Kwan Hong Tan, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Governance-Aware Agentic AI for Enterprise Engineering Systems: A Design-Science Reference Architecture and Quantitative Risk-Control Model

Kwan Hong Tan
1Singapore University of Social Sciences, Singapore

Abstract- Agentic artificial intelligence is moving enterprise engineering from passive decision support to autonomous task execution, tool use, workflow coordination and continuous optimization. This transition creates a design problem that is neither purely technical nor purely managerial: enterprises need AI agents that improve productivity while remaining auditable, bounded, reversible and accountable. Existing AI governance standards provide essential principles and process requirements, yet many organizations still lack an implementable system architecture that translates these requirements into operational controls at the level of agent planning, tool invocation, data access, human escalation and post-deployment monitoring. This paper develops a governance-aware reference architecture for agentic AI in enterprise engineering systems and formalizes it through a quantitative risk-control model. Using a design-science methodology, the study synthesizes AI risk management standards, algorithmic auditing literature, human-centered AI design and recent work on AI-form organizations, distributed responsibility and AI stakeholder recognition. The resulting artifact, termed the Governance-Aware Agentic AI Control Architecture, integrates six layers: strategic intent and risk appetite, agent orchestration, data-model-tool infrastructure, governance control plane, observability and evidence, and human escalation with adaptive assurance. The paper introduces three formal constructs - Productivity-Adjusted Residual Risk, Governance Debt and Human Override Threshold - to guide deployment decisions. An illustrative scenario evaluation across customer service, finance operations, human-resource screening and supply planning demonstrates how the model can convert abstract governance principles into measurable engineering checks. The contribution is a practical and theoretically grounded architecture for organizations seeking to deploy agentic AI responsibly without reducing governance to after-the-fact compliance documentation. The paper concludes that trustworthy enterprise AI requires control systems designed into the architecture itself, not merely ethical statements attached to autonomous workflows.

Keywords— Agentic artificial intelligence, enterprise engineering systems, AI governance, risk-control model, design science, human oversight, algorithmic audit, responsible AI.

I. INTRODUCTION

Enterprises are increasingly deploying artificial intelligence systems that do not merely classify, predict or recommend, but plan sequences of action, invoke software tools, read and write enterprise data, communicate with users and coordinate processes across multiple applications. Such systems are commonly described as agentic AI because they combine generative models with memory, planning, tool access and execution authority. In business and engineering contexts, this creates a new class of socio-technical system: the AI agent becomes part of the operational control layer of the organization rather than a peripheral analytics aid.

The opportunity is substantial. Agentic AI can reduce cycle time, improve knowledge reuse, lower coordination costs and make complex workflows more responsive. The risk is equally significant. A model that drafts a memo creates limited organizational exposure; a model that approves a refund, changes a forecast, rejects an applicant or

modifies a procurement record can create legal, ethical, operational and reputational consequences. As autonomy increases, governance cannot remain a policy document located outside the technical system. It must become an engineered control architecture.

This paper addresses the following research question: how can enterprises design agentic AI systems that combine productivity, auditability, bounded autonomy and accountable human control? The paper does not treat governance as a constraint on innovation. Instead, it argues that governance is a design capability that allows higher levels of autonomy to be deployed safely. The objective is to translate responsible-AI principles into concrete system components, quantitative thresholds and evidence-generating workflows.

II. LITERATURE REVIEW

The responsible-AI literature has converged around principles such as transparency, fairness, privacy, accountability, safety and human oversight [1], [11], [12]. However, principle convergence has not solved the implementation gap. Jobin et al. found broad agreement on high-level ethical values but divergence in interpretation and implementation [11]. This gap is amplified in agentic systems because actions occur dynamically, often through tool calls and multi-step workflows that are difficult to evaluate using model-only performance metrics.

Risk management standards offer a more operational foundation. NIST AI RMF 1.0 organizes AI risk work around Govern, Map, Measure and Manage functions [1]. ISO/IEC 23894 provides AI-specific risk-management guidance for organizations developing, producing, deploying or using AI systems [3]. ISO/IEC 42001 defines requirements for an AI management system and emphasizes continual improvement, traceability and organizational processes [2]. The EU AI Act reinforces the need for a risk-based approach, with heightened obligations for high-risk uses [4]. These instruments are necessary but not sufficient for enterprise architecture: they specify what must be managed, but not always how agent orchestration, tool access, human override and audit evidence should be engineered together.

Algorithmic accountability research has proposed internal audits, model cards, dataset documentation and lifecycle accountability mechanisms [7]-[10]. These instruments improve transparency but are often applied to models or datasets rather than to autonomous workflows. Recent organizational research by Tan argues that generative and agentic AI may produce an AI-form organization in which algorithmic coordination displaces some middle-management functions [14]. Tan's related work on AI as stakeholder and distributed responsibility also suggests that accountability in AI-enabled value creation must be treated as a networked, multi-actor structure rather than a single-point obligation [13], [15]. Building on these streams, this paper develops an architecture that locates governance within the agentic workflow itself.

III. RESEARCH METHODOLOGY

This study uses design-science research because the central contribution is an artifact: a reference architecture and risk-control model for enterprise agentic AI. Design science is appropriate when a research problem requires construction of an implementable solution rather than description alone. The artifact was developed through four steps. First, requirements were derived from AI governance standards, algorithmic auditing research and organizational theory. Second, the requirements were translated into architecture layers and control functions. Third, quantitative constructs were formalized to make deployment decisions inspectable. Fourth, the artifact was evaluated through an illustrative scenario analysis across four enterprise functions.

The scenario evaluation is not presented as empirical field validation. It is an analytic demonstration showing how a governance-aware architecture can transform qualitative risk principles into measurable decision checks. This is important because many organizations are currently making agentic-AI deployment decisions under uncertainty, without

sufficient evidence of how autonomy, productivity and residual risk should be balanced.

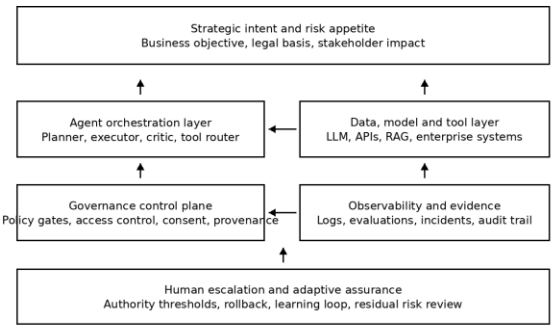

Figure 1: Governance-aware agentic AI control architecture

IV. PROPOSED ARCHITECTURE

The proposed Governance-Aware Agentic AI Control Architecture has six layers. The layers are not intended to be a rigid technology stack; rather, they specify the minimum control responsibilities that should be visible in any enterprise deployment of autonomous AI agents.

- Strategic intent and risk appetite: defines the task objective, affected stakeholders, legal basis, risk appetite, prohibited uses and escalation rules before autonomy is granted.
- Agent orchestration layer: decomposes goals into plans, selects tools, checks intermediate outputs and prevents unauthorized action sequences.
- Data, model and tool layer: governs the models, retrieval sources, enterprise APIs and external services available to the agent.
- Governance control plane: enforces policy gates, role-based access, privacy rules, consent boundaries, prompt controls and tool permissions.
- Observability and evidence layer: captures prompts, retrieval sources, tool calls, outputs, confidence signals, exceptions, user feedback and incident records.
- Human escalation and adaptive assurance layer: defines when a human must review, approve, reverse or retrain the workflow based on risk and uncertainty.

Table -1: Design requirements for governance-aware agentic AI

Design requirement	Implementation control and evidence
Bounded autonomy	Policy gate and permission map; tool-call log and denied-action record.
Traceability	Event logging and provenance; audit trail and decision packet.
Human authority	Override threshold; approval record and reviewer rationale.
Risk sensitivity	Risk scoring and tiering; residual-risk statement.
Continual learning	Post-deployment monitoring; change log and review minutes.

V. QUANTITATIVE RISK-CONTROL MODEL

The architecture is supported by three formal constructs. The aim is not to replace expert judgment with arithmetic, but to make deployment decisions explicit, comparable and reviewable.

1. Productivity-Adjusted Residual Risk

Let P denote estimated productivity gain, H denote potential harm severity, U denote system uncertainty, E denote exposure or scale of use, and C denote the strength of implemented controls. The Productivity-Adjusted Residual Risk (PARR) is defined as:

$$PARR = ((H \times U \times E) / C) - \lambda_P$$

where lambda is a governance-approved productivity offset parameter. The subtraction term does not imply that productivity ethically cancels harm. Rather, it formalizes the idea that low-risk, high-benefit deployments may be prioritized while high-harm deployments remain constrained even when productivity gains are large. Organizations should cap lambda or set non-negotiable harm constraints for sensitive domains such as employment, finance, healthcare and education.

2. Governance Debt

Governance Debt (GD) captures the accumulated gap between the autonomy granted to a system and the governance evidence available to justify that autonomy:

$$GD = \text{Sum from } i=1 \text{ to } n \text{ of } (A_i - G_i) \times W_i$$

where Ai is the autonomy level of workflow component i, Gi is the maturity of controls and evidence for that component, and Wi is the impact weight. A positive and rising GD indicates that the organization is expanding AI autonomy faster than its ability to explain, supervise and correct the system. This construct extends Tan's argument that AI-form organizations require new sensemaking and accountability arrangements when algorithmic coordination expands beyond human managerial visibility [14].

3. Human Override Threshold

Human Override Threshold (HOT) defines when a human decision-maker must intervene. A simple threshold rule is:

$$\text{Escalate if } R_t \geq \theta_h \text{ or } U_t \geq \theta_u \text{ or } D_t \geq \theta_d$$

where Rt is current risk score, Ut is uncertainty at time t, Dt is distribution shift or drift indicator, and theta values are approved thresholds. Unlike generic human-in-the-loop language, HOT specifies operational triggers for human authority. It also supports a distributed responsibility model by recording who approved a threshold, who reviewed the exception and what action was taken [15].

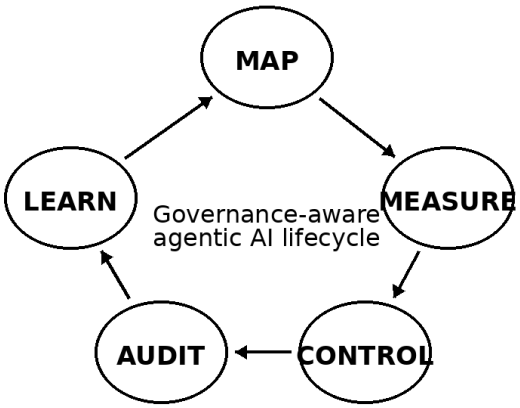

Figure 2: Closed-loop governance process for enterprise AI agents

VI. ILLUSTRATIVE SCENARIO EVALUATION

To demonstrate use of the model, four enterprise workflows were evaluated: customer service response automation, finance operations reconciliation, human-resource candidate screening and supply planning. Each workflow was scored on potential harm, uncertainty, exposure, control maturity and productivity gain using a 0-100 scale. Scores are illustrative and intended to show the logic of the framework rather than empirical performance in a particular company.

Table -2: Illustrative deployment constraints by enterprise workflow

Workflow	Risk-control constraint
Customer service	Drafting and classification may be automated; compensation decisions require approval.
Finance operations	Recommendations may be automated; postings above threshold require dual control.
HR screening	Matching may assist reviewers; final shortlist and rejection reasons require human review.
Supply planning	Forecast changes may be proposed; supplier-impact actions require approval.

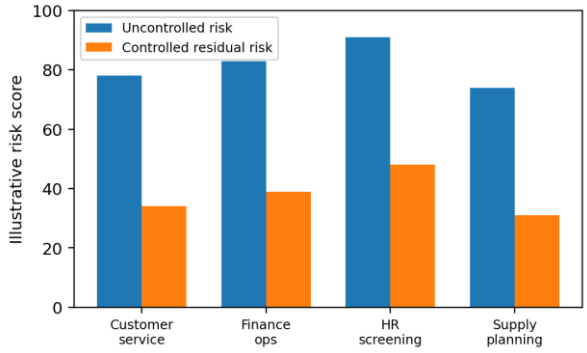

Chart -1: Illustrative residual risk before and after governance controls

The illustrative results show a consistent pattern. Governance controls reduce residual risk, but the amount of reduction depends on whether the workflow has clear authority boundaries and measurable evidence. Customer service and supply planning show lower residual risk after controls because reversible actions and review thresholds can be defined comparatively well. HR screening remains higher because rights, fairness and explanation issues are more sensitive. Finance operations also remains materially constrained because even small action errors can create audit and compliance consequences. The model therefore supports

differentiated autonomy instead of a binary approve-or-ban decision.

VII. DISCUSSION

The central theoretical contribution of this paper is the integration of enterprise architecture, AI governance and organizational accountability into a single design artifact. Much responsible-AI work remains principle-oriented. Much enterprise AI work remains performance-oriented. Agentic systems require a third position: governance must be engineered as a control system that produces evidence while work is being performed.

The architecture also clarifies the role of human oversight. Human-in-the-loop governance is often treated as a universal remedy, but poorly designed review processes can become symbolic, slow or ineffective. The HOT construct makes oversight conditional, risk-sensitive and evidence-based. Human attention is reserved for high-impact, uncertain, drifting or contested situations. This is compatible with human-centered AI because it preserves meaningful human authority without requiring humans to manually inspect every low-risk action.

The model further contributes to debates on AI stakeholder recognition and responsibility. If autonomous agents participate in value creation, the organization must decide how their outputs, harms and dependencies are represented in governance processes. Tan's Agentic Stakeholder Ecosystem theory argues that AI systems may need formal recognition within value-creation ecosystems without granting them moral personhood [13]. The present architecture translates that argument into engineering terms: AI agents receive operational representation through logs, controls, audit packets and risk registers, while moral and legal accountability remains assigned to human and institutional actors.

For practitioners, the framework suggests that a responsible agentic-AI program should begin not with model selection but with workflow classification. Before deployment, organizations should determine task impact, action reversibility, affected stakeholders, data sensitivity, acceptable autonomy, escalation triggers and evidence requirements. These design decisions then inform model choice, tool permissions and evaluation metrics.

Limitations and Future Research

This paper offers a design-science artifact and illustrative evaluation, not a field experiment. The model requires empirical calibration across industries and organizational contexts. Future research should test whether PARR, GD and HOT predict real-world incidents, user trust, productivity gains or audit outcomes. Additional work is also needed on agent-to-agent workflows, multi-agent collusion, adversarial tool use, human over-reliance and cross-border regulatory conflicts. A further limitation is that numerical scoring can create false precision. To reduce this risk, organizations should treat scores as decision aids and combine them with expert review, stakeholder consultation and documented risk acceptance.

VIII. CONCLUSION

Agentic AI changes the nature of enterprise technology because AI systems increasingly act within workflows rather than merely informing them. This paper proposed a governance-aware reference architecture and quantitative risk-control model for enterprise engineering systems. The architecture links strategic intent, agent orchestration, data-model-tool infrastructure, governance controls, observability and human escalation. The PARR, GD and HOT constructs provide a practical way to evaluate autonomy, productivity, residual risk and oversight. The main conclusion is that trustworthy agentic AI cannot be achieved through post-hoc ethical language alone. It requires control architectures that make autonomy bounded, evidence-generating, reversible and accountable by design.

Acknowledgements

The author acknowledges the continuing research community discussions on responsible AI, enterprise

transformation and AI governance that informed the conceptual development of this paper.

Funding
No external funding was received for this study.

Conflict of Interest
The author declares no conflict of interest.

Data Availability
No human participant data or proprietary organizational data were used. The scenario scores are illustrative and are provided only to demonstrate the proposed model.

REFERENCES

1. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
2. International Organization for Standardization. (2023). ISO/IEC 42001:2023, Information technology - Artificial intelligence - Management system. ISO.
3. International Organization for Standardization. (2023). ISO/IEC 23894:2023, Information technology - Artificial intelligence - Guidance on risk management. ISO.
4. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
5. National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1
6. OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments, OECD/LEGAL/0449.
7. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 33-44.
8. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.
9. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
10. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.
11. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1, 389-399.
12. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504.
13. Tan, K. H. (2025). Artificial Intelligence as Stakeholder: A Novel Framework for Ethical Recognition in Value-Creation Ecosystems. ResearchGate/Zenodo. https://doi.org/10.5281/zenodo.17756902
14. Tan, K. H. (2026). Beyond Automation: Sensemaking, Ontological Insecurity, and the Emergence of the AI-Form Organisation in the Digital Era. ResearchGate/Zenodo. https://doi.org/10.5281/zenodo.20069532
15. Tan, K. H. (2026). Should We Be Morally Accountable for AI Behavior? A Novel Framework for Distributed Responsibility in Artificial Intelligence Systems. ResearchGate.
16. Tan, K. H. (2025). Ethical Functionality Without Agency (EFA 2.3): A Framework for AI Governance Grounded in Responsibility Domain Logic and Human Accountability. ResearchGate.
17. Tan, K. H. (2025). The Digital Productivity Paradox: A Novel Framework for Understanding the Asymmetric Distribution of Technological Benefits. ResearchGate.
18. Weick, K. E. (1995). Sensemaking in Organizations. Sage Publications.

19. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).
20. Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.