Management Analytics and Social Insights

www.masi.reapress.com

Manag. Anal. Soc. Insights. Vol. 3, No. 4 (2026) 230–243.

Paper Type: Original Article

Decision-Calibrated Human–AI Collaboration: A Governance Framework for Generative and Agentic AI in Management

Sepideh Nafei*

Department of Business Management, National Taipei University of Technology, Taipei, Taiwan; sepideh9251@gmail.com.

Citation:

Received: 06 September 2026 *Revised: 29 October 2026* *Accepted: 01 December 2026*	Nafei, S. (2026). Decision-calibrated human–AI collaboration: A governance framework for generative and agentic AI in management. *Management Analytics and Social Insights, 3*(4), 230-243.

Abstract

Organizations are rapidly moving from predictive Artificial Intelligence (AI) toward Generative Artificial Intelligence (GenAI) and increasingly agentic systems that can recommend, create, coordinate, and sometimes execute managerial actions. Yet higher model capability does not by itself produce better organizational decisions: underreliance wastes useful machine competence, whereas overreliance can amplify hallucination, bias, automation complacency, accountability diffusion, and skill erosion. This conceptual study develops a decision-calibration framework for human–AI collaboration that treats the central design problem as aligning the degree of AI influence with verified AI competence, task risk, decision reversibility, evidentiary verifiability, and human expertise. The study integrates research on organizational decision design, trust in automation, algorithm aversion and appreciation, explainable AI, responsible AI governance, and recent evidence on GenAI-enabled knowledge work. The resulting framework distinguishes four operating modes, human-led, AI-augmented, AI-led with human veto, and dual-control, and links them to a lifecycle governance architecture comprising task qualification, evidence-grounded generation, confidence and uncertainty signaling, independent verification, escalation, audit logging, and post-decision learning. Eight propositions explain how calibration quality should affect decision accuracy, speed, accountability, organizational learning, and resilience. The paper further introduces practical metrics for calibration error, override quality, verification coverage, and decision reversibility. The framework shifts managerial attention from the simplistic question of whether humans or AI should decide toward the more consequential question of when, how, and under what governance conditions each should influence a decision. This perspective offers a theoretically grounded and operationally actionable basis for designing reliable human–AI decision systems in contemporary organizations.

Keywords: Agentic AI, Decision calibration, Generative artificial intelligence, Human–AI collaboration, Organizational governance, Responsible AI.

1 | Introduction

Artificial Intelligence (AI) has moved from a specialized analytical technology to a general-purpose component of organizational decision processes. Earlier managerial applications primarily emphasized

Corresponding Author: sepideh9251@gmail.com
https://doi.org/10.22105/masi.v3i4.110
Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

prediction, classification, optimization, and recommendation; current Generative Artificial Intelligence (GenAI) systems additionally create text, code, plans, explanations, and alternatives, while emerging agentic configurations can decompose goals, use tools, interact with enterprise systems, and execute multi-step workflows. This shift makes the organization of decision rights, not merely model accuracy, a central management problem. Human–AI symbiosis can exploit the computational scale of AI while retaining human contextual judgment [1], but automation and augmentation are not clean substitutes: They form an interdependent organizational tension whose benefits depend on how tasks, authority, and learning are configured [2].

Existing work provides several foundations for this problem. Organizational decision structures can range from full delegation to sequential and aggregated human–AI arrangements [3]. Managing AI also requires explicit attention to autonomy, learning, and inscrutability [4], and human–AI collaboration can be understood as an organization-design problem involving division of labor and learning configurations [5]. At the same time, learning algorithms reshape expertise, occupational boundaries, coordination, and control [6], creating new theoretical questions for organizations [7] and new forms of algorithmic control at work [8]. These contributions establish that the quality of AI-enabled management depends on institutional and organizational design rather than on technical performance alone.

The practical difficulty is that organizations frequently treat trust as the target variable. This is conceptually insufficient. High trust in an unreliable system is dangerous, while low trust in a reliable system is wasteful. Research on levels of automation shows that the assignment of information acquisition, analysis, decision selection, and action implementation can be designed separately [9]. Research on trust in automation similarly emphasizes appropriate reliance rather than maximal reliance [10]. Accordingly, the core managerial objective should be calibration: AI influence should increase when the system is demonstrably competent for the task and when errors are detectable or reversible, and decrease when uncertainty, impact, opacity, or irreversibility rise.

This issue has become more acute with GenAI and agentic AI because capability is uneven across tasks, outputs can appear fluent despite being unsupported, and a single model can shift from adviser to semi-autonomous actor through tool access. Traditional governance approaches that classify an AI application once, at deployment, are therefore too static. A GenAI system may be reliable for summarization but unreliable for causal inference; an agent may be safe when drafting a purchase request but unsafe when authorizing a payment. The unit of governance must move from the model to the decision episode: the specific task, context, evidence base, requested action, and consequences.

The motivation for this paper is thus both theoretical and managerial. Theoretically, research on automation, trust, algorithmic reliance, explainability, governance, and GenAI is rich but fragmented. Managerially, organizations need a parsimonious mechanism for translating those literatures into decision rights, verification duties, escalation rules, and measurable controls. The main research question is: How should organizations allocate and dynamically adjust human and AI influence so that AI-enabled decisions are fast and analytically strong without sacrificing accountability, verification, and organizational learning?

The paper makes four contributions. First, it introduces decision calibration as a governance construct distinct from trust, acceptance, or the level of automation. Decision calibration is defined as the degree of alignment between AI influence in a decision and the AI system's verified, task-specific competence after accounting for risk, reversibility, verifiability, and available human expertise. Second, it integrates behavioral and organizational research into a four-mode decision-rights architecture: Human-led, AI-augmented, AI-led with human veto, and dual-control. Third, it develops an end-to-end governance process that treats evidence, uncertainty, verification, escalation, logging, and post-decision learning as one connected system rather than separate compliance activities. Fourth, it proposes measurable indicators, calibration error, verification coverage, override quality, unsupported-claim rate, and reversible-action share, that can support empirical testing and managerial monitoring.

The paper is conceptual rather than empirical. It does not fabricate organizational data or claim experimental effects that have not been observed. Instead, it synthesizes established findings and recent evidence to derive propositions and a testable framework. Section 2 reviews the relevant literature. Section 3 develops the theoretical foundations and the decision-calibration construct. Section 4 presents the governance architecture and propositions. Section 5 discusses managerial implications and a research agenda. Section 6 concludes with limitations and future research.

2 | Literature Review

2.1 | From Automation to Organizational Human–AI Collaboration

The intellectual roots of human–AI decision design precede contemporary machine learning. Automation research distinguishes stages such as information acquisition, information analysis, decision and action selection, and implementation, allowing different levels of machine involvement at each stage [9]. This decomposition remains valuable for GenAI because a system can be permitted to retrieve information and draft alternatives without being permitted to select or execute an action. The literature on trust in automation further argues that reliance should correspond to the system's capabilities and the task environment, not to generalized positive attitudes [10]. Longitudinal trust is shaped by prior experience, system performance, user disposition, and situational factors [11]. Empirical work also demonstrates that trust affects whether people rely on automated advice [12], [13].

In organizational research, the central shift has been from 'technology use' to joint decision production. Shrestha et al. [3] identify delegation, sequential hybrid arrangements, and aggregated human–AI decision structures.

Puranam [5] shows that division of labor and learning configurations are organization-design choices, while Raisch and Krakowski [2] caution that automation and augmentation are mutually entangled rather than mutually exclusive. The implication is that the same organization may need different combinations of AI and human authority across tasks and over time. This is particularly important when AI changes the knowledge distribution itself: learning algorithms may alter who is considered expert, which activities remain visible, and how decisions are monitored [6], [8].

2.2 | Algorithm Aversion, Appreciation, and the Problem of Miscalibrated Reliance

Behavioral research demonstrates that reliance on algorithms is neither uniformly low nor uniformly high. Dietvorst et al. [14] show algorithm aversion after people observe an algorithm make errors. In contrast, Logg et al. [15] find algorithm appreciation in several judgment contexts, with people sometimes weighting algorithmic advice more than human advice. Giving users even limited control over algorithmic forecasts can reduce aversion [16], and task characteristics matter: algorithms are trusted less for tasks perceived as subjective than for tasks perceived as objective [17]. A systematic review by Burton et al. [18] organizes causes and remedies around expectations and expertise, autonomy, incentives, cognitive compatibility, and divergent rationalities.

These findings imply that adoption is a poor proxy for decision quality. A manager can overrule a superior algorithm because of aversion, or accept a weak output because of automation bias or the persuasive fluency of a generated explanation. The managerial problem is therefore bidirectional: Preventing both disuse and misuse. Calibration requires the organization to identify when the human is likely to be wrong about the AI, not only when the AI is likely to be wrong about the task.

2.3 | Explainability, Cognitive Effort, and Complementary Performance

Explainability is frequently proposed as a remedy for inappropriate reliance, but evidence is more nuanced. Alufaisan et al. [19] found that AI predictions could improve human decision accuracy in their experimental

setting, whereas adding explanations did not yield conclusive additional gains. Bansal et al. [20] similarly show that explanations can increase acceptance of AI recommendations without necessarily improving complementary team performance. Buçinca et al. [21] demonstrate that cognitive forcing functions can reduce overreliance, but at a user-experience cost. These studies suggest that explanation should not be treated as a universal trust-building mechanism; it is a decision-control intervention whose value depends on whether it helps users discriminate correct from incorrect AI advice.

Human–AI interaction guidelines emphasize expectation setting, feedback, correction, and appropriate control across the interaction lifecycle [22]. The human-centered AI perspective likewise argues that high levels of automation can coexist with high levels of meaningful human control [23]. Together, these literatures support a governance design in which users are not merely shown an explanation; they are given the evidence, uncertainty cues, checkpoints, and authority needed to verify consequential outputs.

2.4 | Responsible AI Governance and the Operationalization Gap

Responsible AI research has achieved substantial convergence on broad principles while remaining less settled on operationalization. A global review of AI ethics guidelines identifies recurring principles such as transparency, justice and fairness, non-maleficence, responsibility, and privacy [24]. The AI4People framework similarly articulates beneficence, non-maleficence, autonomy, justice, and explicability [25]. Yet principles alone cannot guarantee ethical AI because implementation requires institutional mechanisms, professional norms, incentives, and accountability processes [26].

Reviews of responsible-AI tools show a continuing challenge in translating high-level principles into concrete methods and practices [27]. End-to-end internal algorithmic auditing is one response because it embeds documentation and accountability throughout the system lifecycle [28]. The National Institute of Standards and Technology (NIST) AI risk management framework provides a complementary governance vocabulary organized around govern, map, measure, and manage activities [29].

Recent information-systems scholarship reinforces this operational focus. Responsible AI governance can be understood through structural, relational, and procedural practices across the AI lifecycle [30]. The remaining gap is to connect governance practices directly to decision rights and reliance behavior. A policy may require fairness testing and audit logs while leaving unanswered a basic operational question: When a specific AI recommendation is uncertain, who is required to verify it, who may overrule whom, and what evidence is necessary before execution? Decision calibration addresses this missing link.

2.5 | Generative AI, Agentic Systems, and Changing Knowledge Work

The diffusion of GenAI raises the stakes of calibration because evidence shows both substantial productivity gains and task-dependent limitations. Noy and Zhang [31] report that access to a generative AI tool improved speed and quality in professional writing tasks.

Brynjolfsson et al. [32] find productivity gains from generative AI assistance in customer support, with particularly strong benefits for less experienced workers. A large field experiment on product innovation shows that individuals with AI can match the performance of human teams without AI and that AI can help bridge functional knowledge silos [33].

However, survey evidence among knowledge workers indicates that higher confidence in GenAI is associated with lower reported critical-thinking effort, while task stewardship increasingly shifts toward verification and integration [34].

These results imply a central paradox. GenAI can democratize expertise and compress coordination costs, yet the same convenience can reduce the cognitive effort applied to verification. In team settings, AI is also evolving from tool to teammate [35]. As organizations move further toward agentic AI, the locus of risk changes from generated content to generated action. Decision calibration must therefore govern not only whether a recommendation is believed, but whether an AI system is permitted to act.

Table 1. Synthesis of literature streams and the unresolved governance problem.

Literature Stream	Primary Construct	Established Insight	Limitation for GenAI/Agentic AI	Calibration Implication
Automation and trust	Appropriate reliance	Reliance should match automation capability and context [9–13].	Often assumes relatively stable functions and bounded automation.	Calibrate influence at the decision-episode level.
Algorithm reliance	Aversion/appreciation	Users may underuse or overuse algorithms depending on errors, control, and task perception [14–18].	Adoption does not indicate decision quality.	Measure both disuse and misuse.
Explainable AI	Interpretability and acceptance	Explanations do not automatically improve complementary performance [19–21].	Fluent GenAI explanations can themselves persuade.	Prioritize discriminative evidence and verification.
Human-centered AI	Control and interaction design	Automation and meaningful human control can coexist [22], [23].	Control is often described at system rather than decision level.	Assign explicit veto, approval, and escalation rights.
Responsible AI	Principles, auditing, risk governance	Lifecycle governance and auditing operationalize ethical principles [24–29], [36].	Often weakly linked to day-to-day decision rights.	Bind governance controls to action authority.
GenAI and AI teamwork	Productivity, expertise integration	GenAI can improve performance and reduce expertise gaps [31–35].	Capability is jagged; critical-thinking effort may fall.	Use task-specific competence tests and verification duties.

3 | Theoretical Foundations and the Decision-Calibration Framework

3.1 | Decision Calibration as a Distinct Governance Construct

Decision calibration is the alignment between the influence granted to AI in a specific decision and the system's verified competence for that decision, conditional on the decision's risk, reversibility, verifiability, and the availability of competent human oversight. This definition differs from trust calibration in two ways. First, the unit of analysis is not only the user's psychological trust; it is the organizationally designed influence that the AI actually exerts. Second, the target is not a subjective belief alone but a governance outcome: whether decision rights, verification effort, and execution authority are proportionate to demonstrated capability and consequence.

Let I denote AI influence on a bounded scale from 0 (no decision influence) to 1 (full autonomous authority). Let C denote verified task-specific AI competence on a comparable scale. A simple descriptive calibration error can be written as $\mathrm{CE} = |\mathrm{I} - \mathrm{C}|$. This quantity is not a complete risk measure because the same mismatch matters more in a high-impact irreversible decision than in a low-impact reversible one.

Accordingly, organizations should use a Risk-weighted Calibration Error (RCE), $\mathrm{RCE} = \mathrm{W} \times |\mathrm{I} - \mathrm{C}|$, where W is a context weight derived from impact severity, irreversibility, uncertainty, and the weakness of independent verification. The purpose of these expressions is managerial rather than predictive: they make explicit that excessive AI authority and excessive human exclusion are governance choices that can be monitored.

Competence must be task-specific and evidence-based. A single benchmark score for a foundation model is insufficient because managerial tasks combine factual retrieval, interpretation, synthesis, causal reasoning, forecasting, stakeholder judgment, and execution. The relevant question is whether the configured system, model, prompt, retrieval layer, tools, data, and safeguards perform adequately on the class of decision episodes it is being asked to influence. Competence should therefore be estimated from validated historical cases, shadow-mode testing, adversarial tests, and post-deployment monitoring rather than vendor claims or user impressions.

Influence also has multiple components. An AI system can influence a decision by framing the problem, selecting evidence, generating alternatives, ranking alternatives, recommending an action, drafting a justification, or executing the selected action. Organizations should avoid collapsing these into one binary variable called 'AI use.' A system may legitimately have high influence over information search but low influence over execution, especially when the decision is consequential or difficult to reverse.

3.2|Five Contingencies that Determine Appropriate AI Influence

Verified task competence

The stronger and more stable the system's validated performance on the relevant task distribution, the greater the potential role for AI. Competence should include not only average accuracy but failure modes, calibration of uncertainty, robustness to distribution shift, and the rate of unsupported claims.

Decision consequence

High-impact decisions increase the cost of both model error and human overreliance. Consequence includes financial loss, safety, legal exposure, employee or customer harm, reputational damage, and strategic lock-in.

Reversibility

A decision that can be cheaply inspected and reversed can tolerate more machine autonomy than an irreversible action. Drafting, simulation, and sandboxed execution are therefore structurally safer than direct external commitment.

Verifiability

AI influence should be higher when outputs can be independently checked against authoritative data, deterministic rules, or observable outcomes. It should be lower when claims are difficult to validate, causality is contested, or tacit social context dominates.

Human oversight quality

Human involvement is protective only when the reviewer has relevant expertise, sufficient time, independence, and meaningful authority. A nominal 'human in the loop' who routinely clicks approve is not a governance control.

Fig. 1 summarizes the proposed framework. AI influence is not set by model capability alone; it is the output of a calibration decision that combines competence with contextual risk and oversight conditions. The framework treats governance controls as moderators that can increase the safe operating envelope of AI.

For example, retrieval from authoritative sources, independent verification, and reversible execution can justify greater AI influence without assuming that the underlying model has become intrinsically more trustworthy.

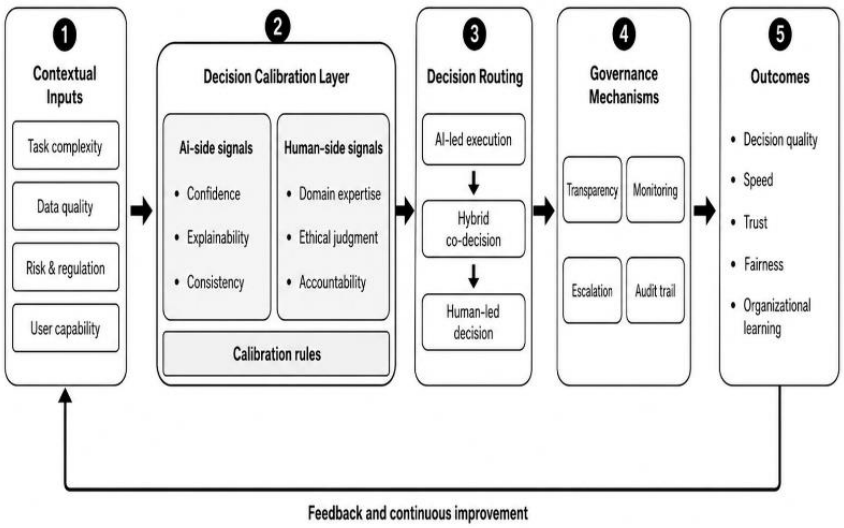


Fig. 1. Decision-calibration framework for allocating AI influence in managerial decisions.

3.3 | Four Operating Modes

The framework identifies four operating modes. In the human-led mode, AI may retrieve, summarize, or simulate, but the human frames the decision and selects the action. This mode is appropriate when task competence is uncertain, contextual, or moral judgment is dominant, or consequences are difficult to reverse. In the AI-augmented mode, AI provides recommendations or alternatives and the human makes the final choice after specified verification. This is likely to be the default mode for many managerial decisions because it combines machine scale with human accountability.

In the AI-led with human-veto mode, the system selects or executes an action unless a human intervenes. This mode is appropriate only when competence is well validated, monitoring is continuous, and actions are reversible or tightly bounded. Examples include routine replenishment within limits, classification with post-hoc review, or workflow routing. Finally, dual-control requires both AI and an independent human (or two independent controls) to authorize an action. It is appropriate for high-impact or high-uncertainty decisions where neither fully autonomous AI nor unstructured human discretion is acceptable. Dual-control is not simply 'more oversight'; it is a designed separation of proposal, verification, and authorization roles.

4 | A Governance Architecture for Calibrated Human–AI Decisions

4.1 | Lifecycle Stages

Stage 1 (task qualification). Define the decision class, its consequence level, reversibility, evidence requirements, excluded uses, and minimum human expertise. No model should be granted authority before the task itself is qualified.

Stage 2 (evidence-grounded generation). Constrain AI to approved information sources when factual accuracy matters. Record provenance for retrieved evidence and distinguish retrieved facts from model-generated interpretation.

Stage 3 (uncertainty and limitation signaling). Require the system interface to disclose uncertainty, missing evidence, conflicts between sources, and conditions outside the validated task envelope. Confidence should not be inferred from linguistic fluency.

Stage 4 (independent verification). Specify what must be checked, by whom, and against what source. Verification coverage should rise with consequence and irreversibility. Critical outputs should be checked independently rather than by asking the same model to self-confirm.

Stage 5 (decision-right allocation and escalation). Route the episode to one of the four operating modes. Escalate when confidence falls, evidence conflicts, the task is novel, policy limits are reached, or the proposed action crosses an impact threshold.

Stage 6 (execution safeguards). Use sandboxing, transaction limits, staged commitments, human approval gates, and reversible actions to prevent a reasoning error from becoming an uncontrolled real-world action.

Stage 7 (audit logging and post-decision learning). Log prompts, evidence, recommendations, overrides, approvals, tool calls, and outcomes. Use these records to update task competence estimates and detect drift, recurring human override failures, or emerging automation bias.

The architecture is depicted in *Fig. 2* as a closed-loop system. The feedback loop is essential: governance should update when either the model changes or the organization learns that users are interacting with it differently than expected. A model upgrade can invalidate earlier competence evidence; similarly, experienced users may become more efficient verifiers over time, or conversely may become complacent.

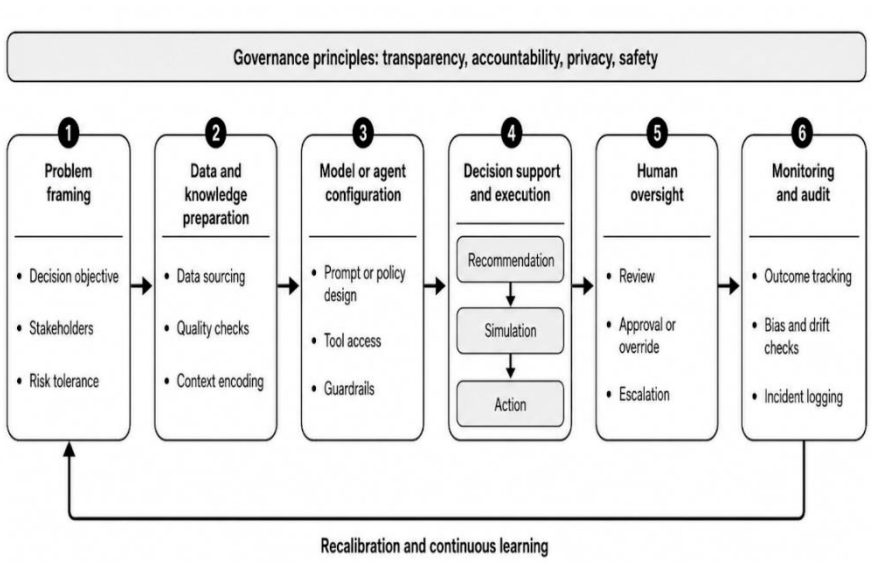


Fig. 2. Closed-loop governance architecture for calibrated GenAI and agentic-AI decisions.

4.2 | Propositions

Proposition 1. Decision performance will be higher when AI influence is calibrated to verified task-specific competence than when influence is determined by generalized trust, user preference, or a fixed organization-wide automation policy.

Proposition 2. The negative effect of AI competence errors on decision outcomes will increase as decision consequence and irreversibility increase.

Proposition 3. Independent verifiability will increase the safe level of AI influence because it lowers the probability that an unsupported output propagates into action.

Proposition 4. Explanation features will improve human–AI decision performance only to the extent that they increase users' ability to discriminate correct from incorrect AI recommendations rather than merely increasing acceptance.

Proposition 5. Human oversight will reduce AI-related error only when reviewers possess relevant expertise, adequate time, independence, and genuine authority to reject or escalate the AI output.

Proposition 6. Reversible execution and bounded tool permissions will permit higher levels of agentic AI autonomy without proportionally increasing organizational risk.

Proposition 7. Organizations that record and analyze both human overrides and non-overrides will achieve better calibration over time than organizations that monitor only model accuracy.

Proposition 8. Excessive reliance on GenAI will reduce critical verification effort and weaken organizational learning unless workflows explicitly preserve problem formulation, evidence checking, and post-decision reflection as human responsibilities.

4.3 | Operational Metrics

A framework becomes managerial only when it can be monitored. *Table 2* proposes a compact measurement system. None of the metrics should be interpreted in isolation. For example, a low human override rate can indicate either excellent AI performance or severe automation complacency. The metrics must therefore be read together with verified task performance and decision outcomes.

Table 2. Suggested metrics for decision-calibrated AI governance.

Metric	Definition	Interpretation	Illustrative Managerial Use
RCE	Context-weighted gap between granted AI influence and verified competence.	Higher values indicate governance mismatch.	Compare business units or decision classes.
Verification coverage	Share of required AI claims/actions independently checked.	Low coverage is dangerous in high-impact contexts.	Audit adherence to verification policy.
Unsupported-claim rate	Share of material AI claims lacking authoritative evidence.	Signals grounding or hallucination risk.	Track by model, prompt, and task class.
Override precision	Share of human overrides that improve the final outcome.	Separates useful oversight from reflexive aversion.	Train reviewers and refine escalation rules.
Missed-override rate	Share of harmful AI recommendations that humans failed to stop.	Signals overreliance or weak review.	Redesign interface and cognitive forcing.
Reversible-action share	Share of agentic actions that can be automatically rolled back or safely staged.	Higher values expand safe autonomy.	Prioritize reversible workflow engineering.
Outcome learning latency	Time from observed outcome to governance/model update.	Long latency slows organizational adaptation.	Improve monitoring and feedback ownership.

5 | Discussion and Implications

5.1 | Theoretical Implications

The framework reframes several established debates. First, it moves beyond the automation-versus-augmentation dichotomy by treating both as variable configurations of influence. The relevant theoretical object is not whether AI replaces or supports humans, but how problem framing, evidence selection, recommendation, authorization, and execution rights are distributed. This is consistent with the organizational perspective that AI changes structures of decision making [3–5] while extending it through a measurable calibration construct.

Second, the framework distinguishes trust from governance. Trust research remains essential for predicting reliance [10–13], but an organization cannot safely rely on psychological calibration alone. Users differ in expertise, incentives, risk tolerance, and exposure to automation. Organizational controls must therefore be

designed so that an overconfident user cannot grant unlimited authority to a weak system and an algorithm-averse user cannot silently discard superior evidence without accountability. Calibration becomes a property of the socio-technical system, not only of individual cognition.

Third, the framework clarifies the role of explainability. The goal of explanation should be error discrimination and contestability, not generalized persuasion. This interpretation is consistent with evidence that explanations may increase acceptance without improving complementary performance [19–21]. In management contexts, useful explanations should expose decision-relevant evidence, assumptions, uncertainty, and counterfactual sensitivity in a form that permits a qualified reviewer to challenge the recommendation.

Fourth, the framework links responsible AI to everyday decision operations. Ethical principles and lifecycle governance are necessary [24–29], [36] but organizations ultimately experience AI risk through specific decisions and actions. By tying governance controls to decision rights, the framework offers a bridge between policy-level responsible AI and operational management.

5.2|Managerial Implications

For executives, the immediate implication is to stop treating 'human in the loop' as a sufficient control. A human presence is meaningful only when the role is defined: What must be reviewed? What evidence must be checked? What decisions can be vetoed? How much time is allocated? What happens when the reviewer disagrees with the AI? Organizations should replace generic human-in-the-loop policies with explicit decision-right matrices.

For AI product owners and analytics teams, performance evaluation should move from model-centric dashboards to decision-centric dashboards. Average benchmark accuracy should be supplemented with unsupported-claim rates, error severity, distribution-shift indicators, human override precision, and downstream outcome measures. This is especially important for GenAI systems because linguistic quality can remain high even when factual reliability deteriorates.

For internal audit and compliance, AI logs should record the evidence path and the authority path. Evidence path means what information the system used and which claims were verified. Authority path means who proposed, approved, modified, escalated, and executed the action. Together these records allow an organization to reconstruct not just what the model said, but why the socio-technical system allowed the decision to occur.

For workforce design, GenAI should be deployed in ways that preserve learning. Productivity gains are valuable [31–33], but work should not be redesigned so that junior employees only approve outputs they no longer know how to independently produce or evaluate. The evidence that GenAI use can shift critical-thinking effort toward verification and integration [34] suggests a need for deliberate skill architecture: organizations should preserve opportunities for problem formulation, first-principles reasoning, and independent judgment while using AI to accelerate search, drafting, simulation, and routine synthesis.

Fig. 3 provides a practical routing matrix. As consequence and irreversibility increase, governance should move upward from autonomous or AI-led modes toward dual-control and human-led modes unless verifiability and competence are exceptionally strong. Conversely, low-impact and easily reversible tasks can support more autonomous agentic execution, provided monitoring remains active.

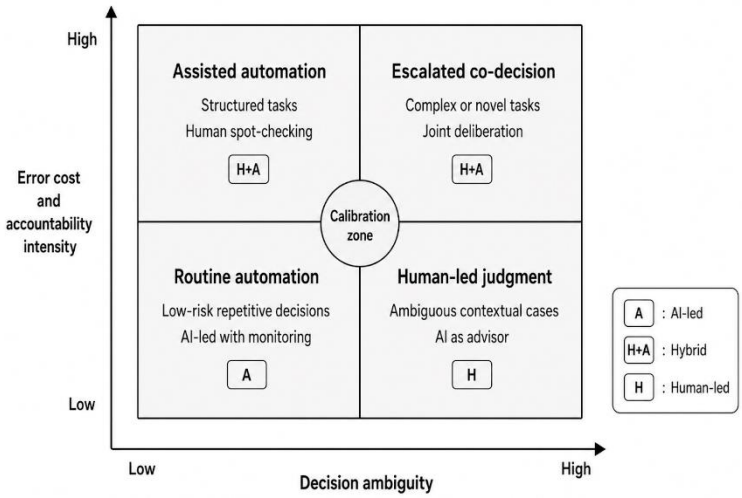


Fig. 3. Decision-routing matrix linking risk, verifiability, and recommended human–AI operating mode.

5.3 | Research Agenda

The framework generates a research agenda across individual, team, organizational, and institutional levels. At the individual level, experiments should compare generalized trust interventions with calibration-oriented interventions such as mandatory evidence checks, uncertainty displays, delayed reveal of AI advice, or justification-before-advice designs. The dependent variables should include complementary accuracy and error discrimination, not only trust or acceptance.

At the team level, research should examine whether AI reduces or redistributes functional diversity. Recent field evidence indicates that AI can bridge expertise silos [33], but it remains unclear whether prolonged use homogenizes reasoning or suppresses minority expertise. Studies could compare teams in which AI is a shared resource with teams in which members use independent agents, testing effects on information pooling, dissent, and coordination.

At the organizational level, longitudinal studies should test whether calibration metrics predict operational incidents, productivity, innovation, and learning. Natural experiments created by staged AI rollouts could compare business units using fixed automation rules with units using dynamic routing based on consequence, reversibility, and verified competence. Researchers should also examine governance drift: whether approval rates rise over time as users become accustomed to AI, even when model reliability does not change.

At the institutional level, future research should study how regulation, professional liability, standards, and procurement rules shape calibration choices. High-level governance principles are increasingly common, but organizations may operationalize them differently depending on sector, country, and risk exposure. Comparative research can identify when structural, relational, and procedural governance practices [30] actually change the influence granted to AI in real decisions.

Finally, agentic AI requires research methods that capture action chains rather than isolated prompts. A consequential failure may emerge from a sequence of individually plausible steps: an agent retrieves incomplete data, forms an incorrect plan, invokes a tool, receives a misleading response, and commits an action. Evaluation should therefore measure chain-level reliability, intervention opportunities, rollback effectiveness, and the degree to which human supervisors can reconstruct and contest the agent's reasoning and actions.

6 | Conclusion

AI-enabled management is entering a phase in which the most important design question is no longer whether an organization uses AI, but how much influence AI should have over particular decisions and actions. This paper developed a decision-calibration framework to answer that question. The framework defines calibration

as alignment between AI influence and verified task competence after accounting for consequence, reversibility, verifiability, and human oversight quality. It integrates behavioral evidence on algorithm aversion and overreliance with organization design, human-centered AI, responsible AI governance, and emerging evidence on GenAI-enabled work.

The framework contributes four practical ideas: a task-specific calibration construct, four operating modes for allocating decision rights, a closed-loop governance architecture, and measurable indicators that reveal both underuse and overuse. Its core principle is simple but consequential: organizations should not maximize AI autonomy, human control, or user trust. They should maximize the fit between capability, context, and authority. GenAI and agentic AI can then be used aggressively where evidence, competence, and reversibility support it, while high-impact and weakly verifiable decisions remain subject to stronger independent judgment and dual control.

This study has limitations. It is a conceptual synthesis and therefore does not estimate causal effects or validate the proposed metrics across industries. The constructs of competence, influence, and consequence will require domain-specific operationalization, and the proposed routing rules may interact with organizational culture, professional expertise, regulation, and technical architecture. Future work should empirically test the propositions using laboratory experiments, field experiments, incident datasets, and longitudinal organizational deployments; develop validated scales for decision-level AI influence and calibration; examine how calibration changes as users and models co-adapt; and extend the framework to multi-agent systems in which several AI agents and several humans jointly produce and execute decisions.

Acknowledgments

The author acknowledges no external assistance or funding for the conceptual development of this manuscript at this stage.

Author Contribution

All authors contributed significantly to this research. All authors have read and approved the submitted version of the manuscript.

Funding

This research received no external funding.

Data Availability

No new empirical data were generated or analyzed in this conceptual study. All literature cited in the manuscript is identified in the reference list.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] 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

[2] 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

[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] Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

[5] Puranam, P. (2021). Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design, 10*(2), 75–80. https://doi.org/10.1007/s41469-021-00095-2

[6] 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

[7] von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries, 4*(4), 404–409. https://doi.org/10.5465/amd.2018.0084

[8] 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

[9] 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

[10] 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

[11] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

[12] 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

[13] Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[14] 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

[15] 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

[16] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[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] Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

[19] Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(8), 6618–6626. https://doi.org/10.1609/aaai.v35i8.16819

[20] Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 81*, 1–16. https://doi.org/10.1145/3411764.3445717

[21] 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, 5*(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

[22] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human–AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 3*, 1–13. https://doi.org/10.1145/3290605.3300233

[23] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[24] 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

[25] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

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

[27] Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

[28] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

[29] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce*. https://doi.org/10.6028/NIST.AI.100-1

[30] Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

[31] 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

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

[33] Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., Mollick, L., ... & Lakhani, K. R. (2026). The cybernetic teammate: A field experiment on generative AI and teamwork. *Organization Science*, *37*(4), 1217-1242. https://doi.org/10.1287/orsc.2025.20702

[34] Lee, H. P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025, April). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-22). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713778

[35] Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G. J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), 103. https://doi.org/10.1016/j.im.2019.103174