Management Analytics and Social Insights

www.masi.reapress.com

Manag. Anal. Soc. Insights. Vol. 3, No. 4 (2026) 258–271.

Paper Type: Original Article

Calibrated Human–Generative AI Decision Intelligence: A Governance-Aware Framework for Managerial Analytics

Mohammadtaghi Azimi[1], **Seyedeh Elnaz Azimimi**[2,*]

[1] Department of Management, Shenzhen University, Shenzhen, China; m.azimi@e.gzhu.edu.cn.

[2] Department of Civil Engineering, Shenzhen University, Shenzhen, China; ElnazAzimii.94@Gmail.Com.

Citation:

Received: 19 May 2026 *Revised: 24 July 2026* *Accepted: 18 September 2026*	Azimi, M., & Azimimi, S. E. (2026). Calibrated human–generative AI decision intelligence: A governance-aware framework for managerial analytics. *Management Analytics and Social Insights, 3*(4), 258-271.

Abstract

Generative Artificial Intelligence (GenAI) is moving from an experimental productivity tool to an embedded component of managerial analytics, yet organizations still lack a decision architecture that explains when managers should delegate, challenge, verify, or override model outputs. This conceptual study develops a calibrated human–GenAI decision intelligence framework that integrates evidence from management, behavioral decision research, human–automation interaction, explainable artificial intelligence, and responsible artificial intelligence governance. The framework treats decision performance as a joint property of task characteristics, human expertise, model capability, calibration mechanisms, and organizational controls rather than as a property of the model alone. It specifies two coupled layers: a calibration layer that makes uncertainty, evidence quality, and model limitations visible, and a collaboration layer that allocates tasks, preserves human contestability, and creates explicit escalation paths. Eight propositions explain how task ambiguity, decision stakes, expertise, verification friction, explanation quality, and accountability shape appropriate reliance. A six-stage operating loop and a five-level governance maturity model translate the framework into implementable managerial analytics practices. The central implication is that organizations should optimize neither automation nor human control in isolation; they should optimize calibrated complementarity, in which GenAI expands the search and synthesis space while humans retain responsibility for framing, contextual judgment, value-sensitive trade-offs, and final accountability. The framework offers a theory-building platform for empirical research and a practical blueprint for deploying GenAI in managerial decision processes without sacrificing decision quality, organizational learning, or responsible governance.

Keywords: Generative artificial intelligence, Human–AI collaboration, Managerial analytics, Decision intelligence, Trust calibration, Responsible artificial intelligence.

1 | Introduction

Artificial Intelligence (AI) has become a central element of contemporary managerial decision support. A recent structured review of the management-decision literature shows that AI is increasingly used for

Corresponding Author: ElnazAzimii.94@Gmail.Com @aihe.ac.ir

https://doi.org/10.22105/masi.v3i4.112

forecasting, classification, knowledge extraction, and decision support across managerial domains [1]. Earlier work had already argued that AI should not be understood only as a substitute for human judgment, because computational systems and managers possess different strengths when organizations face complexity, uncertainty, and equivocality [2]. This complementarity has motivated organizational designs that combine human and machine decisions through delegation, sequential processing, and aggregation [3]. Yet the managerial problem has become more difficult with the emergence of Generative Artificial Intelligence (GenAI): the technology is not limited to prediction but can produce explanations, scenarios, recommendations, summaries, arguments, and executable content in natural language.

The shift from predictive AI to GenAI changes the locus of risk. In traditional analytics, managers often receive a score, forecast, or classification and can inspect the input variables or model documentation. In GenAI-supported work, the system can participate in the reasoning process itself. It may frame a problem, propose alternatives, simulate stakeholders, synthesize evidence, and justify a recommendation. Consequently, managers can experience a fluent output as if it were a reliable argument even when the underlying evidence is weak. This problem intersects with the automation–augmentation paradox: attempts to augment managerial work can silently automate portions of cognition, while attempts to automate can create new human coordination and monitoring demands [4].

Empirical evidence demonstrates why a more precise architecture is needed. Controlled experiments show that GenAI can improve productivity and output quality on professional writing tasks [5], while field evidence from customer-support work indicates substantial productivity gains, particularly for less experienced workers [6]. Research with knowledge workers also reveals a 'jagged technological frontier': performance improves strongly when tasks fall within model capability but can deteriorate when users apply the technology outside that frontier [7]. These findings imply that the managerial objective is not maximal AI use. It is appropriate reliance, using AI when its expected contribution exceeds its risk, and withholding or challenging it when uncertainty, stakes, or contextual complexity make unaudited delegation unsafe.

This problem is also behavioral. Trust in AI is shaped by representation, perceived capability, transparency, and experience [8]. People may reject useful algorithms after observing errors, a phenomenon described as algorithm aversion [9], yet under other conditions they weight algorithmic advice more heavily than equivalent human advice, demonstrating algorithm appreciation [10]. Trust therefore cannot be treated as a stable attitude. It must be calibrated to the system's actual competence on a particular task and to the consequences of error. A manager who uniformly distrusts GenAI wastes augmentation potential; a manager who uniformly trusts it converts a probabilistic assistant into an unaccountable authority.

The motivation of this paper is the gap between rapidly expanding GenAI capability and the slower development of managerial operating logic. Existing scholarship explains important pieces of the problem, human–AI complementarity, trust, explanation, automation levels, productivity, and ethics, but managers still lack an integrated framework that connects these pieces to the sequence of a real decision. The key research question is therefore: How should organizations design human–GenAI decision processes so that productivity and analytical reach improve without degrading judgment, accountability, skill, and governance?

The paper makes four contributions. First, it introduces calibrated complementarity as the organizing principle for GenAI-enabled managerial analytics. Second, it develops a two-layer framework in which calibration mechanisms and collaboration mechanisms jointly determine decision quality and responsible use. Third, it formulates eight propositions that link task conditions, human expertise, uncertainty visibility, verification friction, and accountability to appropriate reliance. Fourth, it translates the conceptual model into a six-stage operating loop and a five-level maturity architecture that can guide organizational implementation. *Fig. 1* previews the proposed framework, which is developed in detail in Section 4.

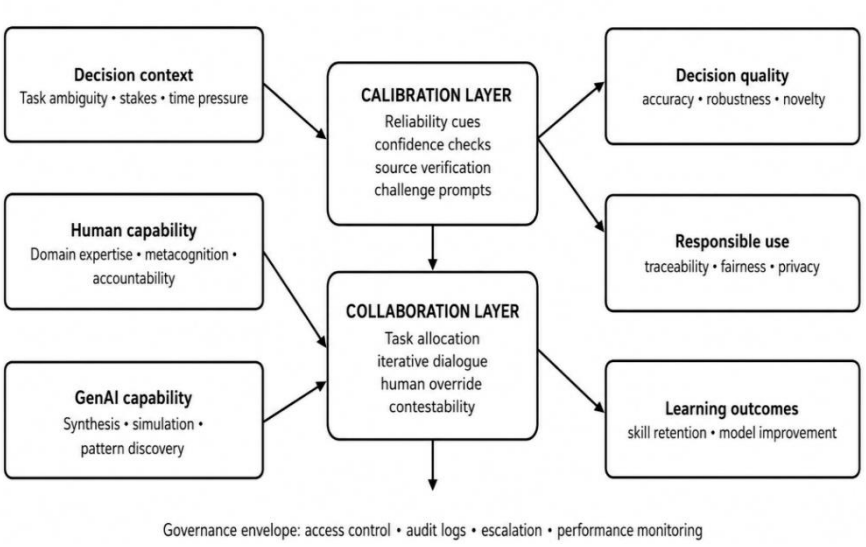

Fig. 1. Calibrated human–GenAI decision intelligence framework.

The remainder of the paper is organized as follows. Section 2 reviews the literature on algorithmic judgment, human–AI collaboration, explanation, GenAI, and responsible governance. Section 3 explains the integrative theory-development approach and defines the core constructs. Section 4 develops the framework and propositions. Section 5 provides an implementation and governance architecture. Section 6 discusses theoretical and managerial implications and concludes with limitations and future research directions.

2 | Literature Review

2.1 | From Algorithmic Judgment to Human–AI Collaboration

The intellectual foundations of AI-assisted decision making predate current machine learning systems. Research comparing clinical and actuarial judgment showed that formal statistical procedures can outperform unaided expert judgment in many prediction settings [11]. The lesson, however, is not that human judgment is dispensable. Rather, formal methods are powerful when the target is sufficiently structured, repeated, and measurable. As automation expands into more complex tasks, the design question shifts from whether to automate to what type and level of automation should be applied to information acquisition, information analysis, decision selection, and action implementation [12].

Trust is the behavioral mechanism through which technical capability becomes actual reliance. Lee and See [13] define appropriate reliance as a match between trust and automation capability, emphasizing that both misuse and disuse can reduce performance. This insight is particularly important for GenAI because model behavior is probabilistic, task-sensitive, and often difficult to infer from surface fluency. Human–AI interaction guidance similarly stresses that systems should communicate what they can do, support efficient correction, make uncertainty visible, and enable users to recover from errors [14].

Organizational research extends these human-factors insights by showing that algorithmic systems reshape expertise, coordination, and control. Learning algorithms can alter occupational boundaries and organizational knowledge because they embed predictions and classifications into workflows [15]. Algorithmic control can also become contested when employees cannot understand, challenge, or influence automated evaluations and assignments [16]. These findings imply that GenAI governance cannot be reduced to model accuracy: it must include authority allocation, contestability, and the preservation of meaningful human agency.

2.2 | Algorithm Aversion, Appreciation, and Calibration

Behavioral evidence shows that reliance on algorithmic advice is conditional rather than uniformly low or high. Dietvorst et al. [9] found that people can become disproportionately averse to algorithms after observing them make mistakes, even when the algorithm still outperforms a human forecaster. In contrast, Logg et al. [10] documented algorithm appreciation, showing that people sometimes place greater weight on algorithmic than human judgment. Castelo et al. [17] reconciled part of this tension by showing that algorithm aversion is task dependent: people are less willing to use algorithms for tasks perceived as subjective than for tasks perceived as objective. The implication for managerial analytics is that perceived task character can distort reliance independently of actual model competence.

Human–AI teams introduce additional dynamics. Updates that improve a model's standalone accuracy can reduce team performance when they violate users' learned expectations about where the model succeeds and fails [18]. Moreover, simply supplying explanations is not guaranteed to correct overreliance. Cognitive forcing interventions that require users to make an initial judgment, inspect alternatives, or actively request AI advice can reduce excessive deference, although such interventions may impose usability costs [19]. Thus, calibration is not synonymous with explanation. It is a broader process by which users learn when to rely, when to verify, and when to reject.

Evidence from public-sector decision making reinforces this point. Experimental studies found no universal automation bias but did observe selective adherence patterns in which decision makers gave different weight to advice depending on context and stereotypes [20]. This suggests that 'human in the loop' is not itself a sufficient safeguard. Human oversight is only valuable when the human has the information, incentives, time, competence, and authority required to scrutinize the AI output.

2.3 | Explainability, Interpretability, and Decision Support

Explainability research provides tools for supporting scrutiny but also cautions against treating explanation as a universal remedy. Doshi-Velez and Kim [21] argue that interpretability requires rigorous, context-specific evaluation rather than intuitive claims. Miller [22] shows that effective explanations are selective, contrastive, and social: people usually want to know why one outcome occurred rather than another, not to receive an exhaustive description of a model. User-centered explainable AI research likewise finds that explanation needs vary across tasks and users, and that designers should begin with the questions users actually need answered [23].

For high-stakes decisions, the opacity problem can be more fundamental. Rudin [24] argues that where interpretable models can achieve adequate performance, they may be preferable to post hoc explanations of black-box models. GenAI complicates this issue because explanations can themselves be generated outputs rather than faithful accounts of internal reasoning. Therefore, managerial use should distinguish between explanatory usefulness and epistemic evidence. A plausible rationale can help a manager interrogate a recommendation, but it should not be treated as proof that the recommendation is correct.

2.4 | Generative AI, Foundation Models, and Knowledge Work

Foundation models are trained on broad data at scale and adapted to many downstream tasks, creating substantial leverage but also propagating common model defects across applications [25]. Large language models can generate fluent content while reproducing biases, privacy risks, representational harms, and misleading information [26]. Hallucination, generation that is unfaithful to source material or unsupported by evidence, has been documented as a persistent natural-language-generation problem [27]. Holistic evaluation therefore requires more than accuracy; calibration, robustness, fairness, toxicity, and efficiency are also relevant [28].

In organizations, GenAI expands the range of tasks that can be augmented. It can summarize long documents, draft managerial narratives, generate alternative hypotheses, produce code, simulate customers, and translate

analytical outputs into natural language. Feuerriegel et al. [29] position GenAI as a general-purpose technology with broad implications for information systems and organizational processes. Dwivedi et al. [30] similarly identify opportunities for productivity and creativity alongside risks related to bias, outdated knowledge, transparency, credibility, and misuse. The empirical productivity literature [5–7] suggests that benefits are real, but the jaggedness of capability means that organizations need mechanisms for identifying where those benefits are reliable.

2.5 | Responsible AI and the Governance Gap

Responsible AI research has converged around recurring principles. The AI4People framework emphasizes beneficence, non-maleficence, autonomy, justice, and explicability [31]. A global review of AI ethics guidelines found broad agreement around transparency, fairness, non-maleficence, responsibility, and privacy, but substantial variation in operationalization [32]. Mittelstadt [33] cautions that principles alone cannot guarantee ethical AI because organizations need concrete practices, incentives, accountability arrangements, and professional norms that translate abstract commitments into behavior.

This implementation gap is especially important for GenAI because the technology is easy to access and difficult to contain within traditional analytics governance. Employees can paste confidential data into external systems, accept fabricated citations, or use generated recommendations without preserving an audit trail. Responsible managerial analytics therefore requires governance-by-design: technical controls, workflow rules, review thresholds, documentation, and escalation paths must be integrated into normal decision processes rather than added after deployment.

Table 1. Literature streams informing calibrated human–GenAI decision intelligence.

Literature Stream	Core Insight	Unresolved Managerial Issue
Algorithmic judgment and automation	Formal models can outperform intuition in structured tasks; automation should be allocated by function and level [11–13].	How should GenAI authority vary across decision stages?
Behavioral reliance	Users may exhibit aversion, appreciation, selective adherence, or overreliance [9], [10], [17], [19], [20].	How can organizations calibrate reliance rather than maximize trust?
Human–AI teams	Team performance depends on compatibility, task allocation, and user understanding [18].	How should roles and checkpoints adapt as models change?
Explainability and interpretability	Explanations must be user-centered and cannot substitute for evidence [21–24].	Which explanations support verification instead of persuasion?
GenAI and foundation models	General-purpose capability creates productivity gains but also hallucination and broad risk propagation [25–30].	How can managers detect the jagged frontier in everyday work?
Responsible AI	Ethical principles require operational controls and accountability [31–33].	How should governance be embedded in decision workflows?

3 | Conceptual Development and Analytical Method

This study uses an integrative conceptual synthesis rather than an empirical dataset. The objective is theory development: to connect research streams that are often studied separately and derive a managerial decision architecture that is internally coherent, testable, and implementable. The synthesis is anchored in peer-reviewed work on management decision making, human–automation interaction, behavioral responses to algorithms, human–AI collaboration, explainable AI, foundation models, GenAI productivity, and responsible AI. Recent management research indicates that AI decision scholarship remains fragmented across forecasting, knowledge management, decision support, and organizational use [1], while emerging work emphasizes that collaboration orientations differ substantially among managers [34]. A broad review of

human–machine collaborative decision making likewise identifies role configuration, interaction, trust, explanation, authority, and responsibility as central mechanisms [35].

The analytical unit is a decision episode rather than a technology adoption event. A decision episode begins when a managerial problem is framed and ends when the decision outcome is monitored and incorporated into organizational learning. This unit of analysis is important because the appropriate role of GenAI can differ within the same episode: the model may be highly useful for generating alternatives, moderately useful for evaluating evidence, and inappropriate as the final decision authority. Recent evidence on human–machine collaboration supports the need to examine decision stages and organizational incentives jointly rather than treating AI use as a binary variable [36].

Five constructs organize the framework. First, task ambiguity is the degree to which the decision lacks a stable objective function, agreed criteria, or well-defined search space. Second, decision stakes capture the expected cost, irreversibility, and distributional consequences of error. Third, model-task competence is the expected capability of a specific GenAI configuration on a specific task, including data access and tool use. Fourth, human epistemic position captures the decision maker's domain expertise, ability to verify, and awareness of uncertainty. Fifth, governance strength captures the technical and organizational capacity to constrain inputs, preserve logs, define review thresholds, and assign responsibility.

From these constructs, the paper derives the idea of calibrated complementarity. Complementarity means that humans and GenAI contribute different capabilities; calibration means that their relative influence changes as evidence about competence and risk changes. The resulting architecture is neither human-centered in the sense of preserving human control at all costs nor AI-centered in the sense of delegating whenever model performance is high. It is decision-centered: authority is assigned to the actor or combination of actors most likely to produce defensible outcomes under the actual task conditions.

4|Calibrated Human–GenAI Decision Intelligence Framework

4.1|The Calibration Layer

The calibration layer determines how much epistemic weight a manager should assign to a GenAI output. It contains four mechanisms: Reliability cues, confidence checks, source verification, and challenge prompts. Reliability cues include known model limitations, historical performance on similar tasks, presence of retrieval or external tools, and the age and provenance of the supporting information. Confidence checks require the system and user to expose uncertainty rather than compress it into a single fluent recommendation. Source verification requires claims that matter to the decision to be tied to accessible evidence. Challenge prompts deliberately search for counterarguments, boundary conditions, missing stakeholders, and failure scenarios.

The calibration layer addresses a central weakness of conversational interfaces: linguistic confidence is not equivalent to epistemic confidence. The system may produce equally polished language when evidence is strong, weak, or absent. Accordingly, organizations should create external reliability signals instead of expecting managers to infer reliability from prose style. This logic leads to the first two propositions.

Proposition 1. The positive effect of GenAI use on decision quality will be stronger when model-task competence is made observable through task-specific performance evidence, source provenance, and explicit uncertainty cues.

Proposition 2. The risk of overreliance will increase as the gap between linguistic confidence and verifiable evidence increases, particularly under time pressure and high cognitive load.

4.2|The Collaboration Layer

The collaboration layer determines how tasks and authority are distributed between human and GenAI. Four mechanisms are central: Task allocation, iterative dialogue, human override, and contestability. Task allocation assigns AI to activities where breadth, speed, pattern synthesis, and alternative generation create value, while

preserving human leadership where contextual interpretation, values, responsibility, and stakeholder legitimacy dominate. Iterative dialogue allows the manager to refine assumptions and expose disagreement. Human override ensures that the manager can reject AI outputs without procedural penalty. Contestability gives affected organizational actors a route to question the data, reasoning, or decision process. The key design choice is not simply whether a human approves the final output. A nominal approval step can become a rubber stamp when the user lacks time or independent evidence. Effective collaboration therefore requires role clarity before the model is consulted. Managers should know whether the AI is acting as a generator, critic, analyst, simulator, verifier, or recommender. The role determines both the expected value and the verification burden.

Proposition 3. Human–GenAI teams will outperform either actor alone when task allocation matches comparative capability and when the human retains sufficient independent information to evaluate AI-generated recommendations.

Proposition 4. Human oversight will have little protective value when verification costs are high relative to available time, even if formal approval authority remains with the manager.

4.3|Task Ambiguity, Stakes, and Expertise as Boundary Conditions

Task ambiguity and decision stakes jointly determine the appropriate level of delegation. Low-ambiguity, low-stakes tasks, such as formatting a report, summarizing a meeting, or generating initial descriptive insights, can tolerate relatively high levels of GenAI autonomy. As ambiguity increases, value judgments and contextual knowledge become more important. As stakes increase, the cost of hallucination, bias, privacy leakage, or misunderstood assumptions rises. The combination of high ambiguity and high stakes is therefore the strongest case for human-led decision making supported, but not governed, by GenAI. Human expertise adds a second boundary condition. Experts can identify subtle model failures, but expertise does not automatically guarantee good oversight. Experts may become complacent when tools perform well repeatedly, while novices may lack the domain knowledge needed to detect errors. Conversely, GenAI may create disproportionate benefits for less experienced workers by providing structured examples and tacit-knowledge-like guidance [6]. This suggests a non-linear relationship: the optimal role of AI depends on whether human expertise is sufficient for verification and whether the AI is being used to extend or substitute that expertise.

Proposition 5. The optimal degree of GenAI delegation will decrease as the joint level of task ambiguity and decision stakes increases.

Proposition 6. GenAI will create larger productivity gains for less experienced managers on structured tasks, but the probability of undetected decision error will also rise when those managers lack independent verification capability.

4.4|Accountability, Learning, and Dynamic Calibration

Calibration should be dynamic. Models change, retrieval sources change, organizational data change, and users adapt their behavior. A system that was reliable last quarter may become unreliable after a model update or after a workflow expands to new tasks. Human–AI team research shows that even performance-improving updates can disrupt established patterns of reliance [18]. Accordingly, organizations should monitor not only model accuracy but also override rates, verification failures, recurring prompt patterns, incident types, and the distribution of AI use across employees and decisions. Accountability is the mechanism that keeps dynamic calibration from becoming optional. When managers know that consequential AI-assisted decisions must be documented and defensible, they have stronger incentives to preserve evidence and avoid uncritical delegation. At the same time, accountability should be attached to organizational design, not merely individual blame. If a firm mandates the use of a system, constrains employee discretion, and fails to provide verification resources, it cannot reasonably treat errors as solely user failures.

Proposition 7. Decision quality will improve over time when organizations treat overrides, detected errors, and near misses as learning signals that update task-specific AI policies and user guidance.

Proposition 8. Clear accountability and contestability will moderate overreliance by increasing the expected value of verification and preserving meaningful human agency.

Table 2. Core constructs and testable propositions.

Construct	Operational Interpretation	Related Propositions
Model-task competence	Observed reliability of a specific GenAI configuration for a specific task.	P1, P3
Evidence visibility	Availability of sources, uncertainty cues, and verifiable support.	P1, P2
Verification friction	Time and cognitive effort required to independently check the output.	P2, P4
Task ambiguity	Lack of stable criteria, objective function, or bounded search space.	P5
Decision stakes	Expected cost, irreversibility, and distributional consequences of error.	P5
Human epistemic position	Expertise and access to independent information needed for evaluation.	P3, P6
Organizational learning	Use of errors, overrides, and outcomes to revise policies and models.	P7
Accountability and contestability	Clarity of responsibility and ability to challenge AI-supported decisions.	P8

4.5 | A Six-Stage Decision Operating Loop

Fig. 2 converts the framework into an operating sequence. *Stage 1* is problem framing: the human defines the decision, affected stakeholders, objective, constraints, and acceptable risk. *Stage 2* is option generation: GenAI expands the search space and can surface alternatives that a manager may not immediately consider. *Stage 3* is evidence verification: Material claims are checked against authoritative sources, internal data, or independent analytical tools. *Stage 4* is judgment and selection: The human integrates evidence with organizational values, commitments, and contextual knowledge. *Stage 5* is execution and monitoring: The organization tracks both operational outcomes and indicators of AI-related failure. *Stage 6* is learning and recalibration: The decision episode becomes training data for policy, workflow, and capability improvement.

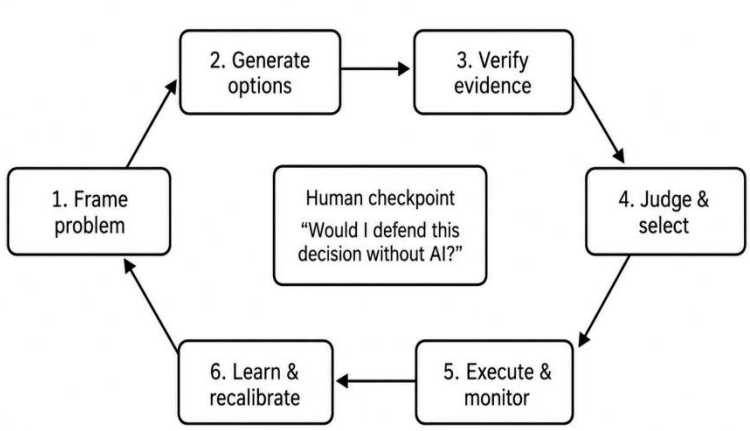

Fig. 2. Six-stage operating loop for calibrated GenAI-assisted managerial decisions.

The human checkpoint at the center of *Fig. 2* is intentionally simple: 'Would I defend this decision without AI?' The question tests whether the manager understands the argument well enough to own it. A "no" response does not always mean that the decision is wrong; it means that additional verification or escalation is required before the decision becomes organizational action.

5 | Managerial Implementation and Governance Architecture

5.1 | Decision-Risk Tiers

Implementation should begin by classifying decisions rather than by classifying tools. A three-tier structure is sufficient for many organizations. Tier 1 contains low-stakes, reversible tasks such as drafting, summarization, brainstorming, translation, and exploratory analysis. These tasks can permit broad GenAI use with basic confidentiality and disclosure rules. Tier 2 contains decisions with meaningful operational or financial consequences, such as supplier evaluation, pricing analysis, staffing plans, or customer segmentation. These require approved data sources, documented verification, and named human ownership. Tier 3 contains high-stakes, regulated, irreversible, or rights-affecting decisions. Here, GenAI should normally function as an advisory or analytical tool rather than as final decision authority, with independent review and auditable evidence. Risk tiering prevents a common governance failure: applying the same controls to every use case. Excessive controls on low-risk use drive employees toward shadow systems, while weak controls on high-risk decisions create avoidable exposure. The appropriate control intensity should therefore scale with decision consequences, not with the novelty of the technology.

5.2 | Workflow Controls and Evidence Standards

A governance-aware workflow should specify what data may enter the model, which tools are approved, when retrieval from authoritative sources is mandatory, which outputs require human verification, and what must be logged. For consequential decisions, the organization should preserve at least the problem statement, material inputs, model or system version, relevant generated outputs, verified sources, human modifications, final rationale, and approval identity. This record transforms AI use from an invisible cognitive aid into an auditable component of managerial analytics. Evidence standards should distinguish three classes of generated content. Class A content is generative but not factual, such as wording alternatives, brainstorming, or stylistic revisions; it normally requires minimal verification. Class B content contains factual or analytical claims that can be checked against authoritative data; verification should be mandatory when the claim affects a decision. Class C content contains recommendations that embed values, trade-offs, or stakeholder consequences; these require human judgment even when the underlying facts are correct.

5.3 | Governance Maturity

Fig. 3 presents a five-level maturity path. Level 1 is ad hoc use, where individuals employ GenAI without shared rules or logging. Level 2 introduces approved tools and baseline policies. Level 3 adds calibration mechanisms, role-based review, and explicit verification thresholds. Level 4 integrates GenAI into managed workflows with secure data access, application programming interfaces, audit trails, and performance monitoring. Level 5 adds adaptive governance in which policies, evaluation sets, and workflow controls are continuously updated from incidents, outcomes, and changes in model capability.

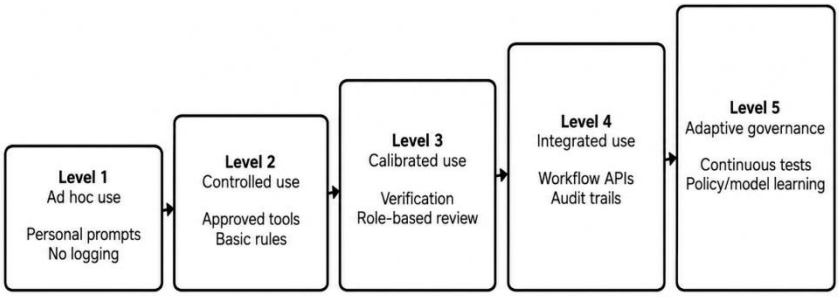

Fig. 3. Five-level governance maturity model for GenAI-enabled managerial analytics.

The maturity model is deliberately cumulative. Organizations should not move to deep integration merely because they have purchased enterprise licenses. Integration without calibration can scale mistakes faster.

Similarly, strong governance without useful capability can generate bureaucratic cost without analytical benefit. Maturity therefore requires simultaneous progress in capability, controls, and managerial accountability.

Table 3. Minimum controls by managerial decision-risk tier.

Control Domain	Tier 1: Low Risk	Tier 2: Material Risk	Tier 3: High/Regulated Risk
Approved system	Recommended	Required	Required; enterprise-controlled
Sensitive-data restriction	Yes	Yes, with role-based access	Strict; legal/privacy review where relevant
Source verification	For material factual claims	Mandatory for decision-critical claims	Independent verification required
Human owner	Task owner	Named decision owner	Senior accountable decision owner
Audit trail	Optional/lightweight	Required	Required and retained per policy
Independent review	Not normally needed	Conditional	Required before execution
Escalation threshold	User judgment	Low confidence or material disagreement	Any unresolved uncertainty or policy exception
Outcome monitoring	Sample-based	Periodic	Continuous/structured

5.4 | Managerial Capability Building

Training should focus less on prompt tricks and more on decision discipline. Managers need to understand how to decompose tasks, identify claims requiring evidence, recognize when the model lacks access to relevant information, design adversarial checks, and document final responsibility. This skill set resembles analytical literacy combined with epistemic risk management. It should be assessed through realistic decision scenarios rather than through generic AI awareness modules.

Organizational incentives also matter. Recent research finds that managerial collaboration orientations toward AI vary considerably and that organizational support for experimentation is associated with more strategic engagement [34]. At the same time, emerging evidence suggests that incentives and AI capability interact to shape human–machine collaboration over time [36]. Consequently, leaders should reward useful challenge and documented correction rather than interpreting disagreement with AI as resistance to innovation. A healthy GenAI culture is one in which employees can exploit the technology aggressively for search and synthesis while remaining conservative about unverified claims and consequential delegation.

6 | Discussion

6.1 | Theoretical Implications

This paper reframes human–GenAI collaboration from a trust problem into a calibration problem. Trust is psychologically important, but managerial performance depends on whether reliance matches actual task-specific competence. The concept of calibrated complementarity extends automation–augmentation theory [4] by specifying how authority should move across stages of a decision episode rather than treating automation and augmentation as organization-level strategies. It also extends human–AI team research by placing evidence visibility, verification friction, and governance strength alongside task allocation and trust.

A second contribution is the separation of explanatory usefulness from evidentiary validity. GenAI can generate explanations that help a manager explore assumptions, but the presence of an explanation does not establish the truth of a claim. The framework therefore positions explanation inside a larger calibration layer that includes provenance, uncertainty, independent verification, and deliberate challenge. This helps connect explainable AI research [21–24] with responsible managerial analytics.

A third contribution is the decision-episode perspective. Organizational AI studies often classify firms, technologies, or use cases. The present framework argues that the unit of design should be the sequence of cognitive and organizational activities that constitute a decision. The same GenAI system can legitimately have different authority at different stages. This provides a theoretically cleaner way to reconcile strong productivity evidence [5–7] with persistent concerns about hallucination, opacity, and responsibility [25–33].

6.2 | Managerial Implications

For managers, the framework yields a direct practical rule: use GenAI to widen the search space before allowing it to narrow the decision space. Generating alternatives, summarizing information, translating formats, identifying missing questions, and simulating scenarios are often high-value activities because errors can still be caught before action. By contrast, allowing the system to make the final choice is much more consequential because a fluent output may compress uncertainty and obscure value judgments. This distinction helps organizations capture productivity without prematurely transferring authority.

A second implication is that verification capacity should be treated as an organizational resource. A firm cannot responsibly scale AI-assisted decisions if employees lack access to authoritative data, domain experts, or time for review. Verification friction predicts whether formal human oversight will be substantive or ceremonial. Investments in retrieval infrastructure, data governance, audit logging, and escalation channels may therefore create as much value as investments in model access itself.

Third, governance should become adaptive. Organizations should maintain task-specific evaluation sets, monitor failure patterns, and update acceptable-use policies when models or workflows change. The literature on human–machine collaboration shows that system updates can alter team behavior even when technical performance improves [18]. Continuous governance is therefore necessary to prevent yesterday's trust calibration from becoming tomorrow's automation bias.

6.3 | Conclusion

Generative artificial intelligence is likely to become a routine component of managerial analytics, but routine use should not be confused with routine reliability. The central challenge is to design decision processes in which the technology's generative breadth, speed, and synthesis capability complement rather than displace contextual judgment, accountability, and organizational learning. The calibrated human–GenAI decision intelligence framework developed in this paper addresses this challenge through two coupled layers: Calibration mechanisms that expose reliability and evidence, and collaboration mechanisms that assign tasks and preserve meaningful human authority. The framework further specifies eight propositions, a six-stage operating loop, and a five-level governance maturity model that together provide a research agenda and an implementation blueprint.

The paper is conceptual and therefore has limitations. It does not estimate causal effects, compare specific commercial models, or validate the proposed propositions across industries and cultures. Future work should test the framework experimentally and longitudinally, measure task-specific calibration error, and examine how expertise, incentives, organizational hierarchy, and regulation affect reliance. Particularly valuable studies would compare different collaboration designs on the same managerial tasks, track how model updates change team behavior, and quantify the trade-off between verification effort and decision risk. Cross-cultural research is also needed because authority, trust, and willingness to challenge automated recommendations can vary across institutional settings. These limitations are deliberate opportunities for empirical development: the framework is intended to provide a falsifiable and operational starting point for a cumulative science of responsible GenAI-enabled managerial decision-making.

Acknowledgments

The authors thank the editor and reviewers for their valuable comments.

Author Contribution

S. M. Azimi: Conceptualization, methodology, formal analysis, writing, original draft, writing, review and editing, and visualization. S. E. Azimi: Visualization, methodology, conceptualization, writing, original draft, writing, review and editing. All authors have read and approved the published version of the manuscript.

Funding

No external funding is declared for this conceptual study.

Data Availability

No new empirical dataset was generated or analyzed in this conceptual study. The article is based on the published literature cited in the reference list.

Conflicts of Interest

The author(s) declare no conflict of interest.

References

[1] Oppioli, M., Sousa, M. J., Sousa, M., & de Nuccio, E. (2025). The role of artificial intelligence for management decision: A structured literature review. *Management Decision, 63*(10), 3262–3280. https://doi.org/10.1108/MD-08-2023-1331

[2] Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

[3] Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review, 61*(4), 66–83. https://doi.org/10.1177/0008125619862257

[4] Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

[5] Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

[6] Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics, 140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

[7] Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., ... & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality*. https://doi.org/10.2139/ssrn.4573321

[8] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

[9] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

[10] Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[11] Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

[12] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286–297. https://doi.org/10.1109/3468.844354

[13] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[14] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300233

[15] Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization, 28*(1), 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

[16] Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

[17] Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*(5), 809–825. https://doi.org/10.1177/0022243719851788

[18] Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human–AI teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[19] Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445492

[20] Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision making: 'Automation bias' and 'selective adherence' to algorithmic advice. *Journal of Public Administration Research and Theory, 33*(1), 153–169. https://doi.org/10.1093/jopart/muac007

[21] Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. https://doi.org/10.48550/arXiv.1702.08608

[22] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[23] Liao, Q. V., Gruen, D., & Miller, S. (2020, April). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376590

[24] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

[25] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the opportunities and risks of foundation models*. https://doi.org/10.48550/arXiv.2108.07258

[26] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

[27] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

[28] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., ... & Koreeda, Y. (2022). Holistic evaluation of language models. https://openreview.net/forum?id=iO4LZibEqW

[29] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*, 111–126. https://doi.org/10.1007/s12599-023-00834-7

[30] Dwivedi, Y. K., Kshetri, N., Hughes, L., et al. (2023). 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[31] Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

[32] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

[33] Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

[34] Leszczyński, G., Gaczek, P., & Munzel, A. (2026). Managerial decision-making and AI: A taxonomy of AI collaboration orientations and strategic integration. *Technovation, 153,* 103519. https://doi.org/10.1016/j.technovation.2026.103519

[35] Luo, Y., Hao, Y., Gong, H., & Li, B. (2026). Applications of human-machine collaborative decision-making: A review of research with recent developments. *iScience, 29*(6), 116056. https://doi.org/10.1016/j.isci.2026.116056

[36] Luong, A., Kumar, N., & Lang, K. R. (2026). Human–machine collaboration and algorithmic decision-making: The role of organizational incentives and AI capabilities. *Information & Management, 63*(6), 104384. https://doi.org/10.1016/j.im.2026.104384