Australian Journal of Science and Technology

ISSN Number (2208-6404)
Volume 10; Issue 2; June 2026

Original Article

The role of humans-in-the-loop in hybrid automation systems based on generative artificial intelligence

Nitin Garg*

Amazon Web Services, Seattle, Washington, USA

ABSTRACT

The article examines how human-in-the-loop mechanisms shape reliability and controllability in hybrid automation systems built on generative artificial intelligence (AI) within enterprise workflows. The practical motivation is anchored in a specific claimed contribution, Quick Automate, treated here as an applied reference point for designing decision-control gates in agentic workflow automation. Scientific novelty is associated with an integrated analytical view that connects agent orchestration architectures, escalation and approval governance, and evaluation criteria for human–AI collaboration under operational constraints. The work formulates a design-oriented interpretation of human participation as a structured decision-control layer that can be configured, measured, and audited. The analysis relies on peer-reviewed studies and technical reports on large language model alignment, multi-agent systems, enterprise modeling, and human–AI teaming. The results identify how to operationalize review gates, exception routing, and post hoc audits in workflow products such as quick automate to reduce high-impact errors while preserving throughput. The conclusion outlines implementation implications for product teams building agentic workflow automation with accountable human oversight.

Keywords: Agentic artificial intelligence, decision approval, enterprise workflow automation, generative artificial intelligence, governance gates, human-in-the-loop, hybrid automation, Quick automate, risk control, workflow orchestration

Submitted: 04-05-2026, **Accepted:** 06-06-2026, **Published:** 30-06-2026

INTRODUCTION

Large enterprises maintain cross-functional processes in which execution quality depends on policy interpretation, exception handling, and a traceable decision rationale. Generative artificial intelligence (AI) systems-large language models (LLMs)-extend automation beyond rigid rule-based workflows by generating plans, synthesizing text artifacts, and invoking external tools, thereby increasing process coverage but creating uncertainty tied to hallucinations, brittle reasoning, and output variability.[1,2] In this setting, human-in-the-loop participation is not reducible to "manual checking"; it becomes a designed control circuit that defines when autonomy is permitted, when escalation occurs, and how accountability is preserved across handoffs between automated agents and business owners. The problem becomes sharper as organizations move from single-assistant deployments to orchestrated collections of specialized agents that coordinate tasks across systems, documents, and approvals, because error propagation and responsibility diffusion rise with orchestration depth.[3,4]

In this article, the discussion is tied to an applied enterprise reference point – Quick Automate – considered a workflow automation product in which agentic components draft actions, route work items, and coordinate multi-step execution, while humans retain explicit decision rights at defined control gates. The aim is to use the contribution as a concrete anchor for analyzing how approval routing, exception handling, and auditability can be designed in hybrid automation systems that operate under enterprise accountability requirements.

The tasks of the study are:

1) To systematize human participation patterns across the lifecycle of agentic workflow automation;
2) To relate escalation logic to technical mechanisms of agent planning, tool use, and alignment;

Address for correspondence: Nitin Garg, Amazon Web Services, Seattle, Washington, USA. E-mail: garg.ng@gmail.com

3) To derive implementation-oriented implications for enterprise product teams, focusing on governance, evaluation, and operational metrics.

The derived implications are presented in a form suitable for workflow products such as Quick Automate, where the central engineering question is the placement and configuration of human control gates that govern irreversible actions, policy-sensitive decisions, and exception resolution.

MATERIALS AND METHODS

The analytical base relies on peer-reviewed, widely cited works on human–AI collaboration and control in systems that integrate large foundation models. Vats *et al*. surveyed human–AI teaming with large pre-trained models, offering a cross-domain map of collaboration modes and risks.[1] Krakowski analyzed human–AI agency at the organizational level in the era of generative AI, motivating governance and human-centered value distribution.[2] Fourney *et al*. proposed a generalist multi-agent architecture with an explicit Orchestrator, clarifying how coordination and error recovery are implemented in multi-agent systems.[3] Wu *et al*. introduced AutoGen as a multi-agent conversation framework that explicitly supports human inputs and tool-use coordination.[4] Ferguson *et al*. examined how verbal explanations affect user perceptions in collaborative decision-making, supporting the need for explanation quality in review steps.[5] Mosqueira-Rey *et al*. provided a consolidated taxonomy of human-in-the-loop machine learning interactions, enabling structured treatment of control placement and human effort allocation.[6] Nast *et al*. evaluated LLM use in enterprise modeling and requirements engineering, demonstrating practical limits of substituting domain experts and motivating "human + LLM" operating modes.[7] Natarajan *et al*. contrasted human-in-the-loop with AI-in-the-loop paradigms, supporting a more precise separation between automation-first and human-first decision regimes.[8] Ouyang *et al*. documented reinforcement learning from human feedback (RLHF) for instruction-following behavior, grounding the alignment side of human contribution.[9] Yao *et al*. introduced ReAct as an interleaving of reasoning and acting, formalizing agent tool-use patterns relevant to workflow automation.[10]

The article applies analytical synthesis of recent literature, comparative conceptual analysis of human control placement across system architectures, and design-oriented generalization from empirical findings reported in the selected sources. The reasoning follows a structured mapping between the technical components of agentic systems (planning, tool calls, orchestration, alignment), the control mechanisms (confidence gating, guardrails, approvals, audits), and the organizational requirements (accountability, operational scalability, preservation of domain expertise). In addition, the article uses Quick Automate as an applied reference point to illustrate how the proposed control-gate logic maps to a workflow automation product.

RESULTS

Hybrid automation systems based on generative AI can be represented as socio-technical pipelines in which an LLM (or a team of LLM-based agents) performs interpretation, planning, and text generation. At the same time, enterprise tools execute state changes via APIs, ticketing actions, approvals, or data updates. In contrast to conventional automation, the generative component does not merely select among predefined actions; it constructs intermediate artifacts – summaries, justifications, extracted requirements, drafted communications – that become operational inputs for downstream steps. This property increases the value of automation in knowledge-intensive processes, yet it introduces a failure class where plausible language masks invalid assumptions. Consequently, human participation becomes structurally tied to decision validity, not only to error correction after deployment. Findings across human–AI teaming and HITL-ML research support viewing human involvement as a resource that is scheduled and routed, with explicit policies specifying when human attention is consumed and when automation proceeds autonomously.[1,6]

A first result concerns the placement of humans-in-the-loop along the workflow timeline. In agentic pipelines, human control points cluster into three functional categories: (i) Pre-action review before irreversible tool calls, (ii) exception handling when the agent detects low confidence or conflicts, and (iii) *post hoc* audit for sampling-based quality control and continuous improvement. The literature indicates that these placements correspond to distinct technical needs: Pre-action review requires intelligible rationales and traceability; exception handling requires robust detection of uncertainty and policy violations; *post hoc* audit connects to feedback mechanisms that correct systematic errors and inform alignment updates. The separation matters because "human oversight" without specifying temporal placement leads to either excessive friction (humans approve everything) or fragile autonomy (humans are consulted too late).[5,6,8]

A second result concerns orchestration depth and escalation design. Multi-agent frameworks show that a coordinating entity – explicitly implemented as an Orchestrator – can allocate tasks to specialized agents and re-plan when failures occur. This architecture clarifies where human escalation can be inserted: either at the orchestrator level (blocking the entire plan) or at the tool boundary (approving a specific action). The distinction is operationally consequential: orchestrator-level escalation reduces the risk of cascading errors across agents but increases latency; tool-boundary escalation preserves throughput but requires precise guardrails and granular risk

classification. Empirical discussions of multi-agent systems emphasize that evaluation and containment become harder when agents take actions with side effects, which further supports human checkpoints at risk-bearing transitions rather than at every conversational turn.[3,4,10]

Quick Automate can be interpreted as an enterprise workflow automation setting where agentic components generate intermediate artifacts (summaries, classifications, proposed next steps) and may propose tool calls that trigger state changes in enterprise systems. In such a product, the practical meaning of "human-in-the-loop" is expressed through explicit control gates: Pre-action approval for irreversible or policy-sensitive actions, exception routing when the agent detects ambiguity, missing mandatory fields, or rule conflicts, and *post hoc* audit for sampled quality control and governance reporting. The orchestrator-level versus tool-boundary escalation distinction becomes operational in this setting: Blocking an entire plan at the orchestration layer protects against cascading errors in multi-step workflows, while fine-grained tool-boundary approvals preserve throughput but require calibrated risk classification and stable explanation payloads for reviewers. This applied framing clarifies why human control gates are a design surface rather than an *ad hoc* safety overlay: Quick Automate-like products succeed when review is selective, evidence-bearing, and routable, not universal and manual.

In enterprise operations, Quick Automate-like workflows commonly concentrate risk at a small number of transitions. Examples include:

1) Drafting and routing customer-impacting communications that require a human owner before delivery;
2) Proposing changes in ticketing or case-management systems where misclassification can trigger policy violations;
3) Generating access or configuration requests that must be approved before execution;
4) Coordinating multi-step remediation actions where a single incorrect tool call can propagate side effects across dependent services.

These scenarios underscore a design implication derived from the literature: Review gates should be placed at risk-bearing transitions, while low-stakes steps remain autonomous under monitoring and audit sampling.

A third result links human-in-the-loop design to alignment and feedback economics. RLHF research demonstrates that systematic human preference data can shift model behavior toward instruction following and reduced undesirable outputs, yet it does not eliminate runtime uncertainty in specific enterprise domains. The implication for hybrid automation is a two-layer human contribution model: offline human feedback shapes the agent's general behavior. In contrast, online human intervention governs domain-specific exceptions and high-stakes actions. Treating these layers as substitutes is analytically incorrect: Offline alignment reduces baseline error rates, while online human control manages residual risk under domain constraints and changing policies.[6,9]

A fourth result concerns enterprise process knowledge and the non-substitutability of domain expertise. Evidence from enterprise modeling and requirements engineering shows that LLM outputs can assist domain experts but do not reliably replace them in constructing "as-is" models or capturing nuanced constraints. For hybrid automation systems, this implies that humans-in-the-loop are not only validators but also sources of tacit constraints that are difficult to encode fully in prompts or static knowledge bases. In practical product deployment, this supports designs where human reviewers do not merely accept/reject outputs, but actively inject missing constraints, correct latent assumptions, and update process artifacts that the agent will reuse (templates, reusable prompts, structured schemas).[7]

A fifth result addresses explainability as an enabling condition for efficient review rather than as a compliance ornament. Research on verbal explanations in collaborative decision-making indicates that explanation characteristics influence perceived trustworthiness and acceptance. In human-in-the-loop automation, review quality depends on whether the human can reconstruct why the agent proposed a step and whether the explanation supports rapid detection of misalignment with policy. Hence, explanation generation becomes part of the control loop: it is evaluated not only for clarity but for its ability to reveal assumptions and expose uncertainty, enabling humans to intervene selectively rather than universally.[5]

A sixth result clarifies conceptual ambiguity between "automation-first with human intervention" and "human-first with AI assistance." The human-in-the-loop versus AI-in-the-loop distinction formalizes that some enterprise use cases require humans to own the decision, with AI providing drafting and analysis. In contrast, other use cases permit autonomous execution with human escalation only for anomalies [Figure 1].

This distinction maps naturally to enterprise governance: Regulated or reputation-sensitive actions often fall into human-first regimes, while low-stakes repetitive steps can shift to automation-first regimes with monitoring. Designing hybrid automation without declaring which regime applies leads to misaligned expectations and inconsistent accountability.

DISCUSSION

The analytical results support a product-oriented interpretation of humans-in-the-loop as a configurable governance layer that trades off throughput, error cost, and accountability.

Anchoring the discussion in Quick Automate strengthens the interpretability of the results as a contribution-driven design argument. In workflow automation products, stakeholder acceptance and deployability are shaped less by the sophistication of the agent's reasoning and more by governance ergonomics: who approves which actions, how exceptions are routed, what evidence supports review, and how traces enable audit and accountability. The proposed mapping (technical components → control mechanisms → organizational requirements), therefore, functions as a blueprint for implementing Quick Automate-like systems where controllability and responsibility assignment must remain explicit under multi-step agentic execution. Multi-agent orchestration frameworks strengthen this interpretation by creating explicit routing points (orchestrator planning, tool boundaries, error recovery) where escalation logic can be implemented with measurable effects on operational performance.[3,4]

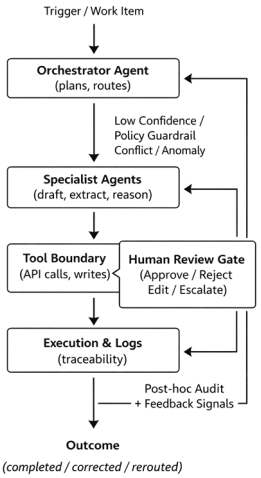

Figure 1: Human-in-the-loop control gates in an agentic hybrid automation workflow (design mapping for workflow products such as Quick Automate; conceptual synthesis based on orchestration and tool-use frameworks).[3,4,10]

Table 1 operationalizes the discussion by aligning intervention patterns with technical mechanisms and enterprise consequences. The mapping follows HITL-ML taxonomies, human–AI teaming surveys, and descriptions of agentic orchestration, treating "human involvement" as a set of design choices rather than a single checkbox.[1,4,6,10]

Table 1 implies that an enterprise "agentic automation program" benefits from explicitly declaring which parts of a workflow operate under automation-first regimes and which remain human-first. This aligns with the human-in-the-loop versus AI-in-the-loop distinction and reduces stakeholder ambiguity in responsible, accountable, consulted, informed -like operational ownership: Humans own outcomes where accountability or regulation dominates, while agents own execution where reversibility and monitoring dominate.[2,8]

Table 1: Human-in-the-loop intervention patterns in generative-AI hybrid automation (synthesized from the literature)[1,4,6,8,10]

Intervention pattern	Typical trigger in enterprise workflows	Supporting technical mechanism	Expected enterprise effect
Pre-action approval at the tool boundary	Irreversible updates (payments, customer-impacting communications, compliance-sensitive changes)	Tool-call gating with rationale/ explanation payloads; orchestrated routing	Reduced high-impact error probability; increased cycle time with bounded risk
Exception handling for anomalies	Conflicting policy signals, ambiguous inputs, and missing mandatory fields	Uncertainty/guardrail detection and escalation routing	Concentration of human effort on edge cases; improved accountability trace
Human-first decision with AI assistance (AI-in-the-loop regime)	High-stakes judgment tasks (interpretation of policy, sensitive approvals)	Human-owned decision flow with AI drafting/analysis support	Straightforward responsibility assignment; controlled adoption in regulated domains
Post-hoc audit and sampling	High-volume, low-stakes automations where full approval is infeasible	Logged traces+audit sampling; feedback capture for updates	Sustainable scaling; continuous quality improvement without full friction
Offline human feedback for behavior shaping	Systematic failure modes across tasks; misalignment with user intent	Preference data for alignment (RLHF)	Lower baseline error rate; reduced manual load over time, without eliminating runtime checks

RLHF: Reinforcement learning from human feedback, AI: Artificial intelligence

Table 2: Enterprise product metrics for managing human-in-the-loop in agentic workflow automation (literature-grounded operationalization)[1,3,4,5,7,8]

Metric	Definition in hybrid automation terms	Why is it diagnostic for HIL design
Escalation rate	Share of work items routed to humans per policy/ uncertainty triggers	Quantifies human load and sensitivity of guardrails
Irreversible-action gate hit rate	Share of workflow steps classified as irreversible and therefore routed through pre-action approval	Quantifies how much of Quick Automate-like execution remains under strict human decision rights; supports calibration of risk taxonomy
Override rate	Fraction of escalated items where humans reject or materially edit agent output	Signals mismatch between agent assumptions and enterprise constraints
Time-to-decision	Latency added by review gates relative to autonomous execution	Captures the throughput cost of control placement
Error recovery incidence	Frequency of re-planning/rerouting after failed actions	Reflects orchestration resilience and quality of exception handling
Explanation usefulness (review efficiency proxy)	Human-reported or behavior-inferred adequacy of rationale for approval decisions	Determines whether selective review is feasible without full manual work
Domain model drift indicators	Recurring corrections tied to changing policies or evolving process definitions	Justifies continual human input into reusable artifacts and schemas

Table 2 translates the discussion into measurement and operating cadence, emphasizing metrics that connect directly to human control placement and agentic behavior. The listed metrics are not proposed as a new experimental battery; they are derived as implementation-relevant constructs grounded in evaluation and governance arguments present in the surveyed literature (orchestration, explanation quality, enterprise modeling limits, and human–AI teaming evaluation).

The discussion yields two implementation implications for enterprise product strategy. First, human-in-the-loop design should be treated as a first-class product surface: Escalation policies, explanation formats, and audit interfaces are as central as model choice, because they determine where accountability rests and how operational teams experience the system.[5,8] Second, orchestration-centric architectures (orchestrator + specialized agents) provide a natural substrate for separating low-risk autonomous segments from high-risk gated segments, enabling incremental deployment and measurable reduction of manual effort while maintaining clear ownership.

CONCLUSION

Human-in-the-loop mechanisms in generative-AI hybrid automation function as a decision-control layer that determines where autonomy is permitted, how exceptions are managed, and how accountability is maintained across enterprise workflow handoffs. The synthesis shows that control placement along the workflow timeline separates pre-action approvals, anomaly-based escalation, and *post hoc* auditing, each tied to different technical requirements and operational trade-offs. Agentic orchestration architectures clarify insertion points for escalation and error recovery, supporting governance designs that prevent cascading failures while preserving throughput. Alignment via human feedback reduces baseline undesired behaviors, yet it does not replace runtime control for domain-specific constraints and high-stakes actions; therefore, offline feedback and online intervention address different risk layers. For enterprise product teams, the practical outcome is an implementation logic where regime selection (automation-first vs. human-first), explanation usefulness for reviewers, and measurable escalation/override dynamics jointly determine deployability and business value. Interpreted through the lens of Quick Automate, the presented design logic clarifies how enterprise workflow automation can scale agentic execution while keeping decision rights, traceability, and auditability explicitly enforced through configurable human control gates.

REFERENCES

1. Vats V, Nizam MB, Liu M, Wang Z, Ho R, Prasad MS, *et al*. A Survey on Human-AI Teaming with Large Pre-Trained Models (arXiv:2403.04931); 2024. Available from: https://arxiv.org/abs/2403.04931 [Last accessed on 2026 May].
2. Krakowski S. Human-AI agency in the age of generative AI. Inform Organ 2025;35:100560.
3. Fourney A, Bansal G, Mozannar H, Tan C, Salinas E, Zhu E, *et al*. Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks (arXiv:2411.04468); 2024. Available from: https://arxiv.org/abs/2411.04468 [Last accessed on 2025 Sep].
4. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, *et al*. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation (arXiv:2308.08155); 2023. Available from: https://arxiv.org/abs/2308.08155 [Last accessed on 2025 Nov].
5. Ferguson S, Aoyagui PA, Rizvi R, Kim YH, Kuzminykh A. The explanation that hits home: The characteristics of verbal explanations that affect human perception in subjective decision-

making. Proc ACM Hum Comput Interact 2024;8:517.
6. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: A state of the art. Artif Intell Rev 2023;56:3005-54.
7. Nast B, Görgen L, Müller E, Triller M, Sandkuhl K. Exploring large language models in enterprise modeling. Discov Artif Intell 2025;5:293.
8. Natarajan S, Mathur S, Sidheekh S, Stammer W, Kersting K. Human-in-the-loop or AI-in-the-loop? Automate or collaborate? Proc AAAI Conf Artif Intell 2025;39:28594-600.
9. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, *et al*. Training Language Models to Follow Instructions with Human Feedback. In: Advances in Neural Information Processing Systems (NeurIPS 2022). Available from: https://arxiv.org/abs/2203.02155 [Last accessed on 2024 Feb].
10. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, *et al*. ReAct: Synergizing reasoning and acting in language models. In: The International Conference on Learning Representations (ICLR 2023). Available from: https://arxiv.org/abs/2210.03629 [Last accessed on 2024 Jun].