, I., Parent-Rochelleau, X., Boasen, J., Sénécal, S., and Léger, P.-M., 2023. Responsible artificial intelligence in human resources management: A review of the empirical literature. *AI and Ethics*, 4, 1185–1200. <https://doi.org/10.1007/s43681-023-00325-1>
- Creswell, J.W., 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 4th ed. Thousand Oaks, CA: Sage Publications.
- Hunkenschroer, A. L., and Luetge, C., 2022. Ethics of AI-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, 178(4), 977–1007. <https://doi.org/10.1007/s10551-022-05049-6>
- Kellogg, K.C., Valentine, M.A. and Christin, A., 2020. Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), pp. 366–410.
- Laux, J., 2024. Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & Society*, 39, 2853–2866. <https://doi.org/10.1007/s00146-023-01777-z>
- Malin, C., Fleiß, J., and Thalmann, S., 2025. Stakeholder-specific adoption of AI in HRM: Workers' representatives' perspective on concerns, requirements, and measures. *Frontiers in Artificial Intelligence*, 8, Article 1561322. <https://doi.org/10.3389/frai.2025.1561322>
- Mujtaba, D. and Mahapatra, N., 2025. Fairness in AI-driven recruitment: Challenges, metrics, methods, and future directions. arXiv preprint arXiv:2405.19699. [online] Available at: <https://arxiv.org/abs/2405.19699>
- Rai, A., 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), pp. 137–141.
- Sony, M. M. A. A. M., Amin, M. B., Ashraf, A., Islam, K. M. A., Debnath, N. C. and Debnath, G. C., 2025. Bias in AI-driven HRM systems: Investigating discrimination risks embedded in AI recruitment tools and HR analytics. *Social Sciences & Humanities Open*, 12, Article 102082.
- Trist, E., 1981. The evolution of socio-technical systems. *Occasional Paper*, 2, pp. 1–67.
- Varma, A., Dawkins, C. and Chaudhuri, K., 2023. Artificial intelligence and people management: A critical assessment through the ethical lens. *Human Resource Management Review*, 33(1), Article 100923. <https://doi.org/10.1016/j.hrmr.2022.100923>
- Zabalawi, I., Kordahji, H. and Aftimos, S., 2024. Digital transformation in universities: Strategic framework, implementation tools, and leadership. In: *Higher Education in the Arab World: Digital Transformation*. Cham: Springer Nature Switzerland, pp. 145–210.