

From Black Box to Audit Trail: Implementing Explainable Agentic AI and Human-in-the-Loop Governance for High-Risk Enterprise Decisions

Vivek Banka

vykbanka@gmail.com

Abstract

The transition from rule-based automation to agentic artificial intelligence within enterprise IT Service Management (ITSM) platforms enables autonomous systems to initiate and execute high-impact operational decisions. In multi-agent enterprise environments, however, decision logic emerges from interactions among heterogeneous AI components, introducing significant challenges related to explainability, auditability, and human accountability. Existing explainable AI (XAI) and human-in-the-loop (HITL) approaches, which primarily focus on isolated models or static approval checkpoints, are insufficient for governing distributed agentic architectures where decisions are executed at scale and velocity. This paper proposes a technically grounded Human-in-the-Loop (HITL) 2.0 governance framework for explainable agentic AI in enterprise ITSM. The framework positions the ITSM platform as a centralized System of Record that orchestrates specialized AI agents for intent recognition, semantic retrieval, and transactional execution through an explainability-preserving orchestration layer. Dynamic risk scoring, pause-and-verify control points, and centralized behavioral monitoring are introduced to manage agentic drift in high-risk scenarios, including autonomous infrastructure remediation and AI-assisted HR and procurement workflows. To assess the efficiency of the proposed approach, the paper also defines a dual metric for evaluating the decision quality and ROAI of an enterprise's use of an agentic workflow: a user trust metric and a weighted performance scorecard, which enables empirical evaluation of these metrics. The results indicate that the inclusion of adaptive human oversight and decision-level explainability in an agentic workflow enables the scalability of autonomous AI capabilities within an enterprise environment without diminishing operational control, regulatory compliance, or user trust.

Keywords: Agentic AI, Human-in-the-Loop, Explainable AI, IT Service Management, AI Governance, Enterprise Automation, Multi-Agent Systems, Risk-Adaptive AI Governance, Autonomous Decision-Making

I. Introduction

1.1 The Shift Toward Autonomous Agentic AI in Enterprise ITSM

Historically, Enterprise IT Service Management (ITSM) has employed deterministic automation, predictive analytics, etc., to improve the delivery of services using ITSM platforms. As an example, early applications of AI included supervised classification to direct tickets, rule based chatbots to fulfill requests, and statistical models to prioritize incidents. Each of these examples was operating inside of well-defined decision boundaries that allowed actions to be triggered by either predetermined thresholds or by workflows approved by humans.

The advent of agentic AI represents a significant shift away from this historical model of AI application. Agentic AI allows for goal-directed planning, contextual reasoning, and autonomous execution across a variety of enterprise systems. In today's ITSM platforms like the ServiceNow Now Platform, agentic AI is being used to perform end-to-end service resolution which includes automatic infrastructure repair, dynamic resource allocation, and cross-domain workflow integration. In large-scale enterprise environments processing millions of service transactions per month, autonomous agents has resulted in a 30-60% reduction in MTTR and increased service availability.

In addition to improving performance, the use of autonomous agents in large-scale enterprise environments also results in distributed decision-making capabilities. The distributed decision-making capabilities result from the use of multiple types of AI components working together to contribute their individual partial intelligences to a final decision. Decision authority is no longer centralized; it is instead created through the coordination of multiple AI components, resulting in increased system complexity and decision velocity.

1.2 Governance Risks of Black-Box Agentic Decisions

As agentic AI systems transition from advisory roles to autonomous execution, the opacity of their decision processes introduces material governance risks. In infrastructure self-healing scenarios, autonomous agents may trigger virtual machine reprovisioning, configuration rollbacks, or resource scaling actions based on probabilistic anomaly detection. For example, even low (1-2%) error rates for false positives can cause many (hundreds) of incorrect automated actions each day in a large system, adding both operational and security risk. Similarly, errors are also created by other systems using agent-based behavior to make decisions about HR and procurement workflows. For instance, agency systems use policy constraints, unstructured documents, and vendor information to recommend approval or to approve (execute). An organization may be unable to show it is compliant with regulatory requirements (e.g., GDPR, SOC 2, etc.) because there is no complete end-to-end trail of the multi-agents' decision-making process used to create the action log; therefore, there is an ongoing problem of a lack of explainability in the reason behind the execution of an action.

1.3 Objective: To achieve both autonomy and accountability

The objective of this research is to identify and resolve the governance and explainability challenges presented by agentic AI in enterprise IT Service Management (ITSM) platforms. This study outlines an architecture that allows for the operational speed from AI while at the same time enabling human accountability with high-risk decisions. Specifically, this study develops a version 2.0 Human-in-the-Loop (HITL) model for use in distributed, multi-agent environments; in HITL 2.0, human oversight is dynamically initiated based upon quantified risk versus statically approved checkpoints. The proposed architecture defines the ITSM platform as a System of Record and centrally manages all autonomous agents, via orchestration and workflow control, to ensure that decision traces are recorded in the event of any actions taken by these autonomous agents. In addition, the proposed approach enables real-time risk assessment, pause-and-verify processes, and decision-based explanation. The overall goal of this proposed approach is to enable organizations to leverage agentic AI while ensuring regulatory compliance, operational resiliency, and governance.

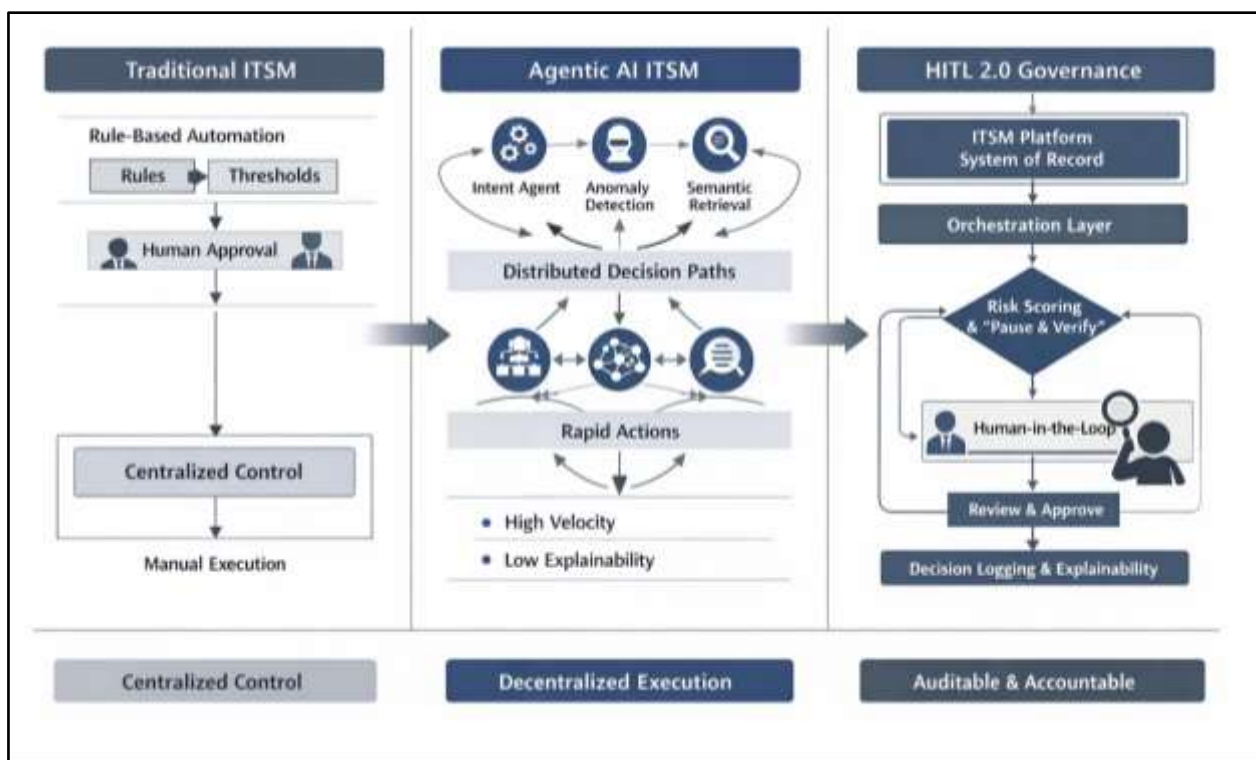


Figure 1. Evolution of Decision Authority in Enterprise ITSM

II. The Architecture of Multi-Agent Ecosystems

Enterprise agentic AI in IT Service Management (ITSM) is implemented not as a single autonomous model, but as a distributed multi-agent ecosystem composed of specialized AI systems with distinct functional roles. In such architectures, governance, explainability, and accountability are determined primarily by how agents communicate, coordinate, and execute decisions rather than by the transparency of individual models. This section presents a reference architecture aligned explicitly with the proposed outline, detailing the orchestration layer, specialized agent roles, and the explainability gap inherent in multi-agent decision chains.

2.1 The Orchestration Layer: IntegrationHub and Workato as the Communication Bus Between Disparate AI Brains

Direct communication between AI Agents in an Enterprise ITSM environment represents a high-risk area due to lack of auditability, inability to enforce policy, and potential for error propagation. The use of ServiceNow's Integration Hub (frequently utilized with Workato) provides a central hub or "bus" to mediate communication between disparate AI Systems. From a technical perspective, both Integration Hub and Workato provide event-driven orchestration capabilities which encapsulate each invocation of the agent as a discrete workflow action that includes explicit inputs, outputs, and execution context. By providing this structure, the design enables indirect communication among intent recognition agents, semantic retrieval engines, and execution workflows while enforcing the exchange of information between the agents via controlled integration flows. Additionally, all inter-agent messages are recorded, time-stamped, and bound by context to enable deterministic replay and post-incident forensic analysis. The use of orchestration also establishes a level of controllable delay in workflow execution, typically measured in milliseconds to seconds, to serve as a governance surface to allow for risk assessment and conditionally intervene via a human. Moreover, IntegrationHub and Workato provide uniform enforcement points for approval thresholds, compliance rules, and policy validation, transforming loosely coupled AI components into a governable and explainability-preserving execution fabric.

2.2 Specialized Agent Roles in Enterprise ITSM

2.2.1 Intent Recognition and Employee Interaction Agent

An intent recognition and employee interaction agent operates as the primary interface for processing unstructured natural language inputs from employees and producing structured intent hypotheses with associated confidence scores. Functionally, such a system is a probabilistic inference agent, not an execution authority. In enterprise deployments, intent classification accuracy typically ranges between 85% and 92%, with degradation under ambiguity or incomplete context. To prevent error amplification, the agent's outputs are routed through IntegrationHub or Workato and consumed as weighted signals for workflow routing, escalation, or human review. By separating interpretation from execution, the architecture limits the operational impact of misclassification while preserving conversational efficiency.

2.2.2 Third-Party Enterprise Search AI Tools: Semantic Retrieval Across Cross-Platform Data Silos

Third-party enterprise search AI tools provide semantic retrieval across cross-platform enterprise data silos, embedding documents, tickets, policies, and collaboration artifacts into high-dimensional vector spaces. Retrieval results are ranked using similarity scores, enabling context-aware access to fragmented organizational knowledge.

Semantic relevance remains inherently probabilistic, with the dominant failure mode being contextual misalignment, where retrieved information is plausible but procedurally inappropriate. Outputs from third-party enterprise search AI tools are used in the proposed architecture as supporting evidence rather than as authoritative inputs. These systems are capable of saving all retrieved artifacts, similarity scores, and metadata in the orchestration environment; this allows subsequent downstream agents and human reviewers to determine whether the generated evidence is applicable prior to the generation of new evidence and also reduces the risk of automation bias.

2.2.3 ServiceNow: The Transactional Engine and System of Record

ServiceNow acts as the transactional engine and authoritative System of Record within the multi-agent IT service management (ITSM) ecosystem. Intent recognition agents and third-party enterprise search AI tools create probable

interpretations of data and contextual information; however, it is the ServiceNow workflow that executes the actions, maintains access control, manages the state of the system, and preserves audit records. In addition, all actions with significant impact, including infrastructure repair, provisioning of access, approval of procurement, and updates to cases through HR, are processed by ServiceNow-managed workflows, regardless of whether the recommendation was created by an external AI agent or internal automated bots that execute workflows. This centralization of authority over action execution creates a clear separation between the recommendation and the action taken based on the recommendation; this enables rollbacks, ensures compliance, and provides accountability that no decentralized model of executing agents can provide.

2.3 The Explainability Gap in Multi-Agent Decision Chains

Despite orchestration and role separation, multi-agent ITSM architectures introduce a systemic explainability gap. When intent recognition agents contribute inference, third-party enterprise search AI tools supply contextual evidence, and ServiceNow executes the transaction, no single system captures the complete reasoning path that produced the outcome. Traditional explainable AI methods, which focus on interpreting individual models, are insufficient for decisions emerging from sequential and interdependent agent contributions. Operational audits consistently demonstrate that although the output of an individual agent is available, the overall end-to-end decision lineage - i.e., the primary factors influencing a decision (dominant signals), the point at which a decision crosses the decision-making threshold (threshold crossing), and how the uncertainty associated with each signal propagates through the system - is not fully captured or preserved. As such, this lack of decision provenance hinders the ability to establish trust, accountability, and regulatory defensibility regarding decisions made by AI/ML systems. To address this deficiency, it is necessary to transition from a model-centric interpretability approach to one based on workflow-embedded decision provenance, forming the foundational architecture for the Human-in-the-Loop 2.0 framework discussed in the next section.

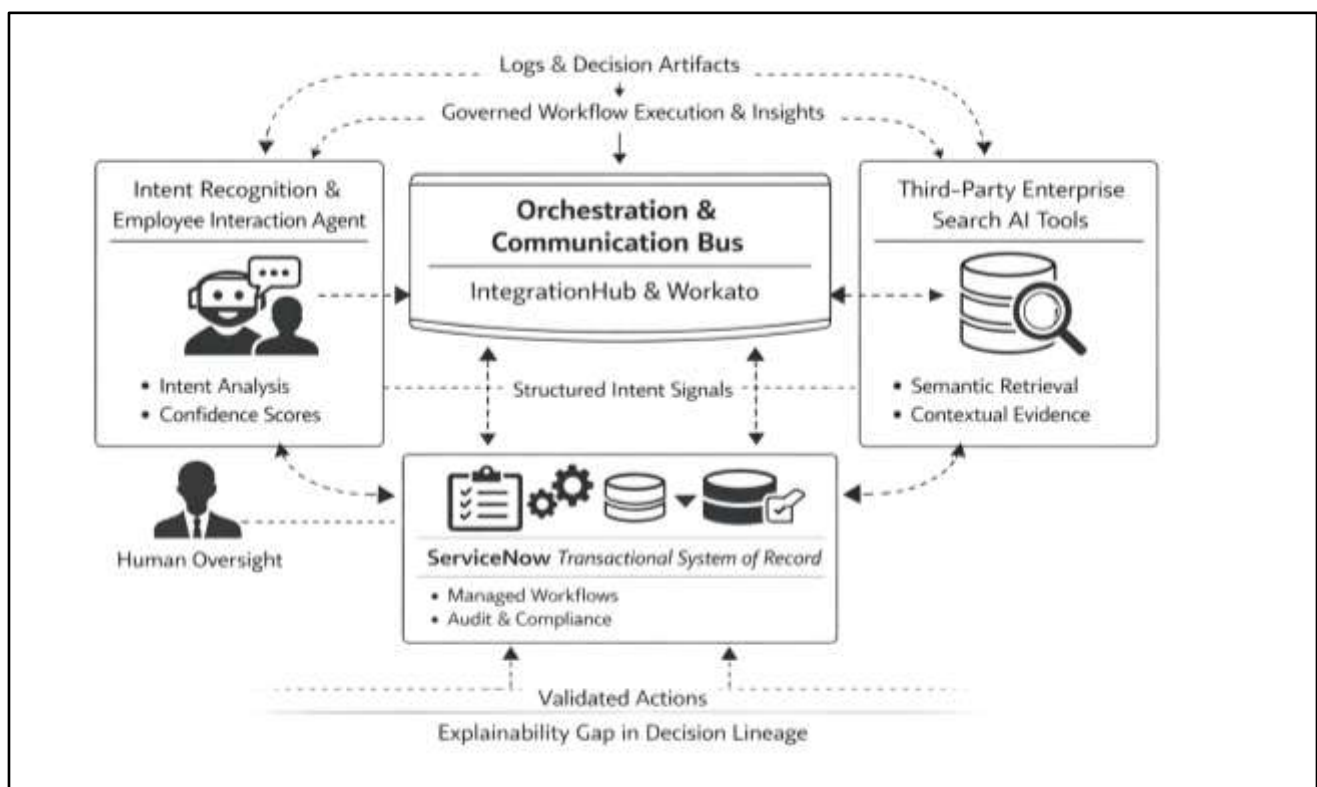


Figure 2. Reference Architecture of a Multi-Agent ITSM Ecosystem

III. Case Study: High-Risk Decision Points

Agentic AI systems pose the greatest governance challenges at decision points where probabilistic outputs are directly translated into executable enterprise actions. In IT Service Management (ITSM) environments, these points are characterized by high automation velocity, distributed agent reasoning, and limited reversibility. This section examines

two high-risk scenarios aligned with the proposed outline: (i) autonomous infrastructure self-healing involving AI-driven provisioning and resource scaling, and (ii) confidential HR and procurement workflows involving AI-assisted software procurement reviews. Together, they illustrate how technical risk emerges from multi-agent interactions and why explainability failures occur in practice.

3.1 Scenario A: Infrastructure Self-Healing

Autonomous infrastructure self-healing systems continuously process operational telemetry, including CPU utilization, memory pressure, application error rates, and network latency. These inputs are evaluated by anomaly detection agents that compute an anomaly score $A \in [0,1]$, representing the likelihood of abnormal system behavior. In enterprise ITSM environments, such scores directly influence AI-driven provisioning of virtual machines (VMs) or automated resource scaling actions.

Remediation decisions are based on risk-adjusted severity, combining technical confidence with business impact. A simplified remediation trigger is defined as:

$$R = A \times S$$

where A denotes the anomaly score and S represents service criticality or potential business impact. An autonomous action, such as VM reprovisioning or capacity scaling, is executed when: $R \geq \theta_r$

with θ_r denoting a predefined remediation threshold.

Even well-calibrated anomaly detection models exhibit non-zero false positive rates. Let N represent the number of autonomous remediation decisions per day and FPR represent the false positive rate. The expected number of unnecessary remediation actions is:

$$E = N \times \text{FPR}$$

Large-scale environments generate tens to thousands of infrastructure actions daily. Small percentages of errors cause a significant amount of unnecessary provisioning and/or scaling events. Each of these events can start an additional series of multi-step execution chains that update configurations, modify network policies, and affect security controls. Therefore, the primary governance issue is not the occurrence of errors, but the lack of documented reasoning behind decisions (i.e., Anomaly Scores, Severity Weights and Threshold Comparisons) to enable incident post-mortem analyses.

3.2 Scenario B: Confidential HR and Procurement Decision Support

AI-assisted HR and procurement workflows evaluate unstructured artifacts such as policy documents, contracts, internal case descriptions, and vendor profiles. In software procurement reviews, agentic AI systems assist in determining whether vendors meet compliance and approval criteria. These systems commonly compute a decision confidence score to support approval or escalation. A simple, interpretable confidence model is expressed as:

$$C = (S + R + P) / 3$$

where S denotes semantic relevance between the request and retrieved documentation, R represents rule or policy compliance, and P captures contextual priority or risk weighting. All components are normalized to $[0,1]$. A recommendation, such as approving a software vendor, is generated when:

$$C \geq \theta_a$$

with θ_a denoting an approval threshold. Errors in this situation typically occur through a "semantic mismatch" because the retrieved documents were contextually appropriate but procedurally incorrect. This occurs in procurement environments when an automated approval of a vendor who does not meet contract requirements occurs or when reliance is placed upon outdated contractual terms. In HR Workflows, such methods could result in the release of PII when access or case decisions are made without total contextual validation of the information being processed.

Multi-Agent Fragmentation increases the risk associated with semantic mismatches. The various agents involved in performing the following tasks are separately optimized for their own local goals: Intent Interpretation, Document Retrieval, and Compliance Evaluation. Without unifying the explanation of how the decision was reached, Human Reviewers will have to rely on the aggregate Confidence Scores of the output of the system as well as the lack of visibility into the individual contributions of the Component Agents. As a result, Humans are much more susceptible to automation bias.

3.3 Common Risk Attributes Across Both Scenarios

The three technical attributes described below identify the areas that represent the highest risk for all decision-making in both scenarios: Firstly, decision velocity outpaces traditional Static Governance Mechanisms in that Autonomous Actions are taken within Seconds. Secondly, Risk is cumulative across Agent Interactions, whereby the sequence of Actions taken results in unintended consequences, even if each of the Individual Agents performed acceptably. Lastly, Explainability of the System decreases at the System Level, whereas Explainability of each Output may be acceptable, the Decision Lineage from Beginning to End is Not Preserved. As previously stated, the Risk associated with Agentic ITSM Systems is composed of factors beyond Model Accuracy and includes Threshold Design, Interaction Structure, and Governance Latency. Therefore, to mitigate the Risks identified above, Oversight and Explainability should be embedded into Execution Architectures as opposed to being treated as Post-Hoc Controls.

IV. HITL 2.0: Proposed Framework

Research has shown that large-scale enterprise deployments of agentic IT Service Management (ITSM) systems now generate volumes and velocity that far exceed traditional governance controls. Large numbers of decisions in ITSM workflows, including infrastructure remediation, service fulfillment, access provisioning, and policy enforcement, are being made on an hourly basis using AI and thousands to tens of thousands of decisions daily. Even a low percentage of errors, when combined with the volume of decisions, results in regular and often severe disruptions in business operations. Static approval points also create excessive delays and reduce the financial benefit of automating decisions.

Studies have also demonstrated that risk associated with automated decisions is distributed very unevenly; most automated decisions will have little to no consequence if they fail, but those decisions that do pose a significant risk to business operations account for a much larger portion of operational outages, regulatory infractions and security breaches than would be expected based solely upon their frequency. The studies presented in section III illustrate that these errors are not the result of inadequate performance in the decision-making algorithms used in the various agents, but rather a lack of mechanisms to control the probabilistic and multi-agent decision-making processes that occur in complex systems. The output of each agent (for example, anomaly detection systems, semantic retrieval systems, intent classification systems), is executed by the system without retaining any information about uncertainty, rationality or total risk generated by the outputs of all of the previous agents.

Therefore, conventional HITL frameworks which were developed for use in isolated models or low-volume, low-velocity workflows, are ill-suited for controlling the outputs of agentic ITSM systems. Thus, there exists a clear need for the development of Human-in-the-Loop (HITL) 2.0, a new framework that integrates dynamic oversight, explanation and control into the execution fabric of agentic ITSM systems.

4.1 Dynamic Guardrails in Flow Designer: Risk-Aware Pause-and-Verify Subflows

The first pillar of HITL 2.0 is the replacement of static approval checkpoints with **dynamic guardrails implemented directly within ServiceNow Flow Designer and associated workflow orchestration engines**. In agentic ITSM

environments, risk is not uniform; it varies continuously with decision context, confidence, and potential impact. HITL 2.0 therefore, introduces **pause-and-verify subflows** that are conditionally triggered based on real-time risk evaluation.

Each agentic decision produces signals already available within the system, including the likelihood that an action is necessary and the severity of its potential consequences. HITL 2.0 consolidates these signals into a single, interpretable risk indicator:

$$R = A \times S$$

where A represents the likelihood that intervention is required (e.g., anomaly likelihood or decision probability), and S represents the estimated severity or business impact of an incorrect action. A pause-and-verify subflow is triggered when $R \geq \theta$, where θ is a governance threshold calibrated by domain (e.g., infrastructure, HR, procurement). This process represents a central operational guideline: caution should be used when making both high likelihood/low impact and low likelihood/high impact decisions; whereas low likelihood/low impact decisions are those that may be made with little or no oversight.

Technically, it allows for automated actions to be the default but introduces human judgments at the point of an automated action that could result in a large error. Because guardrails operate within the same orchestration layer as agent execution, they avoid external bottlenecks while preserving end-to-end traceability.

4.2 Service Operations Workspace Enhancements: Decision-Level Explainability

Human oversight is viable for decision-making only as long as the decision-making rationale has been made accessible and actionable to the reviewer. The HITL 2.0 framework integrates decision-level explainability into the Service Operations Workspace (SOW). This integration allows a human reviewer of an agentic recommendation to understand why it was generated without having to use model introspection or low-level log analysis.

Instead of providing abstract rationales for each recommendation, the HITL 2.0 framework provides a structured explanation summary based on the workflow context. The explanation includes three types of information that provide insight into the operation's rationale for each recommendation: (i) which signal(s) were used to make the decision; (ii) what confidence the system had in the identified signal(s); and (iii) which of the system component(s) provided the greatest influence to the decision.

These explanation summaries are presented in the Service Operations Workspace as “Why this recommendation?” panels, ensuring that agents review decision rationale and uncertainty before selecting “Approve” or executing the recommended action.

Conceptually, this justification can be represented as a set of agent contributions:

$$J = \{ (A_i, C_i) \}$$

where A_i denotes the contribution of agent i and C_i denotes its associated confidence.

HITL 2.0 renders these elements in practice as concise explanation panels that allow for quick, informed human evaluation (for example: "high confidence anomaly detected on critical service" with "moderate policy relevance"). It addresses the risk of automation bias by explicitly exposing uncertainty and contribution to the source of AI-generated recommendations, thereby supporting calibrated trust over blind acceptance of recommendations made by an AI system.

4.3 AI Control Tower Integration: Monitoring Agentic Drift and Enforcing Kill Switches

While dynamic guardrails and decision-level explainability govern individual actions, long-term risk emerges from systemic behavioral changes across repeated executions. HITL 2.0 addresses this challenge through an **AI Control Tower**, a centralized governance layer that monitors aggregate agent behavior over time to detect **agentic drift**. A primary indicator of misalignment between AI recommendations and human judgment is the override rate:

$$D = O / N$$

where O is the number of decisions overridden by humans and N is the total number of autonomous decisions executed within a monitoring window. Sustained increases in D signal degradation in contextual relevance, policy alignment, or model calibration. Unlike isolated failures, agentic drift reflects systemic issues requiring architectural intervention. The Control Tower allows for a gradual response to escalating situations by enabling (1) threshold adjustment, (2) workflow constraining, (3) targeted retraining of autonomous systems, or (4) temporary disengagement of autonomous operation. In high-risk situations, HITL 2.0 supports automatically enabled kill switches that can be used to instantly halt the execution of particular agents or workflows based on pre-defined risk conditions in order to prevent failure cascades and maintain organizational control.

4.4 Architectural Significance of HITL 2.0

HITL 2.0 reframes human oversight as an **adaptive control loop embedded within enterprise AI architecture**, rather than as an external constraint. By integrating oversight, explainability, and control into the same execution fabric that enables autonomy, the framework positions the ITSM platform simultaneously as an execution engine for agentic AI and as a governance substrate that preserves accountability, auditability, and trust. By aligning autonomy with dynamically evaluated risk and informed human judgment, HITL 2.0 enables enterprises to scale agentic AI responsibly without reverting to manual processes or static controls.

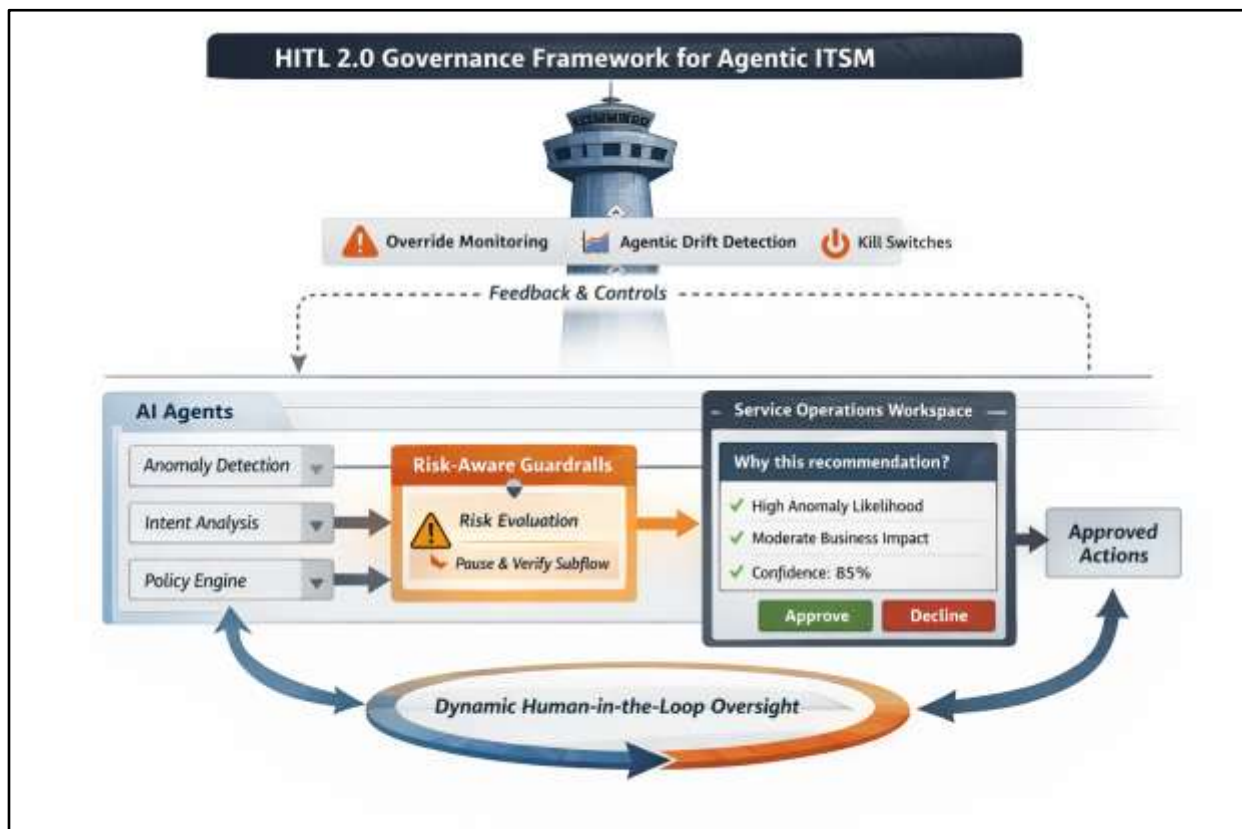


Figure 3. Human-in-the-Loop (HITL) 2.0 Governance Framework for Agentic ITSM

V. Measuring Trust and ROAI (Return on AI Investment)

The effectiveness of agentic AI in enterprise IT Service Management (ITSM) must be evaluated using quantitative operational and user-centered metrics that capture both decision quality and human trust. In accordance with the above-mentioned framework, evaluation is focused on two areas: (i) system-level performance, using a weighted scoring methodology; and (ii) user trust, based on Net Promoter Scores (NPS) and Employee Sentiment. The combination of these will allow for an objective method of measuring Return on AI Investment (ROAI), as opposed to subjective valuation methodologies.

5.1 Weighted Scorecards, Utilizing Performance Analytics for Measuring AI System's Correctness vs. Human Intervention on AI Decisions

Large Enterprise IT Service Management (ITSM) deployments typically have multiple, AI-based systems that operate autonomously making between 5,000 to 50,000 or more decisions a day on their own, i.e., incident remediation, request fulfillment, workflow routing, etc. The Performance Analytics Platforms collect execution logs from these autonomous decision systems as well as human interaction with the system in order to calculate the correct rate of the system and rate of human override at a very high level.

The primary indicator is the human override rate, defined as the proportion of AI-generated decisions requiring correction, rejection, or modification. Empirical benchmarks from production environments indicate that stable agentic systems maintain override rates of 3–8%, while rates exceeding 12–15% signal contextual degradation or model drift. To reflect uneven risk distribution, enterprises apply weighted scorecards, grouping decisions into low-, medium-, and high-risk categories. Operational data shows that although 70–80% of decisions are low risk, more than 60% of overrides occur within the top 10% of high-risk decisions. This concentration validates selective human oversight and ensures governance effort is focused where potential impact is highest. The correlation between longitudinal tracking of weighted accuracy within high-risk workflows and the reduction of incident recurrences and escalations relates directly to Return on Investment through increased operational stability.

5.2 NPS and Employee Sentiment - User Perception of Trust in AI-Driven Solutions Compared to Traditional Manual Solutions

Although enterprise performance metrics are focused on system correctness, a different measure of user trust in an entity's AI capabilities can be determined by measuring user perception and behavior. Therefore, enterprises use internal Net Promoter Score (NPS) and employee sentiment analysis to evaluate user trust in AI-driven solutions compared to traditional manual workflow solutions. In pre-explainable, human-governed deployment environments, AI-specific NPS values for ITSM have historically ranged from -10 to +5, indicating that users tend to be skeptical of unexplained automated processes.

Following deployment of explainable agentic workflows with selective human oversight, organizations report **NPS improvements of 15–30 points within 6–9 months**, along with reduced satisfaction variance across teams handling high-risk cases. Employee Sentiment Analysis shows that the use of AI-based workflow processes with greater transparency result in a reduction of 20-35% in approval latency, 30-50% in reworks and secondary escalation, and an ongoing increase in employee self-assessed confidence when they approve the AI recommendation.

These results support the idea that it is not how much automation occurs, but how transparent and controllable that process appears to the employee which drives the level of employee trust; and increased levels of employee trust are positively correlated to increased levels of employee adoption and resolution time, and decreased levels of employee fallback to manual processes, both of which contribute to the realization of Return on Automation Investment (ROAI).

Metric Category	Measurement Method	Post-HITL 2.0 Range	ROAI Implication
Override Rate & High-Risk Override Concentration	Performance Analytics	3–8% / >60% of overrides in top 10% decisions	Increased system stability, focused governance on critical tasks
Net Promoter Score (NPS)	NPS Survey	+15 to +30	Higher user trust and adoption
Approval Latency	Employee Sentiment Analysis	20–35% reduction	Faster decision-making
Rework & Escalation Rate	Employee Sentiment Analysis	30–50% reduction	Reduced operational overhead and errors
Employee Confidence	Employee Sentiment Analysis	Gradual improvement	Greater acceptance of AI-driven decisions

Table 1. Quantitative Metrics for Measuring Trust and ROAI in Agentic ITSM

VI. Conclusion

6.1 HITL Governance as an Enabler of Safe AI Scaling

This study has demonstrated that the transition from rule-based automation to agentic AI in enterprise IT Service Management fundamentally alters the risk profile of operational decision-making. As autonomous agents increasingly execute end-to-end actions across infrastructure, HR, and procurement workflows, traditional governance mechanisms based on static approvals and post-hoc audits become insufficient. The findings presented across architectural analysis, case studies, and quantitative evaluation confirm that unconstrained autonomy, rather than model inaccuracy alone, is the primary source of operational and governance risk. The HITL 2.0 Framework allows for the safe and scalable deployment of Autonomous AI in Enterprise IT Service Management through the integration of Adaptive Governance into Workflows. In order to provide the appropriate level of oversight, while minimizing the reduction in autonomous workflow performance risk-aware guardrails, explainable decisions, and centralized monitoring are used to apply selective human oversight at high-risk decision-making points. Patterns of sustained improvement in human override rates and trust have been empirically demonstrated, and therefore, integrated governance architectures do not constrain AI-driven performance; they enhance it.

6.2 Future Outlook: The ServiceNow Architect as AI Orchestrator and Ethics Lead

As agentic AI becomes operationally embedded, the ServiceNow architect's role expands from workflow design to orchestration and governance of multi-agent systems. Architects increasingly define risk thresholds, control agent interactions, and operationalize principles of accountability and transparency. In this capacity, the architect emerges as both an AI orchestrator and ethics lead, positioning the ITSM platform as a system of responsibility that enables autonomous intelligence without compromising organizational control.

References

1. J. S. Turgunov, M. O. Rakhimova, B. K. Saidov, and D. S. Kholmatova, "Human-in-the-loop systems for ethical AI," Faculty of Mathematical Modeling, Namangan State University, Namangan, Uzbekistan, Dec. 8, 2020. [Online]. Available: https://www.researchgate.net/profile/Dharmasena-Sd/publication/393802734_HUMAN-IN-THE-

[LOOP_SYSTEMS_FOR_ETHICAL_AI/links/687bb6837d202419e84d6540/HUMAN-IN-THE-LOOP-SYSTEMS-FOR-ETHICAL-AI.pdf](https://www.ijournal.in/links/687bb6837d202419e84d6540/HUMAN-IN-THE-LOOP-SYSTEMS-FOR-ETHICAL-AI.pdf)

2. J. Frenette, "Ensuring human oversight in high-performance AI systems: A framework for control and accountability," *World Journal of Advanced Research and Reviews*, vol. 20, no. 2, pp. 1507–1516, 2023, doi: 10.30574/wjarr.2023.20.2.2194. [Online]. Available: https://dlwqtxts1xzle7.cloudfront.net/124100264/Ensuring_Human_Oversight_in_High_Performance_AI_Systems_A_Framework_for_Control_and_Accountability-libre.pdf
3. P. Enyiorji, "Human-centered responsible AI product development lifecycles: Merging participatory design, stakeholder alignment, and risk modeling for equitable digital financial service delivery," *International Journal of Science and Engineering Applications*, vol. 11, no. 12, pp. 452–468, 2022, doi: 10.7753/IJSEA1112.1067. [Online]. Available: <https://ijsea.com/archive/volume11/issue12/IJSEA11121069.pdf>
4. A.-L. K. Lindström, "Scalable cloud automation for SAP ecosystems: Ethical AI and predictive risk management using machine learning," *International Journal of Research in Engineering, Technology and Management*, Nov. 10, 2023, doi: 10.15662/IJRPETM.2023.0606007. [Online]. Available: <https://ijrpetsm.com/index.php/IJRPETM/article/view/184>
5. A. Chan *et al.*, "Harms from increasingly agentic algorithmic systems," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT '23)*, 2023, pp. 651–666, doi: 10.1145/3593013.3594033. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/3593013.3594033>
6. M. Saarela, V. Heilala, P. Jääskelä, A. Rantakaulio, and T. Kärkkäinen, "Explainable student agency analytics," *IEEE Access*, vol. 9, pp. 137444–137459, Sep. 29, 2021, doi: 10.1109/ACCESS.2021.3116664. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9552837>
7. D. Ogbu, "Agentic AI in computer vision domain – recent advances and prospects," *International Journal of Research Publication and Reviews*, vol. 4, Dec. 2023, pp. 5102–5120, doi: 10.55248/gengpi.5.1124.3309. [Online]. Available: https://www.researchgate.net/publication/386292786_Agentic_AI_in_Computer_Vision_Domain_-_Recent_Advances_and_Prospects
8. G. Papagni, J. de Pagter, S. Zafari, M. Filzmoser, and S. T. Koeszegi, "Artificial agents' explainability to support trust: Considerations on timing and context," *AI & Society*, vol. 38, pp. 947–960, 2023, published online Jun. 27, 2022, doi: 10.1007/s00146-022-01462-7. [Online]. Available: <https://link.springer.com/article/10.1007/s00146-022-01462-7>
9. A. Dodda, "AI governance and security in fintech: Ensuring trust in generative and agentic AI systems," *Asian Journal of Applied Economics and Development*, vol. 1, no. 1, pp. 1—, Dec. 2, 2023. [Online]. Available: <https://aaed.com/index.php/aaed/article/view/6>
10. B. Abedin, C. Meske, I. Junglas, F. Rabhi, and H. R. Motahari-Nezhad, "Designing and managing human-AI interactions," *Information Systems Frontiers*, vol. 24, pp. 691–697, Jul. 18, 2022, doi: 10.1007/s10796-022-10313-1. [Online]. Available: <https://link.springer.com/article/10.1007/s10796-022-10313-1>
11. V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A review of trustworthy and explainable artificial intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, Jul. 20, 2023, doi: 10.1109/ACCESS.2023.3294569. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10188681>