Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Co-Agency in Organizations: Human–Agent Collaboration as a New Management Paradigm

[1]Azmat Islam -Email- azmat24@gmail.com
[*2]Muhammad Ajmal -Email- ajmal.hailian@gmail.com
[1]Department of Business Administration, University of Education, Lahore
*[*2]Department of Management Science, University of Gujrat, Gujrat, Pakistan*

Article Details:

Received on 20 Nov, 2025
Accepted on 25 Dec ,2025
Published on 27 Dec, 2025

Corresponding Authors*:
Muhammad Ajmal

Abstract

The rapid integration of artificial intelligence (AI) into organizational processes is transforming management from a paradigm of tool utilization to one of human–agent co-agency. Rather than functioning solely as automation technologies, intelligent systems increasingly act as collaborative partners that contribute to analysis, coordination, and decision-making. This shift requires organizations to move beyond performance models that optimize AI in isolation and instead design systems that enhance joint human–AI outcomes. Emerging evidence suggests that complementary role allocation, explainability, trust calibration, and adaptive governance structures are central to effective collaboration. This paper introduces co-agency as a new management paradigm in which humans and intelligent agents share decision authority, dynamically distribute tasks based on strengths, and co-create organizational value. We develop a conceptual framework outlining the structural, cognitive, and ethical foundations of co-agency, and examine its implications for leadership, performance management, and organizational design. By reframing AI as a collaborative actor rather than a passive instrument, co-agency provides a foundation for resilient, adaptive, and high-performing organizations in the age of intelligent systems.

Keywords: Co-agency; Human–AI collaboration; Human–agent teams; Artificial intelligence in management; Organizational decision-making; Trust in AI; Explainable AI; Team effectiveness; Complementary intelligence; Digital transformation.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

1. Introduction

Artificial intelligence (AI) is rapidly transforming the foundations of organizational decision-making and management practice. What was once framed primarily as automation is increasingly understood as collaborative augmentation, in which intelligent systems actively participate in analysis, coordination, and problem-solving alongside human actors. Contemporary organizations deploy AI in areas ranging from recruitment and performance evaluation to strategic planning and operational optimization(Ajmal & Suleman, 2015a). In these contexts, AI no longer functions merely as a passive computational tool; instead, it acts as a decision partner whose outputs shape managerial judgment and team dynamics (Dican, 2025; Tummala et al., 2025). This shift calls for a reconceptualization of management paradigms toward what can be described as *co-agency*—a model in which humans and artificial agents jointly contribute to organizational action (Ajmal & Suleman, 2015b).

Emerging research demonstrates that AI-supported decision-making can enhance efficiency, reduce cognitive bias, and strengthen evidence-based management when appropriately integrated (Dican, 2025). At the team level, AI agents are increasingly conceptualized as "team members" capable of influencing shared mental models, communication patterns, and coordination structures (Tummala et al., 2025; Lou et al., 2025). Zercher et al. (2025) show experimentally that teams collaborating with AI possessing centralized knowledge achieve higher decision accuracy than human-only teams, particularly when collaboration processes are structured effectively(Ajmal, Islam, & Islam, 2024b). These findings suggest that AI can improve collective intelligence—not simply by automating tasks, but by reshaping how information asymmetries are resolved within teams.

However, maximizing AI's independent predictive accuracy does not necessarily produce optimal organizational outcomes. Research on AI-advised decision-making demonstrates that models optimized for *team performance* rather than standalone accuracy yield superior joint utility (Bansal et al., 2020; Gao et al., 2023). Complementarity between human and AI capabilities—where humans contribute contextual reasoning, ethical discernment, and creativity, while AI provides large-scale pattern recognition and computational consistency—emerges as a central mechanism for performance gains (Leitão et al., 2022; Gao et al., 2023). This insight challenges traditional management approaches that prioritize efficiency through substitution. Instead, it highlights the importance of structural interdependence between human judgment and algorithmic analysis(Ajmal, Islam, & Khalid, 2025a).

Trust constitutes a critical mediating variable in this evolving collaboration. Managers assign decision weight to AI systems based on their level of trust, which in turn influences reliance, delegation, and oversight (Wen et al., 2025). Empirical evidence indicates that communication breakdowns or perceived opacity in AI systems can significantly reduce trust and impair team performance (Chen & Zhang, 2025). Trust is dynamic and requires calibration mechanisms that continuously monitor transparency, reliability, and behavioral consistency (Hou et al., 2024). Experimental work in human–robot collaboration further shows that calibrated trust—not blind trust—maximizes long-term team performance (Chen et al., 2018). Thus, effective co-agency depends on governance structures that balance algorithmic autonomy with human accountability.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Explainability and interpretability further strengthen human–agent collaboration by enabling users to understand and appropriately evaluate AI recommendations. Explainable AI frameworks enhance transparency and facilitate responsible decision-making in high-stakes domains (Okonkwo et al., 2025). Interface design and cognitive engagement mechanisms also influence how users interpret AI advice, affecting collaborative task performance and trust dynamics (Zichen Chen et al., 2025). Without such mechanisms, automation bias may lead to overreliance on AI recommendations or diminished ethical responsibility (Tolmeijer et al., 2022). Consequently, the integration of AI into organizations requires not only technological sophistication but also thoughtful design of interaction protocols and accountability systems(Ajmal, Islam, & Khalid, 2025b).

Despite growing empirical insights, management scholarship lacks an integrative theoretical framework that synthesizes complementarity, trust calibration, explainability, and adaptive governance into a unified paradigm. Existing studies often examine isolated mechanisms—such as performance optimization, trust, or interpretability—without embedding them within a broader model of organizational design (Ajmal, Islam, & Khalid, 2025c). The concept of co-agency in organizations addresses this gap by framing human and AI actors as interdependent agents who dynamically allocate authority, redistribute tasks based on comparative advantage, and co-create value under structured ethical oversight.

This article develops a comprehensive conceptual framework for co-agency as a new management paradigm. It integrates evidence from human–AI teaming research, trust theory, decision science, and organizational behavior to articulate the structural, cognitive, and ethical dimensions of human–agent collaboration (Ajmal, Islam, & Khalid, 2025d). By moving beyond automation and augmentation narratives, this framework positions AI as a collaborative actor embedded within organizational systems of accountability, leadership, and performance management. In doing so, the paper contributes to the theoretical advancement of management studies and provides practical guidance for organizations navigating the transition toward intelligent, hybrid decision architectures.

2. Literature Review

2.1. From Automation to Human–AI Collaboration

Early organizational applications of artificial intelligence were primarily framed within an automation paradigm, emphasizing efficiency gains through task substitution. However, contemporary research increasingly conceptualizes AI as a collaborative partner embedded within human decision systems rather than as a replacement technology. AI-supported decision-making enhances analytical rigor, reduces cognitive bias, and improves information processing, yet its value is maximized when integrated with human judgment rather than deployed autonomously (Dican, 2025).

Tummala, Burris-Melville, and Eskridge (2025) argue that AI systems are evolving into functional team members capable of influencing shared mental models, coordination, and joint performance. Similarly, Lou et al. (2025) emphasize that AI agents are transitioning from passive tools to adaptive collaborators, requiring new interaction protocols and responsibility allocation frameworks (Ajmal, Khalid, & Islam, 2025b). Empirical evidence from team decision-making experiments shows that human–AI teams can outperform human-only teams when collaboration processes effectively leverage AI's centralized or complementary knowledge (Zercher et al., 2025).

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

This shift reflects a broader theoretical movement toward hybrid intelligence, in which human intuition and contextual reasoning combine with algorithmic computation and scalability. However, effective integration requires organizational structures that explicitly define interdependence between human and artificial agents.

2.2. Complementarity and Performance Optimization

A central theme in the literature is that the performance of human–AI systems depends on complementarity rather than individual accuracy. Bansal et al. (2020) demonstrate that optimizing AI systems for team utility—rather than maximizing standalone predictive accuracy—produces superior collaborative outcomes (Ajmal, Khalid, & Islam, 2025c). Their findings challenge conventional machine learning evaluation metrics that prioritize model performance independent of human interaction.

Extending this logic, Gao et al. (2023) propose learning complementary policies that strategically allocate decisions between human and AI agents based on comparative strengths. Their results show that routing only a subset of decisions to human agents can significantly improve overall team performance. Leitão et al. (2022) further critique "learning-to-defer" frameworks and highlight the complexity of real-world delegation under dynamic conditions.

These studies converge on a key insight: organizational effectiveness depends not merely on AI capability, but on the architecture of decision authority and task distribution. Complementarity emerges as a structural principle for hybrid systems, supporting the argument for co-agency as a deliberate management design choice.

2.3. Trust and Trust Calibration in Human–AI Teams

Trust consistently emerges as a critical determinant of collaboration quality. Wen, Wang, and Chen (2025) show that managerial trust in AI directly influences the decisional weight assigned to algorithmic recommendations in recruitment and performance evaluation context(Ajmal, Khalid, & Islam, 2025d)s. Their findings indicate that willingness to collaborate mediates the relationship between trust and AI reliance.

Chen and Zhang (2025) demonstrate that communication misunderstandings in human–AI teams reduce trust and impair performance, particularly when AI agents omit information. These findings underscore the fragile and context-dependent nature of trust in hybrid teams. Hou et al. (2024) propose a dynamic trust management framework that continuously monitors and calibrates trust levels to maintain effective collaboration(Islam, Ajmal, & Khalid, 2025a).

Earlier foundational work by Chen et al. (2018) in human–robot interaction confirms that calibrated trust—rather than maximum trust—optimizes long-term team performance. This distinction is crucial: blind reliance can lead to automation bias, whereas insufficient trust reduces the benefits of algorithmic support.

Together, these studies highlight that trust is not merely a psychological variable but a structural governance mechanism that shapes authority allocation in organizations adopting AI.

2.4. Explainability, Interpretability, and Accountability

Explainable AI (XAI) has become central to discussions of responsible human–AI collaboration. Okonkwo et al. (2025) argue that interpretability enhances user trust and ethical accountability in high-stakes domains. Transparent explanations allow decision-makers to validate AI reasoning and integrate it with domain knowledge.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Experimental research further demonstrates that interface design influences human engagement with AI recommendations. Chen, Luo, and Sra (2025) find that explanation formats and cognitive forcing mechanisms affect reliance patterns and collaborative task performance. Without appropriate interpretability mechanisms, users may either over-rely on AI or disengage from critical oversight(Islam, Khalid, & Ajmal, 2025a).

Tolmeijer et al. (2022) explore ethical perceptions of AI versus human experts in decision-making. Their findings reveal that AI systems are often perceived as more capable but less morally trustworthy than human experts. Responsibility attribution frequently shifts toward developers or organizations rather than the AI itself. These insights underscore the importance of governance frameworks that embed accountability into hybrid decision systems(Khalid, Islam, & Ajmal, 2025a).

2.5. Ethical Governance and Organizational Implications

The integration of AI into management introduces ethical challenges related to transparency, fairness, responsibility, and workforce adaptation. Dican (2025) emphasizes that AI must augment rather than replace human judgment, highlighting the importance of governance mechanisms that preserve autonomy and ethical reasoning.

Organizational behavior research suggests that hybrid teams require new leadership models capable of coordinating human and algorithmic actors simultaneously (Tummala et al., 2025). The literature indicates that successful integration depends on training, communication protocols, role clarity, and ethical safeguards (Lou et al., 2025).

Despite robust empirical findings across domains, the literature remains fragmented. Studies often focus separately on performance optimization, trust dynamics, explainability, or governance without synthesizing these dimensions into a unified organizational framework. This gap motivates the conceptual development of **co-agency**, which integrates complementarity, calibrated trust, interpretability, and ethical governance into a cohesive management paradigm.

3. Conceptual Framework: Co-Agency in Organizations

3.1. Conceptualizing Co-Agency

Co-agency refers to a structured form of interdependent action in which humans and intelligent agents jointly participate in organizational decision-making, share authority under defined governance conditions, and co-create outcomes through complementary capabilities. Unlike automation (technology substitution) or simple augmentation (technology assistance), co-agency assumes that artificial agents possess functional agency within bounded institutional constraints—meaning they influence decisions, shape communication patterns, and alter organizational dynamics (Tummala et al., 2025; Lou et al., 2025).

The conceptual foundation of co-agency draws from research on human–AI teaming, which demonstrates that AI agents increasingly function as team members capable of participating in shared mental models and coordination structures (Tummala et al., 2025). Zercher et al. (2025) empirically show that teams collaborating with AI achieve higher decision accuracy when knowledge asymmetries are reduced through structured AI integration(Ajmal, Islam, & Islam, 2024b). These findings suggest that AI systems contribute to collective cognition rather than merely processing data.

Accordingly, co-agency is defined here as:

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

A management paradigm in which humans and artificial agents operate as interdependent decision actors, dynamically allocating authority and tasks based on complementary strengths within ethical and governance boundaries.

3.2. Core Dimensions of the Co-Agency Framework

The proposed framework integrates five interrelated dimensions: (1) Complementary Capability Alignment, (2) Dynamic Authority Allocation, (3) Trust Calibration, (4) Explainability and Cognitive Integration, and (5) Ethical Governance and Accountability.

3.2.1 Complementary Capability Alignment

Complementarity forms the structural backbone of co-agency. Research demonstrates that optimizing AI systems for team performance rather than standalone accuracy improves joint decision outcomes (Bansal et al., 2020). Gao et al. (2023) further show that strategically routing decisions between humans and AI based on comparative strengths significantly enhances collective utility (Ajmal et al., 2025).

Humans contribute contextual awareness, ethical reasoning, creativity, and adaptive judgment. AI contributes computational scalability, pattern recognition, and consistency. Leitão et al. (2022) emphasize that real-world delegation requires adaptive mechanisms beyond static "learning-to-defer" models.

In the co-agency framework, complementary alignment is operationalized through:

- Task decomposition based on cognitive complexity
- Risk-based routing of decisions
- Continuous performance feedback loops

Proposition 1: Organizational performance increases when decision tasks are allocated according to complementary human–AI strengths rather than technological substitution logic.

3.2.2 Dynamic Authority Allocation

Co-agency requires structured delegation of decision weight between humans and AI. Wen, Wang, and Chen (2025) demonstrate that trust in AI directly influences the weight managers assign to algorithmic recommendations. Their findings suggest that authority is not binary (human vs. machine) but dynamically negotiated.

Chen et al. (2018) show that trust-aware decision models in human–robot collaboration optimize long-term performance when authority adjustments respond to evolving trust levels. Similarly, Hou et al. (2024) propose dynamic trust monitoring systems that recalibrate collaboration parameters over time.

Dynamic authority allocation in co-agency involves:

- Adjustable decision thresholds
- Escalation protocols for high-risk decisions
- Continuous monitoring of AI reliability

Proposition 2: Adaptive authority allocation mechanisms moderate the relationship between AI capability and organizational performance.

3.2.3 Trust Calibration

Trust functions as the psychological and structural mediator of co-agency effectiveness. Chen and Zhang (2025) find that communication failures in human–AI teams significantly reduce trust and impair performance. Trust is therefore fragile and dependent on transparency and communication quality.

Hou et al. (2024) argue that trust must be continuously calibrated rather than maximized. Excessive trust leads to automation bias; insufficient trust reduces collaboration benefits.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Tolmeijer et al. (2022) further show that AI systems are often perceived as more capable but less morally trustworthy than human experts, complicating reliance patterns.

Within the framework, trust calibration mechanisms include:

- Reliability transparency metrics
- Confidence indicators
- Feedback mechanisms for override decisions

Proposition 3: Calibrated trust mediates the relationship between explainability and effective human–AI collaboration.

3.2.4 Explainability and Cognitive Integration

Explainability enables humans to interpret, validate, and integrate AI outputs into decision processes. Okonkwo et al. (2025) argue that interpretability enhances accountability and trust in high-stakes contexts. Chen, Luo, and Sra (2025) demonstrate that interface design and explanation formats influence cognitive engagement and collaborative performance.

Without interpretability, users may either over-rely on AI (automation bias) or disengage from oversight (Tolmeijer et al., 2022). Thus, cognitive integration requires transparent reasoning pathways that align AI outputs with human mental models.

Explainability mechanisms in co-agency include:

- Model confidence communication
- Visual explanation interfaces
- Structured decision rationales

Proposition 4: Explainable AI strengthens cognitive alignment between human and artificial agents, enhancing decision quality.

3.2.5 Ethical Governance and Accountability

Co-agency operates within institutional and ethical boundaries. Dican (2025) emphasizes that AI must augment rather than replace human ethical reasoning. Lou et al. (2025) identify responsibility allocation and value alignment as unresolved challenges in human–AI teaming.

Tolmeijer et al. (2022) demonstrate that responsibility attribution in hybrid decisions often shifts toward developers or organizations rather than AI systems. This reinforces the need for governance frameworks that clarify accountability structures.

Ethical governance within co-agency includes:

- Responsibility mapping
- Auditability and traceability
- Bias monitoring systems
- Human override mandates

Proposition 5: Ethical governance moderates the relationship between AI autonomy and organizational legitimacy.

3.3. Integrated Co-Agency Model

The conceptual model proposes the following relationships:

1. Complementary capability alignment directly enhances decision quality.
2. Dynamic authority allocation strengthens this effect by adapting to contextual risk and trust levels.
3. Trust calibration mediates collaboration effectiveness.
4. Explainability reinforces trust and cognitive integration.
5. Ethical governance ensures sustainable legitimacy and long-term organizational performance.

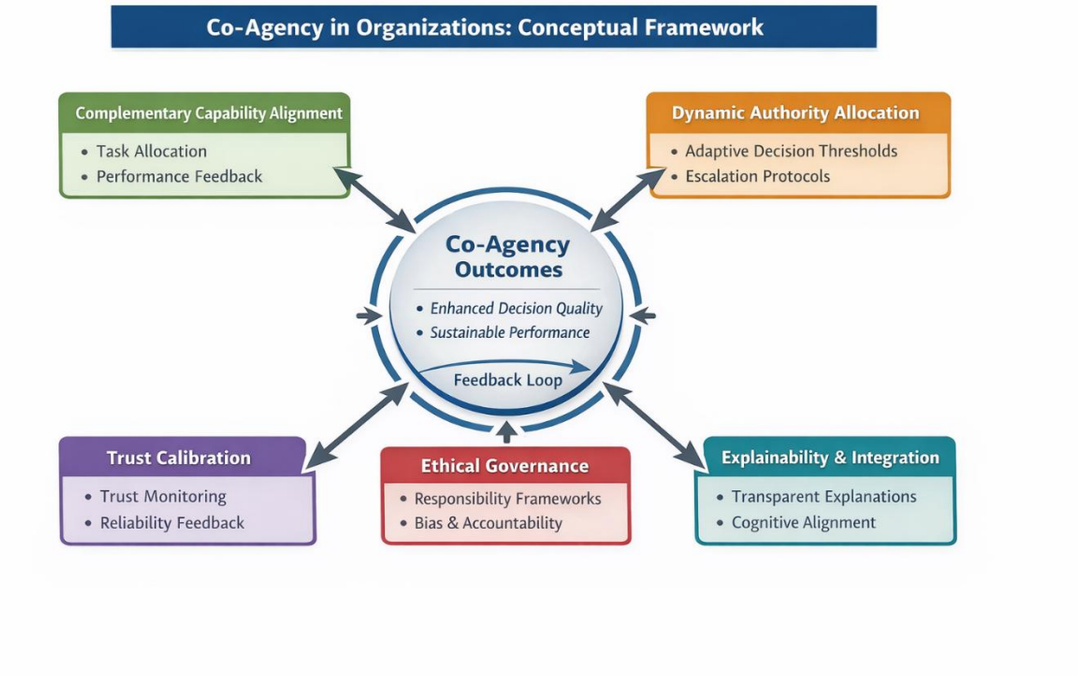
Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Together, these elements form a recursive system in which performance feedback continuously informs authority allocation and trust calibration mechanisms.



4. Explanation of the Co-Agency in Organizations Conceptual Model

The Co-Agency Model proposes that organizational performance in AI-integrated environments emerges not from technological capability alone, but from the structured interaction between human judgment and artificial agents. The framework integrates five interdependent dimensions: Complementary Capability Alignment, Dynamic Authority Allocation, Trust Calibration, Explainability & Cognitive Integration, and Ethical Governance. Together, these dimensions form a recursive system that produces enhanced decision quality and sustainable organizational performance.

4.1. Complementary Capability Alignment

At the foundation of the model lies complementarity. Research consistently shows that human–AI systems outperform either humans or AI alone when tasks are allocated according to comparative strengths.

Bansal et al. (2020) demonstrate that optimizing AI for *team utility*—rather than standalone accuracy—leads to better joint outcomes. Their findings indicate that the most accurate AI model is not necessarily the best teammate. Similarly, Gao et al. (2023) show that routing decisions strategically between humans and AI based on divergent strengths significantly improves overall performance. Leitão et al. (2022) further emphasize that static delegation mechanisms are insufficient in dynamic environments.

In this model:

- Humans contribute contextual reasoning, moral judgment, creativity, and adaptive thinking.
- AI contributes pattern recognition, computational scale, consistency, and probabilistic forecasting.

Performance increases when organizations deliberately design workflows that exploit these differences rather than replacing human tasks wholesale.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Core Mechanism: Structured task decomposition and performance-based routing.

4.2. Dynamic Authority Allocation

Co-agency requires flexible decision authority rather than fixed control structures. Authority is not binary (human vs. machine); it is negotiated and recalibrated over time.

Wen, Wang, and Chen (2025) show that managerial trust directly influences the decisional weight assigned to AI in HR contexts. This demonstrates that authority allocation is psychologically mediated. Chen et al. (2018) further illustrate in human–robot collaboration that adaptive authority models maximize long-term team performance when systems respond to evolving trust and reliability.

Hou et al. (2024) propose a trust management system that continuously monitors and adjusts collaboration parameters, reinforcing the idea that authority must remain dynamic.

In the model:

- Authority adjusts based on risk level.
- High-risk decisions trigger human oversight.
- Performance feedback modifies delegation thresholds.

Core Mechanism: Adaptive escalation and delegation protocols.

4.3. Trust Calibration

Trust functions as the central mediating variable in co-agency effectiveness. It determines whether human decision-makers rely on AI recommendations appropriately.

Chen and Zhang (2025) show that misunderstandings caused by AI significantly reduce team trust and impair performance. Importantly, Hou et al. (2024) argue that *calibrated trust*—not maximum trust—is optimal. Excessive trust leads to automation bias, while insufficient trust prevents leveraging AI capabilities.

Tolmeijer et al. (2022) demonstrate that AI systems are perceived as more capable but less morally trustworthy than human experts. Responsibility attribution often shifts to organizations or developers, reinforcing the need for structured trust governance.

In this framework:

- Trust is continuously monitored.
- Confidence indicators and reliability metrics are communicated.
- Override opportunities preserve human agency.

Core Mechanism: Continuous trust monitoring and recalibration.

4.4. Explainability and Cognitive Integration

Explainability supports trust and cognitive alignment between humans and AI.

Okonkwo et al. (2025) argue that interpretability enhances accountability and strengthens decision legitimacy in high-stakes contexts. Chen, Luo, and Sra (2025) show that interface design and explanation formats directly influence collaborative performance and reliance patterns.

Without explainability, automation bias may emerge, leading to over-reliance or disengagement (Tolmeijer et al., 2022).

In the model:

- AI systems communicate reasoning pathways.
- Confidence levels are made visible.
- Human mental models are aligned with AI outputs.

Core Mechanism: Transparent reasoning pathways enabling validation and override.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

4.5. Ethical Governance and Accountability

Ethical governance acts as the institutional boundary of co-agency. Dican (2025) emphasizes that AI must augment rather than replace human ethical reasoning. Lou et al. (2025) identify value alignment and responsibility allocation as core challenges in human–AI teaming.

The governance dimension ensures:

- Responsibility mapping
- Bias auditing
- Traceability of algorithmic decisions
- Human override mandates

This safeguards organizational legitimacy and ensures sustainable integration of AI systems.

Core Mechanism: Accountability structures embedded in hybrid decision systems.

4.6. Integrated System Dynamics

The five dimensions interact recursively:

1. Complementary capability alignment enhances decision quality.
2. Dynamic authority allocation strengthens this effect under uncertainty.
3. Explainability supports cognitive integration.
4. Trust calibration mediates effective collaboration.
5. Ethical governance ensures long-term legitimacy and sustainability.

Feedback loops connect performance outcomes back to authority thresholds and trust calibration mechanisms, creating a self-adjusting hybrid system.

5. Discussion

The findings synthesized in this study reinforce a growing consensus that organizational performance in AI-enabled environments depends less on technological sophistication alone and more on the architecture of human–AI interaction. Across domains, evidence consistently demonstrates that hybrid systems outperform isolated human or algorithmic decision-making when collaboration is intentionally structured (Bansal et al., 2020; Gao et al., 2023). The present model extends this insight by arguing that performance gains emerge from *co-agency*—a configuration in which humans and artificial agents operate as interdependent decision actors within governance boundaries.

A central observation emerging from the literature is that predictive accuracy does not equate to collaborative effectiveness. Bansal et al. (2020) show that the most accurate AI model may not produce the highest team utility. Similarly, Gao et al. (2023) demonstrate that routing only selected cases to human decision-makers can significantly improve collective outcomes. These findings underscore that hybrid performance is a function of allocation logic rather than raw algorithmic power. The discussion therefore shifts from "how accurate is the AI?" to "how is authority structured within the decision system?"

The evidence further highlights that authority in human–AI systems is dynamic and psychologically mediated. Wen, Wang, and Chen (2025) show that managerial trust directly influences the decision weight assigned to AI outputs. This suggests that organizational structures cannot be designed independently of cognitive and relational processes. Authority allocation is continuously shaped by perceived reliability, prior interaction outcomes, and contextual risk. Chen et al. (2018) demonstrate in human–robot collaboration that calibrated trust—rather than maximal trust—optimizes team

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

performance over time. This aligns with Hou et al. (2024), who argue that trust requires continuous monitoring and recalibration mechanisms to sustain effective collaboration.

Another consistent pattern concerns the role of communication and interpretability. Chen and Zhang (2025) find that misunderstandings in human–AI teams significantly reduce trust and degrade performance. These findings suggest that AI systems must be designed not only for analytical strength but also for communicative clarity. Okonkwo et al. (2025) emphasize that explainable AI enhances accountability and supports responsible integration in high-stakes environments. Meanwhile, Chen, Luo, and Sra (2025) demonstrate that explanation formats and interface design directly influence reliance behavior and task performance. Collectively, this body of work indicates that cognitive alignment between humans and AI is a necessary condition for sustained hybrid effectiveness.

The discussion also reveals tensions surrounding moral responsibility and legitimacy. Tolmeijer et al. (2022) show that AI systems are perceived as more capable but less morally trustworthy than human experts. Responsibility attribution often shifts toward developers or organizations, raising governance concerns. Dican (2025) argues that AI must augment rather than replace human ethical reasoning, reinforcing the need for structured oversight mechanisms. Lou et al. (2025) further identify value alignment and responsibility distribution as unresolved challenges in human–AI teaming. These findings suggest that organizational adoption of AI is not solely a performance issue but also a legitimacy issue shaped by perceptions of fairness, accountability, and transparency.

An additional theme emerging from the literature is the recursive nature of hybrid performance systems. Zercher et al. (2025) show that AI integration can reduce information asymmetries within teams, improving collective decision quality. However, these benefits depend on process design and trust dynamics. Performance feedback influences subsequent trust levels, which in turn reshape authority allocation. This feedback loop supports the argument that co-agency is not a static configuration but an evolving organizational capability.

6. Theoretical Implications

The co-agency framework advances theory in management, organizational behavior, and human–AI collaboration by reconceptualizing artificial intelligence as a structurally embedded decision actor rather than a passive technological artifact. Drawing on empirical and conceptual research, this section outlines five major theoretical contributions.

6.1. Reconceptualizing Agency in Organizations

Traditional management theories implicitly treat technology as an instrument under human control. However, research on human–AI teaming suggests that AI systems increasingly function as quasi-agential participants in decision processes. Tummala, Burris-Melville, and Eskridge (2025) argue that AI systems can operate as “team members,” influencing coordination and shared mental models. Lou et al. (2025) similarly describe AI agents as adaptive collaborators requiring new responsibility allocation frameworks.

The co-agency model extends these arguments by proposing that agency in organizations is distributed across human and artificial actors within governance constraints. This reframes agency theory by introducing *bounded artificial agency*—AI systems possess operational influence but remain institutionally constrained by oversight mechanisms. This perspective bridges organizational theory and socio-technical systems

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

research by positioning AI not merely as infrastructure but as an embedded actor in decision architectures.

6.2. Shifting Performance Evaluation from Algorithmic Accuracy to Hybrid Utility

A second theoretical contribution concerns performance metrics. Conventional AI evaluation prioritizes predictive accuracy, often independent of human interaction. Bansal et al. (2020) demonstrate that the most accurate AI is not necessarily the most effective teammate. Gao et al. (2023) further show that performance gains arise when tasks are routed between humans and AI according to complementary strengths.

The co-agency framework shifts the unit of analysis from isolated algorithmic performance to *hybrid system performance*. This theoretical move challenges dominant efficiency paradigms in operations and management research. It aligns evaluation criteria with organizational outcomes rather than technical benchmarks, thereby integrating decision science with organizational performance theory.

6.3. Integrating Trust Theory with Structural Design

Trust research has historically focused on interpersonal relationships. Human–AI collaboration extends trust theory into socio-technical contexts. Wen, Wang, and Chen (2025) demonstrate that trust influences the decisional weight assigned to AI systems. Chen et al. (2018) show that calibrated trust maximizes long-term performance in human–robot teams. Hou et al. (2024) argue that trust requires continuous monitoring and recalibration mechanisms.

The co-agency model advances trust theory by embedding it within structural governance design. Trust is conceptualized not solely as a psychological state but as a dynamic regulatory mechanism influencing authority allocation. This expands organizational trust theory into hybrid decision systems and introduces the concept of *trust-calibrated authority* as a structural construct.

6.4. Advancing Interpretability as a Cognitive Alignment Mechanism

Explainable AI research has primarily focused on transparency and accountability. Okonkwo et al. (2025) highlight the role of interpretability in high-stakes contexts, while Chen, Luo, and Sra (2025) show that interface design shapes reliance patterns and collaborative outcomes.

The co-agency framework theorizes explainability as a mechanism for *cognitive integration* between humans and AI. Rather than treating transparency as a compliance requirement, the model positions interpretability as a structural bridge aligning algorithmic reasoning with human mental models. This contributes to cognitive and decision-making theory by conceptualizing hybrid cognition as an emergent property of transparent interaction.

6.5. Extending Governance Theory to Hybrid Decision Systems

Organizational governance traditionally addresses accountability among human actors. Human–AI collaboration complicates responsibility distribution. Tolmeijer et al. (2022) show that AI systems are perceived as capable yet morally ambiguous, often shifting responsibility to organizations or developers. Dican (2025) argues that AI must augment rather than replace ethical reasoning.

The co-agency model extends governance theory by incorporating artificial agents into accountability structures. It introduces the concept of *algorithmic accountability embedded within organizational hierarchy*. This contribution integrates ethics, governance, and decision science into a unified framework for hybrid systems.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

6.6. Dynamic Systems Perspective on Organizational Adaptation

Finally, the framework contributes to dynamic capability theory by conceptualizing co-agency as an adaptive organizational capability. Zercher et al. (2025) show that AI can reduce information asymmetries in teams, but effectiveness depends on process design and trust dynamics. Performance feedback loops influence authority thresholds and trust recalibration, creating a recursive system.

This dynamic system perspective positions co-agency as an evolving organizational competence rather than a static technological implementation. It aligns hybrid intelligence research with broader theories of organizational adaptation and learning.

7. Practical Implications

The co-agency framework provides actionable guidance for organizations integrating AI into managerial and operational decision systems. Rather than treating AI deployment as a purely technological initiative, the model emphasizes structural design, authority calibration, trust governance, and ethical oversight as determinants of sustainable performance.

7.1. Design for Complementarity, Not Replacement

Empirical evidence shows that AI delivers the greatest value when it complements human strengths rather than replaces them. Bansal et al. (2020) demonstrate that optimizing AI for team utility produces better outcomes than maximizing standalone model accuracy. Gao et al. (2023) further show that routing only selected cases to human decision-makers significantly improves collective performance.

Practical implication

Organizations should redesign workflows around *comparative advantage mapping*. This involves:

- Identifying decision types suited for AI (high-volume, pattern-based, probabilistic forecasting).
- Reserving context-sensitive, ethically complex, or ambiguous cases for human oversight.
- Implementing decision-routing systems based on risk thresholds.

This shift moves firms from automation-driven cost logic to hybrid performance optimization.

7.2. Implement Adaptive Authority Structures

Authority in human–AI systems should be dynamic rather than fixed. Wen, Wang, and Chen (2025) show that managerial trust influences the weight assigned to AI recommendations. Chen et al. (2018) demonstrate that calibrated delegation improves long-term team outcomes.

Practical implication

Organizations should establish:

- Adjustable decision thresholds based on confidence scores.
- Escalation protocols for high-risk or high-impact decisions.
- Human override mechanisms embedded within AI systems.

Rather than binary control (human vs. machine), firms should adopt flexible authority allocation systems that adapt to contextual risk and performance feedback.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

7.3. Establish Continuous Trust Monitoring Mechanisms

Trust significantly affects collaboration quality. Hou et al. (2024) argue that trust must be continuously monitored and recalibrated to sustain effective human–AI teaming. Chen and Zhang (2025) demonstrate that misunderstandings reduce trust and impair performance.

Practical implication

Organizations should introduce trust calibration infrastructures, including:

- AI reliability dashboards.
- Transparent communication of model confidence levels.
- Feedback loops where humans can rate AI performance.
- Regular audits of AI recommendation accuracy.

Trust should be treated as a measurable organizational variable rather than an assumed outcome of technological adoption.

7.4. Prioritize Explainability in System Design

Explainable AI enhances both trust and decision legitimacy. Okonkwo et al. (2025) emphasize that interpretability strengthens accountability in high-stakes decisions. Chen, Luo, and Sra (2025) show that explanation formats influence user engagement and reliance patterns.

Practical Implication

Organizations should:

- Integrate transparent reasoning pathways into AI interfaces.
- Provide visual or textual explanations aligned with user expertise.
- Avoid “black-box” systems in strategic or ethical domains.

Explainability should be embedded at the design stage rather than treated as an afterthought.

7.5. Clarify Accountability and Governance Structures

Human–AI collaboration complicates responsibility allocation. Tolmeijer et al. (2022) show that AI systems are often perceived as capable but morally ambiguous, shifting responsibility toward organizations. Dican (2025) argues that AI must augment rather than replace human ethical reasoning.

Practical implication:

Organizations should:

- Define clear accountability frameworks specifying who is responsible for AI-supported decisions.
- Implement algorithmic auditing procedures.
- Establish bias detection and correction protocols.
- Maintain documented human oversight in critical decisions.

Embedding accountability mechanisms enhances organizational legitimacy and reduces reputational risk.

7.6. Develop Hybrid Capability as a Strategic Competence

Zercher et al. (2025) demonstrate that AI integration can reduce information asymmetries and improve collective decision-making when collaboration is properly structured. However, benefits depend on process design and trust dynamics.

Practical implication

Firms should treat co-agency as a long-term capability development initiative by:

- Training managers in AI literacy and decision calibration.
- Developing cross-functional AI governance committees.

Vol. 3 No. 12 (2025)

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

- Encouraging iterative learning through pilot implementations.

Hybrid intelligence becomes a strategic organizational asset rather than a one-time technological upgrade.

7.7. Manage Ethical and Legitimacy Risks Proactively

As AI systems gain decision authority, ethical scrutiny increases. Lou et al. (2025) highlight unresolved issues regarding responsibility distribution and value alignment in human–AI teams.

Practical implication

Organizations must proactively manage legitimacy by:

- Aligning AI objectives with organizational values.
- Ensuring transparency with stakeholders about AI usage.
- Engaging compliance, legal, and ethics teams in AI deployment.

Sustainable integration requires balancing performance enhancement with social responsibility.

8. References

Ajmal, M., & Suleman, S. A. (2015a). Organizational consciousness: A new paradigm for sustainable change management. *International Journal of Strategic Change Management, 6*(3), 254–267.

Ajmal, M., & Suleman, S. A. (2015b). Exploring organizational consciousness: A critical approach towards organizational behavior. *Pakistan Journal of Commerce and Social Sciences, 9*(1), 202–217.

Ajmal, M., Islam, A., & Islam, Z. (2024b). Unveiling organizational consciousness: A conceptual framework for nurturing thriving organizations. *Journal of Organizational Change Management, 37*(6), 1361–1381.

Ajmal, M., Islam, A., & Islam, Z. (2024b). Unveiling organizational consciousness: A conceptual framework for nurturing thriving organizations. *Journal of Organizational Change Management, 37*(6), 1361–1381.

Ajmal, M., Islam, A., & Khalid, S. (2025). A socio-technical systems perspective on organizational performance: Integrating soft systems methodology and high-performance work systems. *ASSAJ, 3*(02), 2672–2688.

Ajmal, M., Islam, A., & Khalid, S. (2025). Knowledge transcendence as a catalyst for organizational consciousness development. *Research Consortium Archive, 3*(4), 2336–2252.

Ajmal, M., Islam, A., & Khalid, S. (2025). The future of organizations: Moving from knowledge management toward organizational consciousness through knowledge transcendence. *Research Consortium Archive, 3*(3), 1738–1735.

Ajmal, M., Islam, A., & Khalid, S. (2025). Transforming organizational intelligence: Knowledge management systems and the path to knowledge transcendence. *Research Consortium Archive, 3*(2), 1116–1131.

Ajmal, M., Khalid, S., & Islam, A. (2025). A systems-based perspective on organizations: Integrating soft systems methodology and knowledge management systems. *Journal of Business and Management Research, 4*(5).

Ajmal, M., Khalid, S., & Islam, A. (2025). From knowledge assets to epistemic capital: Human–AI collective intelligence in organizations. *ASSAJ, 4*(01), 4721–4735.

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Ajmal, M., Khalid, S., & Islam, A. (2025). Organizational problem solving as a conscious process: Integrating soft systems methodology and organizational consciousness. *Journal of Business and Management Research, 4*(4).

Ajmal, M., Manzoor Dar, W., Islam, A., & Islam, Z. (2025). Empowering collective intentionality: A framework for social change through group consciousness. *World Futures, 81*(1), 15–34.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., & Weld, D. S. (2020). Optimizing AI for teamwork. *arXiv*.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018). Trust-aware decision making for human-robot collaboration. *ACM Transactions on Human-Robot Interaction, 9*(1), Article 3, 1–23. https://doi.org/10.1145/3359616

Chen, N., & Zhang, X. (2025). When misunderstanding meets artificial intelligence: The critical role of trust in human–AI and human–human team communication and performance. *Frontiers in Psychology, 16*, Article 1637339. https://doi.org/10.3389/fpsyg.2025.1637339

Chen, Z., Luo, Y., & Sra, M. (2025). Engaging with AI: How interface design shapes human-AI collaboration in high-stakes decision-making. *arXiv*. https://doi.org/10.48550/arxiv.2501.16627

Dican, L. (2025). Human-AI collaboration and its impact on decision-making. *International Journal of Multidisciplinary Research and Growth Evaluation, 6*(2), 919–923. https://doi.org/10.54660/.ijmrge.2025.6.2.919-923

Gao, R., Saar-Tsechansky, M., De-Arteaga, M., Han, L., Sun, W., Lee, M. K., & Lease, M. (2023). Learning complementary policies for human-AI teams. *arXiv*. https://doi.org/10.48550/arxiv.2302.02944

Hou, M., Banbury, S., Cain, B., Fang, S., Willoughby, H., Foley, L., Tunstel, E., & Rudas, I. J. (2024). IMPACTS homeostasis trust management system: Optimizing trust in human-AI teams. *ACM Computing Surveys, 57*(3), Article 69, 1–24. https://doi.org/10.1145/3649446

Islam, A., Ajmal, M., & Khalid, S. (2025). Beyond knowledge management: Reframing organizations as knowledge ecologies for a wisdom-based management paradigm. *Pakistan Journal of Social Science Review, 4*(7), 786–798.

Islam, A., Khalid, S., & Ajmal, M. (2025). From complexity to clarity: Merging soft systems thinking with knowledge transcendence in modern organizations. *Journal of Business and Management Research, 4*(3).

Khalid, S., Islam, A., & Ajmal, M. (2025). From academic freedom to algorithmic agency: Knowledge governance in AI-enhanced education. *Journal of Management Science Research Review, 4*(3), 1036–1058.

Leitão, D., Saleiro, P., Figueiredo, M. A. T., & Bizarro, P. (2022). Human-AI collaboration in decision-making: Beyond learning to defer. *arXiv*. https://doi.org/10.48550/arxiv.2206.13202

Lou, B., Lu, T., Santanam, R., & Zhang, Y. (2025). Unraveling human-AI teaming: A review and outlook. *arXiv*. https://doi.org/10.48550/arxiv.2504.05755

Okonkwo, R., Folorunso, A., Ogundipe, F., & Tettey, C. Y. (2025). Explainable artificial intelligence (AI) through human-AI collaborative frameworks: Quantifying trust and interpretability in high-stakes decisions. *Computer Science & IT Research Journal, 6*(5). https://doi.org/10.51594/csitrj.v6i5.1934

Journal of Social Signs Review

Online ISSN 3006-4651

Print ISSN 3006-466X

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3491102.3517732

Tummala, V., Burris-Melville, T. S., & Eskridge, T. C. (2025). AI as a team member: Redefining collaboration. *Journal of Leadership Studies*. https://doi.org/10.1002/jls.70003

Wen, Y., Wang, J., & Chen, X. (2025). Trust and AI weight: Human-AI collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, Article 1419403. https://doi.org/10.3389/forgp.2025.1419403

Zercher, D., Jussupow, E., Benke, I., & Heinzl, A. (2025). How can teams benefit from AI team members? Exploring the effect of generative AI on decision-making processes and decision quality in team–AI collaboration. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2898