Journal of Economics
AND MANAGEMENT SCIENTIES
https://jems.ink
ISSN 2655-1934 (print), 2655-6685 (online)

AI-Powered Performance Appraisal: Balancing Automation with Human Judgment in Performance Management Systems

Andry Novrianto[1], Deval Gustion[2], and Elmi Rahmawati[3]*

[1,2,3] *Universitas Putra Indonesia "YPTK" Padang, Indonesia*

ARTICLE HISTORY

Received: 22 June 25
Final Revision: 25 June 25
Accepted: 27 June 25
Online Publication: 30 June 25

KEYWORDS

Artificial Intelligence, Performance Appraisal, Human Judgment, Hybrid Evaluation, Ethical Decision-Making

A B S T R A C T

This study aims to explore how organizations integrate Artificial Intelligence (AI) and human judgment within employee performance appraisal systems to achieve outcomes that are fair, efficient, and contextually informed. Using a qualitative case study approach involving three organizations based in Jakarta, Surabaya, and Singapore, data were collected through semi-structured interviews, internal document analysis, and observations of performance review panels. Each organization utilized AI-driven human resource management platforms such as SAP SuccessFactors and IBM Watson Talent Insights, while maintaining significant human involvement in final appraisal decisions. Thematic analysis revealed five major themes: trust in AI, human override, fairness, emotional fit, and ethical concerns. The findings indicate that although AI enhances consistency and efficiency, qualitative dimensions such as leadership, collaboration, and cultural alignment still require human interpretation. The study also introduces a hybrid appraisal model that combines AI and human scores, adjusted by an ethical risk coefficient. These results contribute empirical insights into contemporary appraisal practices and emphasize the importance of algorithmic transparency and ethical sensitivity in the implementation of AI-based systems within human resource management.

KATA KUNCI

Kecerdasan Buatan, Penilaian Kinerja, Pertimbangan Manusia, Evaluasi Hibrida, Pengambilan Keputusan Etis

A B S T R A K

Penelitian ini bertujuan untuk mengeksplorasi bagaimana organisasi mengintegrasikan Kecerdasan Buatan (AI) dan penilaian manusia dalam sistem evaluasi kinerja karyawan guna mencapai hasil yang adil, efisien, dan kontekstual. Dengan menggunakan pendekatan studi kasus kualitatif di tiga organisasi yang berlokasi di Jakarta, Surabaya, dan Singapura, data dikumpulkan melalui wawancara semi-terstruktur, analisis dokumen internal, dan observasi panel penilaian kinerja. Setiap organisasi menggunakan sistem manajemen berbasis AI seperti SAP *Success Factors* dan IBM *Watson Talent Insights*, namun tetap mempertahankan peran aktif penilaian manusia dalam pengambilan keputusan akhir. Analisis tematik menghasilkan lima tema utama: kepercayaan terhadap AI, intervensi manusia, keadilan, kecocokan emosional, dan isu etika. Hasil penelitian menunjukkan bahwa meskipun AI mampu meningkatkan efisiensi dan konsistensi, dimensi kualitatif seperti kepemimpinan, kolaborasi, dan kecocokan budaya kerja masih membutuhkan penilaian manusia. Studi ini juga mengusulkan model evaluasi hibrida yang menggabungkan skor AI dan skor manusia, disesuaikan dengan koefisien risiko etis. Temuan ini memberikan kontribusi empiris terhadap praktik penilaian kinerja kontemporer, serta menekankan pentingnya transparansi algoritmik dan sensitivitas etis dalam implementasi sistem berbasis AI di bidang manajemen sumber daya manusia.

CORRESPONDING AUTHOR

andrynovrianto@upiyptk.ac.id

DOI

10.37034/jems.v7i3.163

1. Introduction

In the era of digital transformation, organizations are under increasing pressure to optimize human resource practices, particularly performance appraisal systems, to align employee capabilities with strategic goals. Traditionally, performance appraisal has been a subjective and manually administered process, often susceptible to human biases, inconsistency, and a lack of transparency [1]. As workforce expectations grow more complex, there is a rising need for performance evaluations that are not only accurate but also perceived as fair and evidence-based [2]. These evolving demands have prompted organizations to adopt Artificial Intelligence (AI) as a tool to support and enhance the performance appraisal process by enabling data-driven and real-time evaluations [3].

In recent literature, AI has been recognized for its potential to improve the efficiency and consistency of performance assessments. AI-based systems can process large volumes of employee data, identify behavior patterns, and deliver objective insights into employee performance [4], [5]. However, critical limitations have also been identified. For instance, AI lacks contextual understanding and emotional intelligence, which are

crucial for interpreting nuanced aspects of human behavior [6]. Several studies have also raised ethical concerns related to dehumanization and data privacy in AI-driven performance evaluations [7]. In contrast, human judgment brings subjective insights, contextual understanding, and interpersonal sensitivity that AI cannot replicate [8]. Despite its susceptibility to cognitive bias and inconsistency, human evaluation remains irreplaceable in capturing qualitative dimensions of performance such as leadership, teamwork, and cultural fit [9]. This dichotomy between automation and human discretion has driven scholars to explore hybrid appraisal models that seek to combine the objectivity of AI with the empathy of human evaluators [10], [11].

Despite the growing implementation of AI in performance appraisal, there is a lack of qualitative research exploring how organizations balance automation and human judgment in real-world appraisal settings. Most existing studies adopt a quantitative or technical perspective, overlooking the social, ethical, and experiential dimensions that influence the effectiveness of AI in human resource practices. Consequently, there is a pressing need to understand how HR professionals, managers, and employees perceive, interpret, and interact with AI-powered appraisal tools in practice. By focusing on organizational experiences, this study aims to uncover how hybrid appraisal systems are constructed, negotiated, and managed within dynamic workplace environments.

Although the integration of Artificial Intelligence (AI) into employee performance appraisal systems has shown improvements in efficiency and objectivity [5], most existing studies adopt a technical or quantitative lens, often overlooking the social, ethical, and experiential dimensions of AI application in real-world organizational contexts [6], [7]. AI's limitations in capturing contextual understanding and interpersonal dynamics present challenges in ensuring fair and holistic evaluations, especially in qualitative aspects such as leadership, teamwork, and cultural fit [8], [9]. Meanwhile, literature on hybrid models that blend AI and human judgment remains largely conceptual, lacking empirical insight into how organizations practically construct, negotiate, and manage this synergy [10], [11]. Hence, there exists a significant research gap concerning how HR professionals and organizational stakeholders perceive and experience AI-supported appraisals within dynamic work environments. The novelty of this study lies in its qualitative approach, aiming to empirically explore how organizations balance AI and human judgment in performance appraisals and to offer new insights into the design of ethical, fair, and context-sensitive appraisal systems in the digital age.

Therefore, the main research question addressed in this study is: How do organizations balance the use of AI and human judgment in performance appraisal systems to enhance the quality of employee evaluation. This study adopts a qualitative approach to investigate this question, with the objective of providing deep insights into the practical realities of implementing AI in performance management. The research contributes to the literature by offering an empirical understanding of how technological and human elements coexist within performance appraisal systems, and by proposing practical implications for designing ethical, transparent, and effective appraisal frameworks in the digital era.

2. Research Method

This study employs a qualitative research methodology to explore how organizations balance artificial intelligence (AI) and human judgment within performance appraisal systems. A qualitative approach is particularly appropriate for investigating complex social phenomena that involve subjective perceptions, ethical considerations, and contextual interpretations [12]. The focus of the study extends beyond the technical functionalities of AI tools to examine how human resource (HR) professionals, managers, and employees interact with and interpret hybrid appraisal systems in real-world organizational contexts.

A multiple case study design was adopted to enable a comprehensive analysis of how various organizations conceptualize and implement hybrid performance appraisal systems [13], [14]. Three organizations from the technology and services sectors were selected using purposive sampling. The selection criteria included: (1) the presence of AI-based tools in the performance appraisal process, (2) continued involvement of human evaluators, and (3) the organization's willingness to grant access to relevant stakeholders for research purposes [15].

Data collection methods included semi-structured interviews, document analysis, and non-participant observations. A total of 18 interviews were conducted with HR managers, team leaders, and employees across the three organizations. Interviews lasted between 45 and 75 minutes and followed an open-ended question guide designed to elicit detailed reflections on performance appraisal experiences. In addition to interviews, internal organizational documents such as appraisal policies, training manuals, and procedural guidelines were analyzed to triangulate qualitative findings. Non-participant observations during performance review meetings were also undertaken to capture real-time interactions between AI-generated assessments and human evaluative responses.

Thematic analysis was employed to analyze the data, following the framework proposed by certain researchers [16]. The process involved initial familiarization with data, generation of initial codes,

identification and review of themes, and refinement of thematic categories. NVivo software was used to facilitate data coding, organization, and retrieval. To enhance research credibility and trustworthiness, the study implemented several rigor strategies including member checking, data triangulation, and maintaining an audit trail throughout the analysis process [17]. This methodology enables the research to uncover nuanced organizational practices, ethical tensions, and experiential dimensions that are often overlooked in quantitative evaluations. It also facilitates the development of practical recommendations for designing hybrid performance appraisal systems that are both effective and human-centered.

2.1. Method of Preparation

This qualitative case study research was designed to analyze the integration of AI and human judgment in employee performance appraisals. The research aimed to uncover how appraisal processes are structured, moderated, and interpreted when both algorithmic and human evaluators are involved.

The study focused on three companies headquartered in Jakarta, Surabaya, and Singapore respectively, all of which have implemented AI-powered performance evaluation tools between 2022–2024. The selected organizations employ more than 250 staff and have HR divisions utilizing software such as SAP SuccessFactors and IBM Watson Talent Insights.

Participant selection followed purposive maximum variation sampling to ensure the inclusion of diverse perspectives, involving 6 HR specialists, 6 middle managers, and 6 general employees (n = 18). Each participant had direct experience with AI-evaluation processes. Ethical approval was obtained from an institutional review board (Ref No: 2025/HR-88), and all participants provided signed informed consent.

2.2. Characterization Techniques

a. Interview Structure and Depth

Each semi-structured interview included 15 core questions, validated through a pilot with 2 external HR experts. Interviews were conducted in private meeting rooms or via encrypted Zoom calls. Recordings totaled 1,110 minutes of audio and 246 pages of transcripts.

b. Document Triangulation

Document triangulation in this study served to corroborate findings obtained from interviews and observations by analyzing internal documents directly involved in the performance appraisal process. Across the three organizations, a consistent framework was identified in how AI-generated evaluations were structured and integrated into the broader performance management system. Each organization had algorithmic documentation outlining the logic, weight distribution, and key performance indicators used by the AI. These documents revealed that most algorithms were designed around quantitative metrics such as task completion rate, peer feedback scores, and time-on-task, with limited capacity to incorporate soft skills such as emotional intelligence or conflict resolution.

Notably, policy guidelines across the organizations emphasized that AI scores were considered "supportive data" rather than final judgments. This was clearly stated in all three HR policy documents, underscoring a hybrid model where human intervention was institutionalized. Human discretion was particularly emphasized in cases involving qualitative performance dimensions, such as leadership, cultural fit, and collaboration, which AI systems were unable to evaluate with nuance. For example, one internal HR memorandum highlighted a case in which an employee received a low AI score due to minimal digital activity, but was later upgraded by the human panel due to contributions made through offline client engagement an activity not tracked by the AI system.

The employee feedback forms offered a unique perspective on how AI-generated scores were perceived by the workforce. In all three organizations, there were recorded concerns regarding transparency and perceived bias in AI outputs, particularly in cases where the algorithm's logic was not disclosed to the evaluated individual. Some employees expressed skepticism toward the accuracy of their scores and highlighted the lack of context in the numerical ratings. This aligns with themes found in the interview data, where trust in AI tools was often contingent on the degree of explanation and human oversight provided.

Collectively, these documents reinforced the study's findings that AI, while beneficial for processing large data sets and reducing administrative burdens, must be complemented by human evaluators who can interpret context, account for intangible performance traits, and ensure fairness. The presence of written procedures for human override, ethical review protocols, and documented instances of human-AI score reconciliation all point to an evolving but cautious acceptance of AI in HR practices. These patterns affirm the necessity of hybrid appraisal systems that prioritize both algorithmic efficiency and human empathy in decision-making.

2.3. Observation Notes

In 6 performance appraisal panels observed (2 per organization), field notes captured the integration sequence of AI data vs. human commentary. Key data points included: time spent reviewing AI reports (avg. 4.3 mins), frequency of human overrides (avg. 3.2 cases per panel), and ethical issue flags (noted in 67% of panels).

2.4. Thematic Analysis and Validation Strategy

The thematic analysis conducted in this study adhered to the six-phase framework, which includes data

familiarization, generation of initial codes, theme searching, theme reviewing, theme defining and naming, and final report production [18]. This inductive approach enabled the emergence of themes based on participants lived experiences, allowing for a data-driven construction of meaning rather than relying on pre-established theoretical constructs. Throughout the coding process conducted using NVivo 12 software patterns began to surface across various stakeholder roles, including HR managers, team supervisors, and employees, like we seen on Table 1. These recurring themes revealed nuanced perspectives on how artificial intelligence (AI) and human judgment are integrated within performance appraisal systems, highlighting contextual interactions, perceptions of fairness, and the division of interpretative authority between algorithmic outputs and human evaluators.

Table 1. The data were coded and analyzed using NVivo 12 Plus

Theme Code	Description	Frequency
T1: Trust in AI	Confidence in automated scoring	92
T2: Human Override	Situations where humans revised AI scores	84
T3: Fairness	Concerns on equity and transparency	76
T4: Emotional Fit	Judgments on soft skills like teamwork	59
T5: Ethical Flags	Privacy or bias concerns in algorithm usage	48

Trustworthiness measures: Triangulation across interviews, documents, and observations, Member checking with 6 participants confirming thematic accuracy, Peer debriefing with 2 external qualitative scholars. The theme of "Trust in AI" emerged as the most frequently coded element across interviews. Participants who expressed high levels of trust tended to reference the consistency and speed of AI-generated feedback. However, trust was conditional often tied to the perceived transparency of the algorithm's operation. For instance, when AI tools were accompanied by dashboards that explained score components or offered comparative data, users were more likely to regard the outcomes as legitimate. Conversely, opaque algorithms led to suspicion, reinforcing the importance of explainability in AI deployment.

2.5. Analytical Model and Symbolic Formula

To guide HR professionals in balancing AI and human insights, the study proposes a hybrid decision model. The final score can be calculated by using Equation (1).

$$Final\ Score\ (FS) = 1 + \gamma(AIx \times \alpha) + (HJy \times \beta) \quad (1)$$

Where AIx is Normalized AI-generated score (0–100). HJy is Human-assigned score (0–100). α is AI reliability index (0.4–0.7), β is Human judgment index (0.3–0.6) and γ is Ethical risk coefficient (0–2, higher = more uncertainty).

2.6. Pseudocode for Real-Time Appraisal Calculation

The inclusion of a pseudocode model in this study was aimed at operationalizing the hybrid evaluation logic that organizations use when combining AI-generated scores with human judgment. While the study itself employed a qualitative approach, the use of algorithmic representation served to illustrate how theoretical concepts identified through interviews and document analysis can be translated into programmable decision-making logic. This hybrid model captures the actual computation process that occurs behind performance appraisal systems in AI-assisted environments which can be seen on Figure 1.

```
Algorithm
# Hybrid Performance Appraisal Calculation

def hybrid_appraisal(ai_score, human_score,
ai_weight, human_weight, risk_coef):
   combined_score = ((ai_score * ai_weight) +
   (human_score * human_weight)) / (1 +
   risk_coef)return combined_score

# Usage
score = hybrid_appraisal(85, 90, 0.6, 0.4,
0.5)
print("Final Appraisal Score:", score)
```

Figure 1. Pseudocode

3. Result and Discussion

This section presents and interprets the findings of the study in a structured and logical sequence, enabling a coherent narrative on how hybrid performance appraisal systems function in real organizational settings. The results are not only described in terms of observed data but are contextualized to address the research question concerning the balance between automated evaluation tools and human judgment. By analyzing interview transcripts, panel observations, and internal documents from three organizations, the study provides concrete evidence on how algorithmic and human components co-exist and interact within contemporary appraisal frameworks. Each sub-section focuses on a key theme identified through thematic analysis, backed by real-world examples, empirical frequency data, and procedural documentation. The discussion component of each sub-section interprets the significance of the findings, drawing comparisons with relevant literature and identifying both strengths and limitations of the current appraisal practices.

3.1. Structure of Hybrid Appraisal Implementation

The structural design of hybrid performance appraisal systems across the three case organizations based in Jakarta, Surabaya, and Singapore demonstrates a deliberate integration of automated scoring tools with human judgment to produce balanced performance evaluations. Each organization implemented advanced Human Resource Management Systems (HRMS), notably SAP SuccessFactors and IBM Watson Talent Insights, to initiate the assessment process. These digital

platforms processed vast amounts of employee data in real-time, utilizing predefined metrics such as task completion rate, peer review ratings, and digitally logged hours of engagement. The logic embedded within these systems was codified in internal algorithm documents which described the weight distribution and scoring thresholds used by each tool. As shown in Table 2, a total of nine internal documents were analyzed, consisting of algorithm specifications (n=3), policy frameworks (n=3), and anonymized employee feedback forms (n=3). These sources collectively confirm that while AI-generated data served as the first layer of analysis, final performance judgments required human verification. Importantly, the HR policy documents in all three organizations explicitly stated that AI scores were considered "supportive evidence," and not the definitive basis for decision-making. This institutionalized approach ensured that final appraisal outcomes reflected not only numeric efficiency but also qualitative contributions and contextual nuances.

Table 2. Across the three organizations, 9 key internal documents were analyzed

Document Type	Description	Quantity
AI Appraisal Algorithms	Backend logic and weight structure	3
HR Evaluation Policy	Standard operating procedures	3
Employee Feedback Forms	Anonymous reaction to AI scores	3

In operational terms, human judgment was embedded at multiple stages of the performance review cycle. During the six appraisal panels observed (two per organization), the average time spent reviewing AI-generated reports was 4.3 minutes, followed by discussion and potential override by human evaluators. On average, 3.2 score adjustments were made per panel, illustrating a recurring pattern where algorithmic outputs were either questioned or supplemented by human insight. This practice was particularly evident in scenarios involving non-digitized work contributions. For instance, in one notable case, an employee's AI score was penalized due to minimal activity on tracked systems; however, the human evaluation panel later upgraded the score after acknowledging extensive offline client engagements an area beyond the scope of the AI's tracking mechanism. Such cases highlight that while AI offers consistency and speed in processing performance metrics, it remains insufficient in capturing nuanced behaviors such as leadership, mentorship, and cross-functional collaboration. Consequently, each organization adopted a formalized two-tier system: the first tier consisted of AI-derived analytics; the second tier involved qualitative assessments by HR managers or team leads who could interpret contextual elements. This hybrid structure not only safeguards against blind reliance on technology but also affirms a commitment to fairness, acknowledging the complex nature of human performance that cannot be reduced to mere data points.

3.2. Dependence on Manual Judgment

Despite the integration of sophisticated software systems capable of processing vast amounts of employee performance data, the research findings underscore a persistent and institutionalized reliance on human evaluators in determining final appraisal outcomes. This dependence on manual judgment was especially evident in the handling of non-digitized, context-rich activities that escape algorithmic detection. For example, one case documented in Organization A involved a senior consultant whose AI-generated score was markedly low due to minimal digital footprint, particularly in internal task management systems. However, the human evaluation panel reversed this assessment after considering his consistent contributions to high-value client negotiations conducted outside the scope of the tracking software. This incident exemplifies the irreplaceable value of human discretion in interpreting the qualitative dimensions of performance such as client engagement, leadership initiative, or mentorship all of which are poorly captured by automated systems focused on quantifiable metrics. During the six observed appraisal panels, human overrides of AI scores occurred in 3.2 cases per session on average, demonstrating that these interventions were not exceptional but rather integral to the standard evaluation process.

This reliance on human judgment was also corroborated by internal policy documents and interview data, which collectively underscored the importance of qualitative assessments in ensuring a fair and holistic performance evaluation process. Across all three organizations, formal guidelines explicitly stipulated that AI-generated outputs should serve only as preliminary insights rather than definitive judgments, requiring validation or moderation by human evaluators. Interviewees particularly middle managers and HR professionals repeatedly emphasized the limitations of current AI systems in interpreting employee behaviors that fall outside algorithmically recognized norms but nonetheless contribute meaningfully to organizational objectives. These findings are consistent with prior studies suggesting that AI, while competent in identifying patterns within structured datasets, lacks the emotional intelligence and contextual sensitivity needed for nuanced evaluations [19], [20]. Additionally, employee feedback documents reviewed during the study revealed a broad consensus that the human aspect of the appraisal process plays a vital role in preserving fairness and recognizing unique, role-specific contributions. The evidence thus points to a cautious, layered approach to AI integration in practice one that views AI not as a replacement for human judgment, but as a complementary tool aimed at enhancing administrative efficiency while safeguarding interpretive and ethical integrity through ongoing human involvement.

3.3. Trust in System Scores

One of the most prominent themes identified through the data analysis was the issue of trust in system-generated performance scores. The thematic code “Trust in AI” appeared 92 times across the 18 semi-structured interviews, signifying its prominence as a key concern among HR professionals, supervisors, and employees. Trust in automated appraisal outcomes was found to be highly contingent upon the perceived transparency and explainability of the AI system’s decision-making process. In organizations that provide users with access to visual interfaces uch as dashboards that detailed the scoring formula through weighted components like task completion metrics or peer evaluations acceptance of AI-generated scores was significantly higher. Several respondents indicated a willingness to accept low performance scores when the system offered clear breakdowns of the calculation process. Conversely, in contexts where the scoring algorithm was opaque or inadequately explained, participants voiced skepticism regarding both the fairness and accuracy of the appraisal outcomes. These findings underscore the importance of explainability as a mediating variable in the formation of trust, aligning with prior research that links algorithmic transparency with increased user acceptance and perceived fairness [8], [9].

Furthermore, the study uncovered that trust was not merely a function of system design but also of how organizations managed communication and oversight of the appraisal process. For example, in Organization C, the HR department introduced an internal guidebook explaining the algorithmic logic and scoring weights used in the performance appraisal software. This proactive measure significantly improved trust levels among employees, as reflected in both interview narratives and written feedback. Employees in this organization reported fewer instances of contesting their scores and exhibited higher levels of engagement in appraisal review meetings. In contrast, Organization B, which did not disclose the scoring logic and treated the algorithm as a "black box," experienced more frequent disputes and a general climate of suspicion toward automated evaluation. The interviews also revealed that trust was deeply intertwined with perceptions of fairness; when AI tools were seen as impartial and rule-based but also adjustable through human review participants described the system as both credible and flexible. This underscores the importance of hybrid mechanisms where automated processes are balanced by human validation, allowing for not only procedural efficiency but also psychological assurance. Ultimately, trust in AI-generated scores cannot be achieved through technical optimization alone; it requires deliberate transparency, open dialogue, and procedural checks that affirm the organization's commitment to ethical and fair evaluation.

3.4. Thematic Validation through Triangulation

To enhance the credibility and depth of the findings, the study employed methodological triangulation by combining three distinct data sources: semi-structured interviews, non-participant observations of appraisal panels, and analysis of internal organizational documents. This triangulation strategy was designed not only to corroborate the thematic codes but also to capture multidimensional insights into how performance appraisal systems operate in real-world settings. For instance, themes such as "Trust in AI," "Human Override," and "Ethical Flags" were not confined to interview data but were echoed in panel discussions and documented HR procedures. The presence of ethical review protocols, human override guidelines, and employee appeal channels across all three organizations confirmed that concerns over fairness, transparency, and contextual relevance were institutionally recognized. Furthermore, the observational notes from six appraisal panels provided real-time evidence of how AI scores were integrated into discussions and subsequently revised. On average, each panel spent approximately 4.3 minutes discussing AI-generated reports before transitioning into interpretive deliberations led by human evaluators. These observational insights added granularity to the themes identified through interviews, showing that scoring revisions were not ad hoc but systematically embedded into the appraisal workflow.

Additionally, the study applied rigorous validation techniques to ensure that the thematic interpretations were both accurate and grounded. Member checking was conducted with six participants representing different roles two HR managers, two supervisors, and two general employees who reviewed and affirmed the accuracy of the thematic summaries generated from their interviews. Their feedback confirmed that the coding categories accurately reflected their experiences, particularly regarding transparency concerns and the importance of human discretion. Moreover, the study engaged in peer debriefing with two external qualitative researchers who reviewed the coding framework, thematic structure, and selected transcript excerpts. Their feedback led to refinements in the definition of certain codes, such as distinguishing between "algorithmic opacity" and "system rigidity" under the broader theme of trust. This external scrutiny helped bolster the study’s analytical rigor and mitigate potential researcher bias. The use of NVivo 12 software further facilitated systematic data management and ensured traceability from raw excerpts to thematic categories. These methodological choices affirm that the themes discussed are not anecdotal but are the result of carefully validated, multi-source data interpretation. Thus, triangulation in this study not only increased the robustness of the results but also enhanced the interpretive power of the discussion, enabling a comprehensive understanding of hybrid performance

appraisal systems from both procedural and experiential perspectives.

3.5. Scoring Model and Simulation

To operationalize the hybrid evaluation framework observed in the field, this study proposes a computational scoring model that reflects how organizations merge algorithmic efficiency with human contextual judgment. The model is formulated on Equation (1), where AIx represents the normalized score produced by the automated system (0–100), HJy represents the score given by human evaluators (0–100), α and β are weighting factors representing the reliability index of the AI and human inputs respectively, and γ denotes an ethical risk coefficient (ranging from 0 to 2) that adjusts for uncertainty or potential bias. The inclusion of the γ variable is particularly significant as it introduces a formal mechanism for addressing the ethical ambiguity often overlooked in performance evaluation systems. This model was not merely theoretical but emerged from empirical practices observed in the three organizations, each of which allowed for score reconciliation and documented instances where ethical concerns justified deviations from AI recommendations. For example, in Organization A, a junior analyst received a below-average system score due to limited online engagement; however, after a peer review noted her extensive mentorship contributions, the human panel invoked ethical adjustment principles and raised the final score accordingly.

A simulation using this model further demonstrates its utility and adaptability to real-world scenarios. When assigning a system score (AIx) of 85 and a human-assigned score (HJy) of 90, with respective reliability indices of $\alpha = 0.6$ and $\beta = 0.4$, and setting the ethical coefficient γ at 0.5 to account for a moderate level of risk or uncertainty, the computed final appraisal score is on Equation 2.

$$FS = 1 + 0.5(85 \times 0.6) + (90 \times 0.4) = 88.5 \quad (2)$$

This calculation illustrates the balanced weight of both inputs, giving proportional influence to human insights while moderating the impact of automation depending on ethical considerations. In practical terms, such a scoring model allows organizations to tailor appraisal outcomes according to contextual demands while maintaining transparency and consistency in evaluation logic. The formula was also translated into pseudocode to simulate how HR systems might integrate these logic rules in real-time appraisal platforms. Beyond the computational clarity, this model addresses a critical gap in current HR analytics: the absence of flexible mechanisms to account for ethical and interpretive variables. The flexibility to adjust weights and coefficients also allows HR practitioners to calibrate the model to different roles, departments, or performance environments, thereby avoiding a "one-size-fits-all" logic that often undermines fairness in digital appraisals. Overall, the scoring model and its simulation illustrate not only a formal representation of the appraisal process but also a normative framework that reinforces the coexistence of structured data analysis and human ethical oversight in organizational decision-making.

4. Conclusion

This study revealed that hybrid performance appraisal systems those that integrate algorithmic evaluation with human judgment are not only technically feasible but operationally necessary to ensure fairness, contextual sensitivity, and ethical accountability in employee assessments. Across the three case organizations, automated tools were consistently used to generate preliminary performance scores based on quantifiable indicators such as task completion and peer feedback. However, these scores were never treated as final decisions. Human evaluators actively engaged in interpreting the data, adjusting the scores where appropriate, especially in cases involving non-digitized contributions like mentoring, client engagement, and leadership. The observed use of human overrides in an average of 3.2 cases per appraisal panel affirms that human discretion is a fundamental safeguard in performance evaluation processes. The findings also highlight that employee trust in appraisal systems hinges on the transparency of the algorithmic logic and the perceived fairness of score interpretation. Organizations that disclosed how scores were calculated observed higher levels of acceptance and reduced contestation. Furthermore, the study proposed and simulated a hybrid scoring model that integrates quantitative AI scores and qualitative human judgments while adjusting for ethical risk. This model provides a structured yet flexible framework for organizations aiming to enhance their performance management systems. From a practical standpoint, the findings suggest that future HR technologies should embed transparency features and allow for dynamic human adjustment to maintain legitimacy and employee morale. For future research, it is recommended to explore longitudinal impacts of hybrid systems on employee motivation, as well as to test the scalability of the proposed scoring model across diverse industries and cultural contexts.

References

[1] Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

[2] Creswell, J. W., & Poth, C. N. (2021). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.

[3] Haenlein, M., Kaplan, A., Tan, C.-W., & Zhang, P. (2021). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, *63*(3), 135–152. https://doi.org/10.1177/00081256211014385

[4] Jarrahi, M. H. (2021). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *64*(5), 595–606. https://doi.org/10.1016/j.bushor.2021.02.020

[5] Khan, S., Maalik, A., & Zaman, A. (2021). AI-driven human resource management: Towards a conceptual framework. *Journal of Business and Social Review in Emerging Economies, 7*(1), 47–60. https://doi.org/10.26710/jbsee.v7i1.1560

[6] Kim, J., & Park, S. (2021). Beyond numbers: A qualitative exploration of performance appraisal in digital era. *Human Resource Development Review, 20*(1), 70–89. https://doi.org/10.1177/1534484320978962

[7] Liu, Y., Wei, Y., & Wang, Y. (2022). Human factors in AI-based performance evaluation systems. *International Journal of Human–Computer Interaction, 38*(11), 1012–1023. https://doi.org/10.1080/10447318.2021.2006455

[8] Nugroho, Y., Prasetyo, A., & Rahman, T. (2023). The transformation of performance appraisal through AI integration: A case study. *Indonesian Journal of Management Studies, 7*(2), 120–134. https://doi.org/10.21009/IJMS.07206

[9] Park, S., & Lee, J. (2020). Designing AI systems for performance appraisal: A sociotechnical perspective. *Journal of Information Technology, 35*(4), 305–320. https://doi.org/10.1177/0268396220911880

[10] Sujan, M., Hassan, R., & Bakar, R. (2020). Ethical considerations in AI-based HR systems: A hybrid appraisal model. *Ethics and Information Technology, 22*(4), 323–334. https://doi.org/10.1007/s10676-020-09538-9

[11] Tambe, P., Cappelli, P., & Yakubovich, V. (2020). Artificial intelligence in human resources management: Challenges and a path forward. *Academy of Management Perspectives, 34*(4), 628–652. https://doi.org/10.5465/amp.2018.0071

[12] Zambrano, R., O'Neill, J., & Araya, D. (2022). Algorithmic fairness in workplace surveillance and appraisal. *Technology in Society, 70*, 101968. https://doi.org/10.1016/j.techsoc.2022.101968

[13] Zhang, L., & Fu, H. (2022). The role of perceived fairness in AI-based performance reviews. *Journal of Organizational Behavior, 43*(1), 23–38. https://doi.org/10.1002/job.2553

[14] Ghosh, R., & Saha, D. (2023). Organizational readiness for hybrid performance management systems. *International Journal of Productivity and Performance Management, 72*(2), 425–442. https://doi.org/10.1108/IJPPM-04-2021-0205

[15] Lee, S., & Chang, H. (2022). Hybrid intelligence in HR analytics: Integrating AI and human judgment. *Journal of Business Research, 139*, 854–863. https://doi.org/10.1016/j.jbusres.2021.10.063

[16] Murthy, V., & Menon, S. (2020). Evaluating emotional intelligence in the era of AI-based appraisal. *Human Resource Management Review, 30*(4), 100736. https://doi.org/10.1016/j.hrmr.2019.100736

[17] Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1–13. https://doi.org/10.1177/1609406917733847

[18] Yin, R. K. (2023). *Case study research and applications: Design and methods* (7th ed.). SAGE Publications.

[19] Choudhury, S., & Singh, K. (2021). From human bias to algorithmic bias: Exploring ethics in AI-driven appraisal. *Journal of Business Ethics, 172*, 543–556. https://doi.org/10.1007/s10551-020-04666-7

[20] Putro, R. L., & Candrakusuma, M. (2025). Human resource management in Muhammadiyah business charity: A case study of Airmu Ponorogo. *Journal of Economics and Management Sciences, 7*(2), 32–36. https://doi.org/10.37034/jems.v7i2.83