

Human-in-the-Loop Enterprise AI Architecture

A Comprehensive Enterprise Architecture for Human Approval, Review, Escalation, Override, Auditability, and Safe AI-Assisted Decision-Making

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

Human-in-the-Loop (HITL) architecture is a foundational control pattern for enterprise Artificial Intelligence where AI systems assist with analysis, recommendations, content generation, or task execution while designated people retain authority at critical decision points. As Generative AI and Agentic AI become more autonomous, enterprises require a systematic architecture for determining when human review is mandatory, how work is routed, what evidence reviewers receive, how approvals are recorded, and how overrides and escalations are governed. This paper proposes a Human-in-the-Loop Enterprise AI Architecture integrating AI models, RAG, enterprise data, policy engines, workflow orchestration, identity, approval services, audit trails, security, Responsible AI, and operational monitoring. It defines risk-based review patterns, confidence thresholds, exception routing, separation of duties, and measurable governance controls. The architecture enables organizations to balance automation and human judgment while improving safety, accountability, explainability, compliance, and enterprise trust.

Keywords — Human-in-the-Loop, HITL, Enterprise AI, Human Oversight, Agentic AI, Responsible AI, AI Governance, Workflow, Approval, Explainability, SAP BTP, Decision Intelligence, Zero Trust

I. INTRODUCTION

Human-in-the-Loop architecture combines AI automation with explicit human authority. Instead of allowing an AI system to independently determine all outcomes, the architecture routes selected decisions, exceptions, or high-risk actions to qualified people for review.

This pattern is increasingly important as Generative AI and autonomous agents are connected to enterprise workflows, customer interactions, financial processes, and regulated decisions.

II. BUSINESS PROBLEM

AI can accelerate analysis and reduce manual effort, but probabilistic outputs may be incomplete, biased, hallucinated, or contextually inappropriate. Full automation is therefore unsuitable for many consequential processes.

Ad hoc manual review is also insufficient because it can be inconsistent, slow, poorly documented, and difficult to audit. Enterprises require a standardized architecture for human oversight.

III. ARCHITECTURE OBJECTIVES

The architecture aims to preserve human accountability while gaining automation benefits. It should route work based on risk, uncertainty, policy, and business impact; provide reviewers with relevant evidence; and record decisions for audit.

The framework should also minimize unnecessary review by allowing low-risk, high-confidence cases to proceed automatically when policy permits.

IV. RISK-BASED HUMAN OVERSIGHT

Human involvement should be proportional to consequence, reversibility, uncertainty, and regulatory importance. Low-risk content generation may require post-review, while payments, legal commitments, or safety decisions may require pre-execution approval.

Risk tiers can define mandatory review types, reviewer qualifications, escalation paths, and evidence retention.

V. HUMAN-IN-THE-LOOP PATTERNS

Common patterns include approve-before-act, review-before-release, exception-only review, dual approval, random sampling, confidence-threshold review, and post-action audit.

Different steps in one process may use different patterns depending on risk and operational urgency.

VI. AI DECISION PREPARATION LAYER

AI services prepare recommendations, summaries, classifications, or proposed actions. RAG, Knowledge Graphs, SAP Business Data Cloud, and SAP Datasphere can provide grounded context.

The system should present sources, confidence, assumptions, policy checks, and unresolved uncertainty to the human reviewer rather than only a final recommendation.

VII. WORKFLOW AND ORCHESTRATION LAYER

Workflow services route AI-generated work to the appropriate reviewer based on role, workload, risk, and segregation of duties. Deadlines, reminders, escalation, delegation, and reassignment should be governed.

SAP Build Process Automation or other workflow platforms can coordinate AI tasks with human approvals and enterprise transactions.

VIII. IDENTITY AND REVIEWER AUTHORIZATION

Reviewers should be authenticated and authorized based on role, competence, jurisdiction, and business responsibility. Sensitive decisions may require privileged approval or two-person control.

The architecture should prevent the same identity from generating, approving, and executing a high-risk transaction when separation of duties is required.

IX. EVIDENCE AND EXPLAINABILITY

Review interfaces should provide the evidence needed for an informed decision: source documents, retrieved context, relevant business data, policy results, model confidence, and prior actions.

Explainability should support human judgment rather than create false certainty. Reviewers should

be able to inspect the most important factors and supporting sources.

X. APPROVAL, OVERRIDE, AND ESCALATION

A human should be able to approve, reject, modify, request more information, or escalate. Overrides should require reason codes and, for significant cases, free-text justification.

Escalation paths should handle uncertainty, policy conflicts, unavailable reviewers, and exceptions outside normal authority.

XI. AGENTIC AI AND HUMAN CONTROL

For AI agents, HITL controls should be placed before high-impact tool calls, irreversible actions, or delegation beyond predefined authority. The agent can prepare the action but a policy service should enforce whether human approval is required.

Human approval tokens should be short-lived and scoped to the exact action so approval cannot be reused for unrelated tasks.

XII. SECURITY AND PRIVACY

HITL architecture must protect review queues, sensitive evidence, approval credentials, audit trails, and communication channels. Zero Trust, encryption, least privilege, and secure session management are foundational.

Reviewers should receive only the data required for the decision to minimize privacy exposure.

XIII. AUDITABILITY AND COMPLIANCE

The system should record AI version, prompt or policy version, retrieved evidence, reviewer identity, decision, timestamp, override reason, downstream action, and final outcome.

This creates a defensible chain of accountability for internal audit, regulatory review, incident investigation, and model improvement.

XIV. HUMAN FACTORS AND USABILITY

Poorly designed review interfaces can create automation bias, alert fatigue, rubber-stamping, or excessive cognitive load. Interfaces should clearly distinguish facts, model-generated content, uncertainty, and policy requirements.

Reviewer training should include model limitations, escalation expectations, and the responsibility to challenge AI outputs.

XV. OBSERVABILITY AND PERFORMANCE

Operational metrics should measure review volume, approval rate, rejection rate, override rate, queue time, escalation rate, false positives, missed issues, and reviewer workload.

Monitoring can identify when thresholds are too conservative, when automation is unsafe, or when human review becomes a bottleneck.

XVI. GOVERNANCE AND RESPONSIBLE AI

AI Governance Committees should define which use cases require human oversight, acceptable automation levels, review qualifications, sampling rates, and evidence retention.

Responsible AI principles require meaningful human control rather than nominal approval. Reviewers must have sufficient information, time, and authority to disagree with the AI.

XVII. IMPLEMENTATION METHODOLOGY

Implementation begins by mapping the business process, identifying consequential decisions, defining risk tiers, and selecting appropriate HITL patterns. Pilot testing should measure both AI quality and human review effectiveness.

Reusable approval services, policy engines, reviewer interfaces, audit logging, and observability should then be standardized across AI solutions.

XVIII. INDUSTRY USE CASES

Financial services can apply HITL to credit exceptions, fraud cases, and customer communications. Healthcare can require qualified review of AI-assisted clinical or administrative recommendations.

Manufacturing can use HITL for maintenance decisions and quality exceptions, while enterprise IT can use it for privileged changes, incident remediation, and architecture approvals.

XIX. FUTURE TRENDS

Future HITL systems will use adaptive oversight where review intensity changes dynamically based on risk, model uncertainty, user behavior, and operational context.

Human oversight will increasingly be integrated with agentic workflows, decision intelligence, process mining, and continuous learning while preserving clear accountability boundaries.

XX. CONCLUSION

Human-in-the-Loop architecture provides a practical way to balance AI automation with human authority. The proposed architecture integrates risk-based routing, evidence, workflow, identity, approval, security, auditability, and Responsible AI into a repeatable enterprise pattern.

By designing human oversight as part of the architecture rather than an afterthought, organizations can increase automation while maintaining safety, accountability, compliance, and trust.

REFERENCES

- [1] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
- [2] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
- [3] ISO 31000:2018, Risk management — Guidelines.
- [4] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.
- [5] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
- [6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
- [7] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [9] SAP SE, SAP Business Technology Platform Documentation.
- [10] SAP SE, SAP AI Core and SAP AI Launchpad Documentation.
- [11] SAP SE, SAP Integration Suite Documentation.
- [12] SAP SE, SAP HANA Cloud Documentation.
- [13] SAP SE, SAP Cloud Identity Services Documentation.
- [14] SAP SE, SAP Cloud ALM Documentation.
- [15] SAP SE, SAP Business Data Cloud and SAP Datasphere Documentation.

- [16] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [17] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [18] OECD, Recommendation of the Council on Artificial Intelligence.
- [19] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
- [20] Cloud Native Computing Foundation, Kubernetes Documentation and Cloud Native Security guidance.

AUTHOR BIOGRAPHY

Sanjeeve Kumar Gajadi is an SAP Enterprise Architect and Lead Consultant with extensive experience in enterprise architecture, SAP BTP, data and analytics, integration, cloud architecture, security, governance, and AI-enabled transformation. His professional interests include Human-in-the-Loop AI, Responsible AI, decision intelligence, Agentic AI, SAP BTP, and enterprise governance.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, international standards bodies, AI governance communities, security researchers, and SAP documentation teams whose published guidance informed the concepts synthesized in this paper.