Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00284-8

Original Article

A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems

Sarat Piridi [1], Satyanarayana Asundi [2], Srinivas Kamineni [3], Chaitanya B. Dadi [4]

1. *NPI, Microsoft, San Jose, USA*
2. *IT, TTI Consumer Power Tools Inc, Anderson, USA*
3. *IT, Walmart, Sunnyvale, USA*
4. *IT, National Grid, Atlanta, USA*

Received: May 21, 2026 | Review began: July 26, 2026 | Review ended: September 16, 2026 | Published: September 20, 2026

Abstract

The increasing use of intelligent automation in human-focused workplaces has intensified the need for governance mechanisms that allow the preservation of human power without compromising operational effectiveness. Traditional automation systems are often designed using static oversight models that are not responsive enough to contextual risk, changing system behavior, and dynamic human-AI interaction patterns. A cascaded human-in-the-loop governance architecture is one of the proposed architectures in this paper that incorporates adaptive oversight into intelligent automation processes. The proposed framework formalizes governance intensity as a dynamical system property, which is organized around five coordinated layers, namely: automation execution monitoring, probabilistic risk stratification with uncertainty-sensitive anomaly detection, ethical compliance validation, adaptive human escalation orchestration with configurable risk thresholds, and a feedback-based learning module, which augments the escalation policies over time through human correction as a feedback mechanism. As opposed to understanding governance as an ex post compliance role, the architecture realizes structured human-AI cooperation to maintain accountability, transparency, and labor agency in high-risk decision-making situations. Improvements in decision accuracy, compliance adherence, and risk detection sensitivity over a base of 15,000 simulated workforce workflow decisions are statistically significant and do not incur suboptimal productivity trade-offs. Longitudinal analysis also shows that unnecessary escalations are gradually reduced through feedback-induced optimization of the system, which proves that the system has the ability to achieve adaptive control without reducing human control. The results contribute to the design of intelligent systems that are people-centered, which can be achieved through the establishment of adaptive governance as a structural principle of intelligent system design in order to guarantee ethical, trustworthy, and sustainable intelligent automation across various fields of operation.

Categories: Ethical AI and Responsible Technology, Ethical and Social Implications, Human-Machine Interface

Keywords: human-in-the-loop, ai governance, adaptive escalation, ethical ai, workforce automation, risk stratification, feedback-driven optimization, human–ai collaboration

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

Introduction

The increasing adoption of intelligent automation technologies in organizations has fundamentally changed the environment of workforce decisions, operational planning, and enterprise risk management. With the move toward artificial intelligence (AI) systems designed as more than single analytical tools and capabilities, and toward their broader integration as decision-making agents within deeply embedded workflows and the mission space of workplaces, there is an urgency to establish sound governance mechanisms [1,2]. Contemporary organizations are shifting more and more consequential decisions, including recruitment checks, resource allocation, performance evaluation, regulatory guidelines, and operational safety checks, to AI-enhanced systems, which run with different levels of autonomy. As much as these systems promise a high level of throughput, consistency, and scalability, they also present governance issues that are beyond the structural capability of traditional oversight paradigms. The conflict between business effectiveness and the presence of human purpose in control is one of the most impactful design issues for organizations aspiring to use AI in a responsible manner in workplace settings [3,4].

Current governance models used for AI-enhanced systems largely use static oversight models; that is, these models place human review gateways at regular points or in response to hard-coded, inflexible rule sets. Although largely administratively uncomplicated, these methods have underlying drawbacks, including an inability to be flexible, sensitive to context, and poor scalability. Statistical thresholds cannot adapt to the changing risk distributions of active operational systems, often leading to either too much human intervention in routine tasks (degrading productivity gains) or too little intervention in truly abnormal or otherwise ethically sensitive situations [5,6]. Additionally, traditional human-in-the-loop (HITL) systems view human supervision as an external compliance feature that is added to automated pipelines, instead of an architecturally integrated governance system that can learn and adapt in both directions. This organizational constraint yields oversight mechanisms that are continuously fine-tuned based on operational experience at the time of first deployment and increasingly drift away over time as system behavior, workforce dynamics, and regulatory environments change [7].

The difficulty in effectively architecting human-AI governance systems is further increased by the heterogeneous nature of workforce decision-making contexts. Enterprises make decisions with significantly different risk profiles, ethical sensitivities, and levels of regulatory exposure, such as those related to employee safety, compensation equity, disciplinary actions, and resource allocation to operations. An oversimplified governance strategy cannot reflect this heterogeneity, and, as such, governance intensity is simultaneously oversensitive in low-risk routine decisions and undersensitive in high-stakes ethical resolutions [8,9]. Recent literature has emphasized the importance of risk-stratified governance systems that adjust the intensity of oversight in response to contextual risk evaluation, uncertainty quantification, and ethical conformity assessment [10,11]. Nevertheless, the current proposals are still largely theoretical, and they are not designed as real architectural designs, not validated empirically, and engineered to minimize feedback and optimize their use in complex workforce systems.

These limitations are addressed in this paper by suggesting a cascaded HITL governance architecture, an operationalization of adaptive oversight as a system property and not as a supervisory role. The proposed framework presents five integrated layers of governance, including monitoring of the execution of automation, risk stratification with uncertainty-sensitive anomaly detection, validation of ethical compliance, orchestration of adaptive human escalation, and feedback-driven learning, all of which allow the dynamic, context-specific governance of decisions related to the use of AI-enhanced workforce systems. The architecture makes the intensity of governance a system variable subjected to continued modulation, but controlled by a composite risk score, which is a weighted integration framework with operational, ethical, and uncertainty dimensions. The core of the design is as follows: a feedback-based escalation optimization module, which captures human correction signals where human reviewers override, modify, or confirm AI-generated decisions and introduces these signals into the systematic revision of escalation thresholds, the risk assessment model, and compliance evaluation criteria when the human correction signal is high [12,13].

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

Fourfold are the contributions to this work. First, we provide an architectural specification of a cascaded governance structure organizing human-AI cooperation at several integrated levels of coordinated supervision. Second, we present a risk stratification engine which is a probabilistic ensemble anomaly detector with quantified uncertainty to generate calibrated and context-sensitive risk measurements. Third, we develop feedback-based optimization of the escalation scheme, which makes it possible to gradually perfect the policies of governance by introducing systematic human control indicators. Fourth, we provide a controlled simulation-based evaluation of the proposed architecture using 15,000 synthetically generated workforce workflow decisions, enabling comparison with fully autonomous and alternative HITL baselines under consistent experimental conditions. The simulation-based evaluation shows improvements of 14.3 percentage points in decision accuracy, 19.7 percentage points in compliance adherence, and 22.1 percentage points in risk detection sensitivity relative to the fully autonomous baseline [14-16].

Literature review

HITL Systems and AI Governance Frameworks

HITL integration of automated decision systems has a long intellectual heritage dating back to several decades of research in the fields of human factors engineering, supervisory control theory, and human-computer interaction. Mosqueira-Rey et al. [2] presented a state-of-the-art review of HITL machine learning, identifying three major paradigms: HITL (human intervention in model training and inference), HITL (human monitoring and intervention capabilities), and human-in-command (human control over system boundaries and deployment choices). A critical gap revealed by their taxonomy is the absence of mechanisms for dynamic transition between these paradigms based on contextual risk assessment, but instead, a constant degree of human involvement, despite the importance of the decision being made.

Shneiderman [1] moved the discussion further by introducing a conceptualization of human-centered AI to include human control, transparency, and accountability in the design of AI systems. His framework also stresses the fact that good governance does not only happen when human checks are present, but through the structural incorporation of checks and balances that are auditable, responsive, and able to intervene meaningfully. A similar critical analysis was given by Green [3], who showed that policies that require human control over government algorithms often do not work in practice because of the existence of automation bias, cognitive overload, and a lack of institutional support to encourage meaningful human review. The implications of these findings indicate that governance architectures are needed to actively counter the barriers to effective human oversight arising from the cognitive and organizational limits of human supervision, rather than only to demand that humans should be embedded in automated decision loops.

Floridi and Cowls [5] developed a cohesive ethical framework that included five principles of AI in society, namely beneficence, non-maleficence, autonomy, justice, and explicability. This normative framework has become a common standard for AI governance, but the application of these principles in practice is quite difficult. To translate abstract moral values into concrete government procedures operating in real-time automated processes, it is necessary to employ architectural innovations that mediate normative structures and technical application. This translation problem is directly addressed in the proposed cascaded governance architecture, as ethical compliance validation is implemented as a special architectural layer with formal evaluation criteria and adjustable thresholds.

Risk Stratification and Anomaly Detection in Automated Decision Systems

Automated decision governance through the application of probabilistic risk assessment to automated systems is a new intersection of multiple computational fields. The paper by Xiao et al. [11] reviews human error in risk-informed decision-making, and the authors found the combination of human reliability assessment, AI, and human performance models to be of paramount importance for improving predictive fidelity in high-stakes systems. They found that the integration of cognitive models and AI-based analytics into risk-aware human reliability assessment pipelines can greatly enhance error prediction accuracy but requires richer datasets and more transparent algorithmic execution.

As a framework proposing the use of AI in financial risk assessment, Joshi [10] introduced the A-FRADA framework, which is a multi-layer framework that integrates three types of AI (Bayesian Networks, Gaussian Processes, and deep learning) and four types of large language models into an explainable AI structure. The framework was found to be more accurate

in analyzing the dependencies of financial risk chains while ensuring compliance with regulations by making decisions regarding the chain in view. Although the architectural concepts of layered risk assessment and explainability constraints are oriented toward the context of financial applications, workforce governance is an immediate application of these architectural principles. Hynek et al. [13] conducted a study on the implementation of AI-driven security solutions to protect critical infrastructure and found that there are strong demographic and sectoral differences in experts' perceptions of AI-enhanced governance systems. They emphasize the relevance of context-sensitive governance calibration, taking into consideration stakeholder views and institutional contexts.

Falemi et al. [8] proposed a conceptual model of enterprise risk management that combines HR strategy and operational analytics and operates across four interrelated dimensions: strategic alignment, risk intelligence, analytical integration, and adaptive governance. The adaptive governance layer, which introduces cross-functional communication channels to respond proactively to risks, provides a conceptual background to the governance optimization that is feedback-driven, which has been proposed in this work. Nevertheless, their framework is still theoretical and does not have formal specification and empirical validation similar to the architecture that is described here.

Ethical AI and Workforce Automation

The ethical aspects of using AI in workforce settings have been extensively discussed by scholars due to the increasing acknowledgment that automated decisions in workforce settings have serious consequences in terms of employee well-being, organizational fairness, and compliance with regulations. Budhwar et al. [4] conducted a study on the complex role of generative AI in human resource management and found that, among workforce-related issues, job displacement, algorithmic biases, and loss of human decision-making freedom are critical problems concerning generative AI. Their analysis focuses on the importance of governance models that would maintain meaningful human agency while using AI capabilities to achieve operational efficiency.

According to Whyte [6], the dynamics of human trust in AI-based decision-support systems in high-stakes security environments were examined; it was discovered that technical expertise has a positive effect on assessing the utility and credibility of AI. The sense of human agency in AI-enhanced decision loops, however, is paradoxically associated with increased readiness to work with less information, which means that governance structures should proactively organize human authority to avoid excessive trust in AI suggestions. A systematic review of interaction patterns in AI-assisted decision-making by Gomez et al. [7] showed that existing forms of interaction are characterized by simplistic collaboration paradigms that do not offer enough support for interactive human-AI collaboration. Their interaction pattern taxonomy recognizes huge potential for more advanced governance designs that allow adaptive, situation-specific human-AI interaction.

Al-Bashrawi et al. [12] examined the transformative potential of agentic AI systems for entrepreneurship, introducing the concept of entrepreneur-co-agency, where externalized memory, planning, and inference augment human decision-making. While focused on entrepreneurial contexts, their analysis of the shift from intuition-driven to data-driven decision-making mindsets has direct relevance to workforce governance, where similar cognitive and organizational transformations are required for effective human-AI collaboration. Nadler et al. [17] addressed the regulatory dimension, proposing a framework for human-centric AI regulation that balances innovation with societal resilience, providing normative guidance that informs the ethical compliance validation layer of the proposed architecture.

Feedback-Driven Learning and Adaptive Automation

Ranjan et al. [15] proposed the concept of enterprise agentic AI, and they analyzed how autonomous AI agents can be integrated successfully into organizational workflows without violating proper human control and governance mechanisms. Their analysis states the significance of systematic escalation policies and adjustable boundaries of autonomy - principles that feature in the adaptive escalation orchestration layer of the proposed architecture. Adeyemo [18] proposed a data-driven CAPA framework to improve quality assurance through continuous monitoring and process optimization. Marcellin et al. [19] investigated the intersection of AI and biotechnology governance and suggested 10 principles of governance based on adaptive, pluralistic, and capability-based governance. Their focus on anticipating future challenges, dual-use issues, and participatory decision-making is closely related to the feedback-based mechanism

for optimization of governance suggested here, which also seeks to produce governance systems that change as they experience operation while preserving principled boundaries for oversight. The integration of human feedback in automated system optimization is a developing area of research with a broad range of implications in a variety of application areas.

Barrett and Korostelina [20] highlighted the role of resilient AI-mediated communication in enhancing frontline public safety and decision-making. Srivastava et al. [21] introduced an AI-driven real-time threat detection framework for Zero Trust environments, demonstrating the effectiveness of continuous monitoring in strengthening cybersecurity.

Adaptive learning systems can also be successfully applied in pharmaceutical education. Wang et al. [22] revealed that the modeling of knowledge states and individual learning progression can significantly enhance the learning process. Although their study is specific to educational contexts, its design, which integrates real-time feedback with adaptive system behavior, provides a methodological precedent for the governance feedback loops proposed in this work.

Pomeroy [23] examined the strategic, ethical, and operational readiness required for integrating generative AI into healthcare and proposed a governance-oriented framework to support responsible, transparent, and clinically safe AI deployment.

Dogru et al. [24] reviewed reinforcement learning in the process industries, covering the use of reinforcement learning in manufacturing, energy systems, and robotics. Their discussion points out the opportunities and shortcomings of reward-based optimization in safety-critical settings, citing issues associated with sample efficiency, safety restrictions, and the disconnect between the simulated and real world. These challenges directly guide the design of the proposed feedback-driven learning module in the current work, based on the use of bounded policy updates, as well as safety-constrained optimization to stabilize governance refinement.

The literature review shows that there have been significant advances in human-centered AI, HITL decision-making support, risk-aware automation, ethical considerations of AI governance, and feedback-based adaptation. Current methods usually focus on one of these dimensions or assume more static ways of human oversight. There is a missing link, however, in the application of risk-sensitive escalation, independent ethical compliance enforcement, capability matching of reviewers, and feedback-driven governance adaptation in a cascaded architecture. Further, empirical support for such integrated governance mechanisms is also insufficient in workforce decision contexts, in particular. In this regard, this work aims to propose and test a governance structure in a five-layered cascaded structure in a controlled environment. The study is simulation-based and is therefore viewed as a proof-of-concept evaluation and not as a validation of deployment in real organizational environments.

Materials And Methods

Proposed methodology

The formalization of the HITL governance architecture is presented in this section. It aims to tackle the following key governance challenges in workforce systems with AI dependency: (i) static governance rules that are not flexible enough for changing risk profiles; (ii) the absence of clear ethical compliance checks in automated decision pipelines; and (iii) the lack of feedback mechanisms to update governance rules based on human corrective signals.

The architecture consists of five layers of governance pipeline. Each layer performs a dedicated governance function and passes structured decision artifacts as inputs to the subsequent layer in the pipeline. Governance decisions and risk signals cascade through the process, and human correction signals are added to the feedback-learning module for updating the parameters at the next time step. The overall architecture of the proposed framework is shown in Figure *1*.

All figures presented in this work are original contributions developed by the authors for this study and do not reproduce any previously published diagrams [25-27].

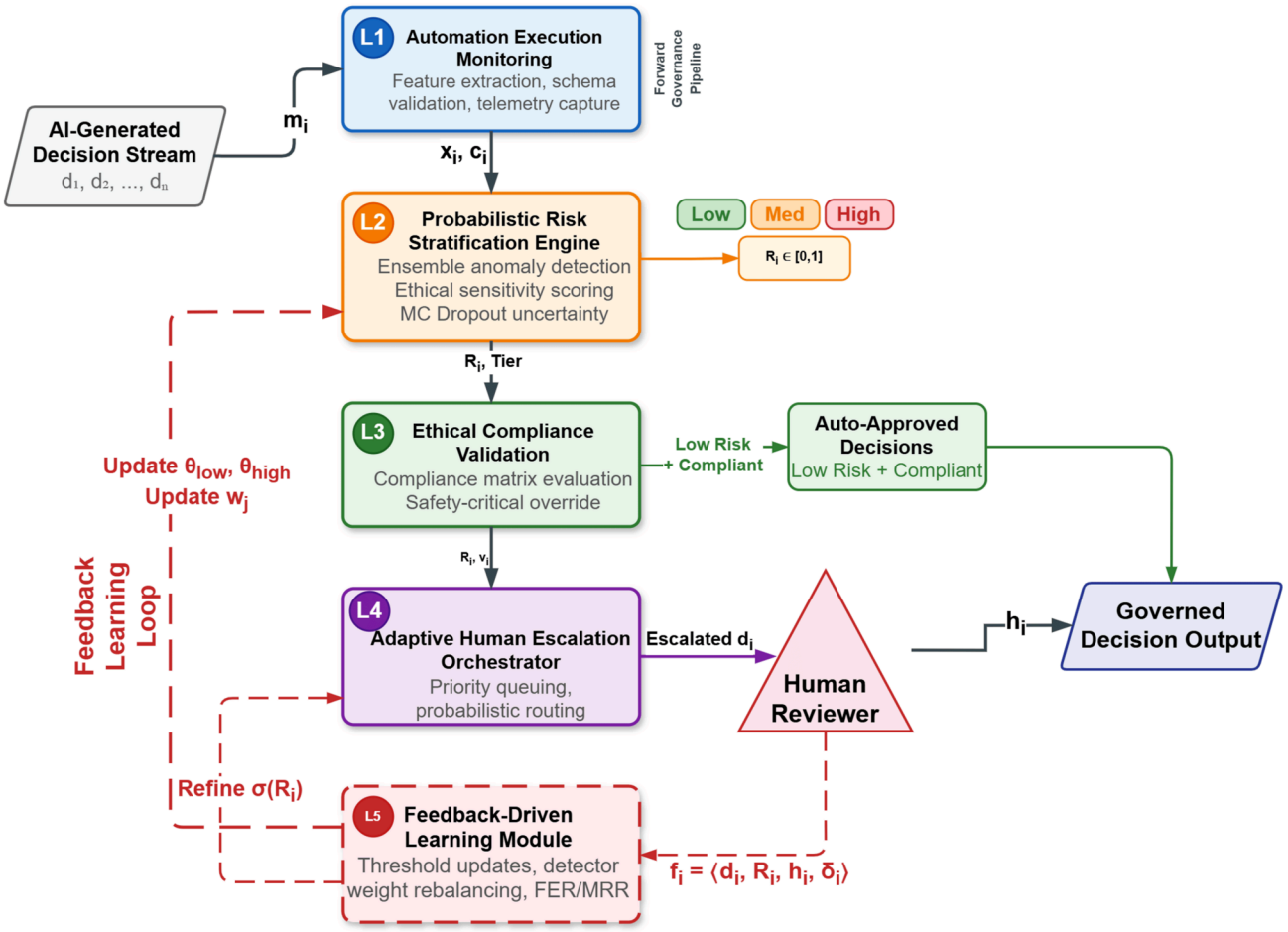

FIGURE 1: High-level architecture of the proposed cascaded human-in-the-loop governance framework showing the five coordinated governance layers and their interconnections

Original figure developed by the authors.
FER, False Escalation Rate; MC, Monte Carlo; MRR, Missed Risk Rate

As illustrated in Figure *1*, the five layers operate in a sequential cascade, with governance decisions and risk signals propagating downward while human correction feedback propagates upward through the learning module.

Layer 1: Automation Execution Monitoring

The automation execution monitoring layer offers real-time data on all AI-made workforce decisions, and structured telemetry data is produced, which can be further analyzed as input for downstream governance assessment. This layer is the entry point of the governance pipeline and is responsible for ensuring that every automated decision is fully observable before any risk or compliance evaluation is applied. Given an automated decision d_i at time *t* obtained by the monitoring layer, a feature vector $x_i \in \mathbb{R}^n$ comprising decision attributes, contextual parameters, and system state indicators is extracted. Formally, the monitoring output for decision d_i is defined as:

$$m_i = \langle x_i, c_i, s_i, \tau_i \rangle \tag{1}$$

where each parameter is defined as follows: $x_i \in \mathbb{R}^n$ is the *n*-dimensional decision feature vector capturing operational attributes of decision d_i ; $c_i \in \mathbb{R}^k$ is the contextual metadata vector of dimension *k*, encoding domain-specific context such as workforce sector, regulatory jurisdiction, and historical decision precedent; $s_i \in \{0,1\}$ is a binary syntactic

validity flag, where s_i = 1 indicates a well-formed decision record; and $\tau_i \in \mathbb{R}^+$ is the Unix timestamp at which decision d_i was generated. Schema validation, completeness checking, and format normalization are enforced in the monitoring layer on the input data to guarantee uniform quality for the downstream governance layers.

Layer 2: Probabilistic Risk Stratification Engine

The risk stratification engine constitutes the computational heart of the governance framework, converting the monitoring telemetry into risk measurements by assessing risk on a scale through a multi-component evaluation framework. A composite risk score $R_i \in \{0,1\}$ is calculated by the engine for each decision d_i by aggregating three orthogonal risk dimensions: operational risk r_i^{op}, ethical sensitivity risk and predictive uncertainty u_i. These three dimensions are orthogonal by design: operational risk captures anomalous deviations from normal decision patterns; ethical risk captures sensitivity across protected workforce attributes; and predictive uncertainty captures the AI model's own confidence level, which directly influences how much a given decision should be trusted without human review.

The operational risk element is calculated using an ensemble anomaly detection system that includes Isolation Forest, One-Class SVM, and Local Outlier Factor detectors. The ensemble anomaly score is obtained through weighted averaging:

$$r_i^{\text{op}} = \sum_{j=1}^{J} w_j \cdot a_j(x_i), \quad \text{subject to} \quad \sum_{j=1}^{J} w_j = 1 \tag{2}$$

where each parameter is defined as follows: $a_j(\mathbf{x}_i) \in \{0,1\}$ is the normalized anomaly score assigned by detector *j* to decision feature vector x_i, with higher values indicating greater deviation from the learned normal decision distribution; $w_j \geq 0$ is the non-negative weight assigned to detector *j*, optimized through held-out validation performance; and $J = 3$ is the total number of ensemble detectors (Isolation Forest, One-Class SVM, and Local Outlier Factor). The unit-sum constraint $\sum_{j=1}^{J} w_j = 1$ ensures that r_i^{op} remains a properly normalized score in {0.1}.

The ethical sensitivity risk r_i^{eth} is evaluated against a predefined ethical evaluation space $E = \{e_1, e_2, \ldots, e_K\}$

$$r_i^{\text{eth}} = \frac{1}{K} \sum_{k=1}^{K} \phi_k(x_i, c_i) \tag{3}$$

where each parameter is defined as follows: $\phi_k(\mathbf{x}_i, \mathbf{c}_i) \in \{0,1\}$ is the ethical sensitivity evaluation function for dimension *k*, mapping the decision feature vector x_i and contextual metadata c_i to a normalized sensitivity score; and *K* is the total number of ethical evaluation dimensions configured for the deployment domain (e.g., *K* = 4 for fairness, privacy, transparency, and safety). The arithmetic mean over all *K* dimensions produces an aggregate ethical risk score $r_i^{eth} \in \{0,1\}$.

The predictive uncertainty measure u_i measures the confidence of the AI system in the generated decision using a Monte Carlo Dropout approximation to Bayesian inference. In a *T* stochastic forward pass neural decision model:

$$u_i = \frac{1}{T} \sum_{t=1}^{T} \|\hat{y}_i^{(t)} - \bar{y}_i\|^2, \quad \bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_i^{(t)} \tag{4}$$

where each parameter is defined as follows: *T* denotes the number of stochastic forward passes performed through the neural network with Monte Carlo dropout active at inference time (default *T* = 50); y^t_i is the model output at the *t*-th stochastic forward pass for decision d_i; $\hat{y}_i^{(t)}$ is the scalar or vector output of the *t*-th stochastic forward pass for decision d_i; $\bar{y}_i$ is the Monte Carlo mean prediction computed across all *T* passes; and $\|\cdot\|^2$ denotes the squared Euclidean norm. The resulting $u_i \geq 0$ measures the empirical variance of the model's output distribution, with larger values indicating higher predictive uncertainty and, therefore, lower confidence in the AI decision.

These components are combined in the composite risk score by a parameterized combination:

$$R_i = \alpha \cdot r_i^{\text{op}} + \beta \cdot r_i^{\text{eth}} + \gamma \cdot u_i, \quad \alpha + \beta + \gamma = 1 \tag{5}$$

where each parameter is defined as follows: $\alpha \in \{0, 1\}$ is the weight assigned to operational anomaly risk; $\beta \in \{0, 1\}$ is the weight assigned to ethical sensitivity risk; $\gamma \in \{0, 1\}$ is the weight assigned to predictive uncertainty; and the constraint $\alpha + \beta + \gamma = 1$ ensures that $R_i \in \{0, 1\}$ remains a properly normalized composite score. Default values are α = 0.40, β = 0.35, and γ = 0.25, reflecting the relative governance priority of operational correctness, ethical compliance, and model confidence in workforce decision contexts (see Table 1). Table 1 summarizes the risk stratification parameters.

Parameter	Description	Default	Range
α	Operational risk weight	0.40	(0, 1)
β	Ethical sensitivity weight	0.35	(0, 1)
γ	Uncertainty weight	0.25	(0, 1)
J	Ensemble detector count	3	{2, 3, 4, 5}
T	Monte Carlo Dropout passes	50	{20, 50, 100}
θ_{low}	Low-risk threshold	0.30	(0.1, 0.4)
θ_{high}	High-risk threshold	0.70	(0.6, 0.9)

TABLE 1: Risk stratification engine parameters and default configurations

Risk stratification partitions decisions into three governance tiers based on configurable thresholds θ_{low} and θ_{high}:

$$\text{Tier}(d_i) = \begin{cases} \text{Low Risk (Auto-approve)} & \text{if } R_i < \theta_{\text{low}} \\ \text{Medium Risk (Flagged Review)} & \text{if } \theta_{\text{low}} \leq R_i < \theta_{\text{high}} \\ \text{High Risk (Mandatory Escalation)} & \text{if } R_i \geq \theta_{\text{high}} \end{cases} \tag{6}$$

where each parameter is defined as follows: θ_{low} = 0.30 is the lower risk boundary below which decisions are automatically approved without human review, and θ_{high} = 0.70 is the upper boundary above which human escalation is mandatory. Decisions falling in the interval (θ_{low}, θ_{high}) are routed to probabilistic review, as described in Layer 4. Tier(d_i) denotes the governance tier assigned to decision d_i, categorizing it as Low Risk, Medium Risk, or High Risk according to the composite risk score. Figure 2 details the internal data flow of the risk stratification engine, illustrating how the three orthogonal components (operational anomaly detection, ethical sensitivity scoring, and Monte Carlo uncertainty estimation) are independently computed and then fused into the composite risk score R_i via Equation 5.

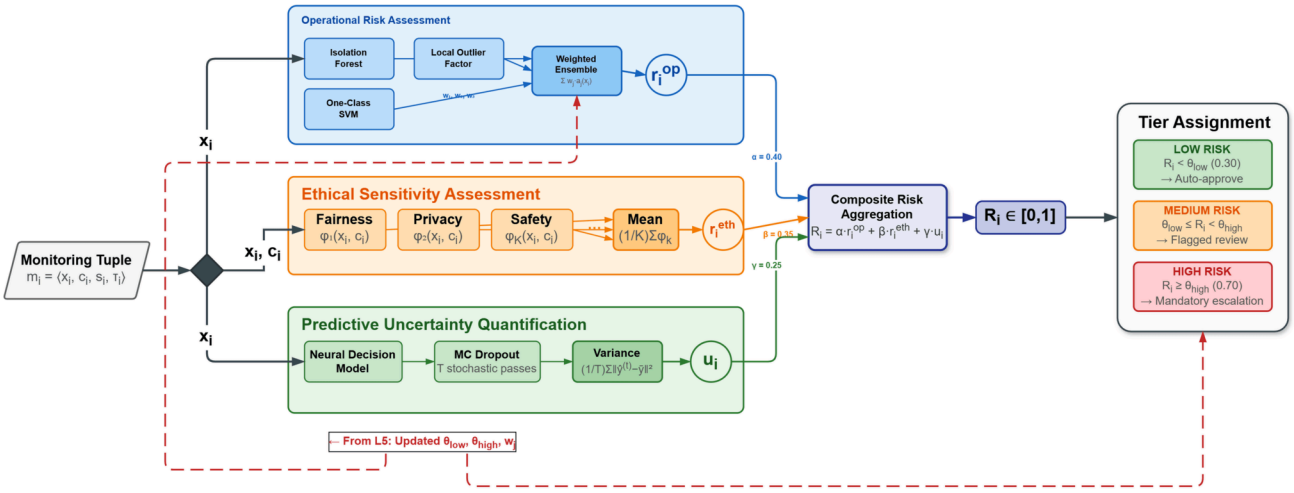

FIGURE 2: Internal architecture of the probabilistic risk stratification engine showing the three orthogonal risk assessment components and their aggregation into composite risk scores

Original figure developed by the authors

Layer 3: Ethical Compliance Validation

Each decision is tested against an organized compliance matrix, which is used to encode domain-specific regulatory requirements, organizational policies, and ethical guidelines by the ethical compliance validation layer. This layer operates independently of the composite risk score and enforces hard ethical boundaries: any detected compliance violation unconditionally triggers human escalation, regardless of how low the value of R_i may be. This safety-critical design prevents risk score aggregation from obscuring violations of fundamental ethical or regulatory requirements. The compliance analysis produces a binary compliance vector:

$$v_i = \left[v_i^{(1)}, v_i^{(2)}, \ldots, v_i^{(M)}\right] \in \{0,1\}^M \tag{7}$$

where each parameter is defined as follows: $v_i^{(m)} \in \{0, 1\}$ indicates the compliance status of decision d_i with respect to requirement m, with $v_i^{(m)} = 1$ denoting compliance and $v_i^{(m)} = 0$ denoting a violation; and M is the total number of compliance criteria configured in the deployment domain's compliance matrix. A decision is deemed compliant if and only if $\sum_{m=1}^{M} v_i^{(m)} = M$. Any compliance breach will trigger compulsory escalation independent of the composite risk score, and a safety-crucial bypass mechanism is provided such that the aggregate of risk scores never may be used to breach ethical limits.

Layer 4: Adaptive Human Escalation Orchestration

The escalation orchestration layer deals with the routing of decisions that involve human review, which applies adaptive workload balancing, priority-based queuing, and matching reviewer expertise. The decision at the escalation layer combines the results of the risk stratification layer and the compliance validation layer:

$$\text{Escalate}(d_i) = \begin{cases} 1 & \text{if } R_i \ge \theta_{\text{high}} \text{ or } \sum_{m=1}^{M} v_i^{(m)} < M \\ p_i & \text{if } \theta_{\text{low}} \le R_i < \theta_{\text{high}} \\ 0 & \text{if } R_i < \theta_{\text{low}} \text{ and } \sum_{m=1}^{M} v_i^{(m)} = M \end{cases} \tag{8}$$

where each parameter is defined as follows: Escalate(d_i) = 1 denotes mandatory escalation to a human reviewer; Escalate(d_i) = 0 denotes automatic approval without human intervention; p_i ~ Bernoulli ($\sigma(R_i)$) is a probabilistic escalation flag for medium-risk decisions, with $\sigma(\cdot)$ denoting the sigmoid function that maps the composite risk score to an escalation probability; and $\sum_{m=1}^{M} v_i^{(m)} < M$ denotes the presence of at least one compliance violation that unconditionally overrides the risk score and mandates escalation.

The reviewer-assignment functionality transfers capability-matched decisions to competent reviewers according to a capability-matching matrix:

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,L} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{P,1} & c_{P,2} & \cdots & c_{P,L} \end{bmatrix} \tag{9}$$

where each parameter is defined as follows: $C \in \{0,1\}^{P \times L}$ is the reviewer capability matrix; $c_{p,l} \in \{0,1\}$ represents the domain proficiency score of reviewer $p \in \{1, \ldots, P\}$ in decision domain $l \in \{1, \ldots, L\}$; P is the total number of available human reviewers in the reviewer pool; and L is the number of distinct decision domain categories (e.g., HR, safety, compliance). Reviewer assignment selects the reviewer $p^* = \arg\max c_{p,l(d_i)}$ for the domain $l(d_i)$ of escalated decision d_i, subject to availability and queue load constraints.

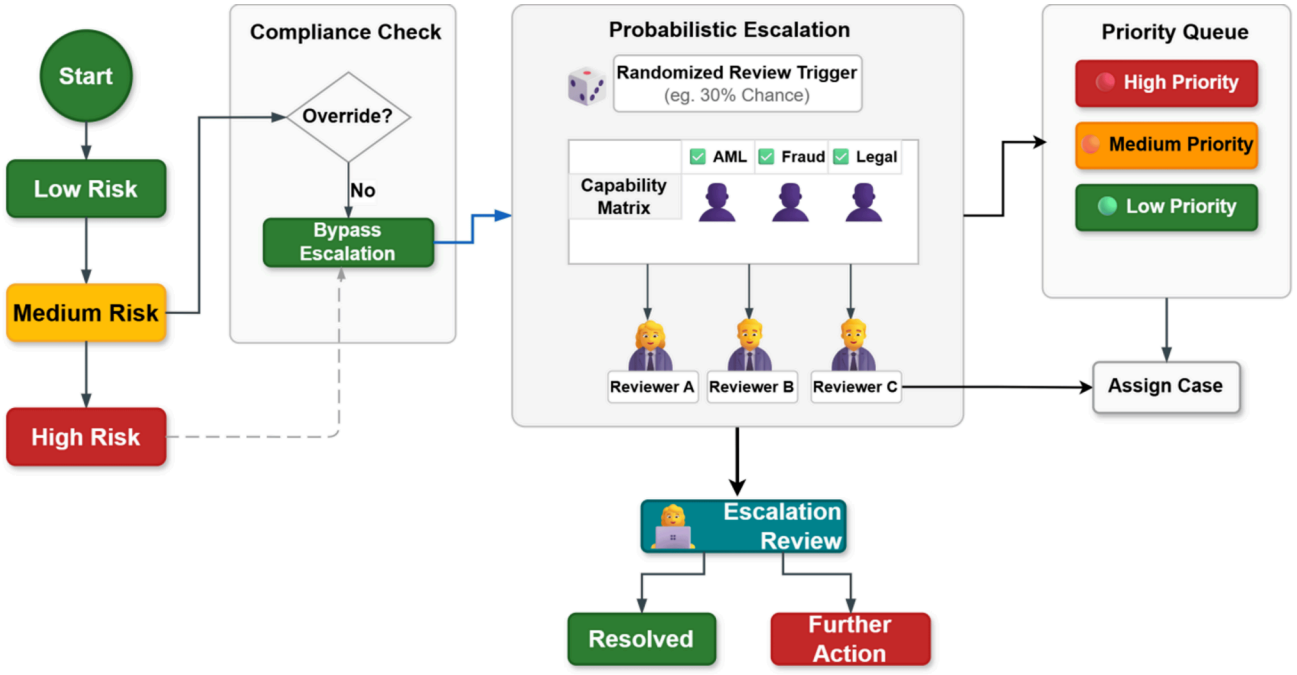

FIGURE 3: Adaptive human escalation orchestration workflow showing the decision routing logic, reviewer capability matching, and priority-based queue management

Original figure developed by the authors.

As shown in Figure 3, escalated decisions are routed through a three-stage pipeline: (i) risk- and compliance-based routing using Equation 8; (ii) reviewer matching using the capability matrix C from Equation 9; and (iii) priority-based queue management that orders pending reviews by composite risk score in descending order.

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

Layer 5: Feedback-Driven Learning Module

The feedback-driven learning module captures the actions of human reviewers on escalated decisions and uses this feedback as corrective signals to continually improve governance parameters. For every escalated decision d_i that is examined by a human, a feedback tuple is recorded by the module:

$$f_i = \langle d_i, R_i, h_i, \delta_i, \tau_i^{\text{rev}} \rangle \tag{10}$$

where each parameter is defined as follows: d_i is the escalated decision identifier; $R_i \in \{0, 1\}$ is the composite risk score assigned to d_i at the time of escalation; $h_i \in$ {confirm, modify, override} is the human reviewer action, indicating whether the reviewer agreed with, partially changed, or fully reversed the AI decision; $\delta_i \in \{0, 1\}$ quantifies the magnitude of any modification applied by the reviewer (with $\delta_i = 0$ for confirmations and $\delta_i = 1$ for full overrides); and $\tau_i^{rev} \in \mathbb{R}^+$ is the timestamp at which the human review was completed.

The feedback-learning algorithm adjusts the risk-stratification thresholds θ_{low} and θ_{high}

using an exponentially weighted moving average based on the false escalation rate (FER) and the missed risk rate (MRR):

$$\theta_{\text{low}}^{(t+1)} = \theta_{\text{low}}^{(t)} + \eta \cdot \left(\text{FER}^{(t)} - \text{MRR}^{(t)} \right) \tag{11}$$

$$\theta_{\text{high}}^{(t+1)} = \theta_{\text{high}}^{(t)} - \eta \cdot \text{MRR}^{(t)} + \lambda \cdot \text{FER}^{(t)} \tag{12}$$

where each parameter is defined as follows: $\theta_{low}^{(t)}$ and $\theta_{high}^{(t)}$ are the current lower and upper risk thresholds at update step t; $\eta > 0$ is the learning rate controlling the step size of threshold updates (default $\eta = 0.01$); $\lambda > 0$ is a regularization parameter that penalizes excessive threshold relaxation to prevent the system from becoming under-escalating over time (default $\lambda = 0.005$); $\text{FER}^{(t)}$ is the false escalation rate at step t (proportion of escalated decisions confirmed by reviewers without modification); and $\text{MRR}^{(t)}$ is the missed risk rate at step t (proportion of auto-approved decisions subsequently found to contain errors during post-review auditing). These rates are computed over a sliding window of W recent decisions.

The FER and MRR are formally defined as:

$$\text{FER}^{(t)} = \frac{|\{d_i : \text{Escalate}(d_i) = 1 \wedge h_i = \text{confirm}\}|}{W} \tag{13}$$

$$\text{MRR}^{(t)} = \frac{|\{d_i : \text{Escalate}(d_i) = 0 \wedge \text{PostReview}(d_i) = \text{error}\}|}{W} \tag{14}$$

where each parameter is defined as follows: W is the sliding window size (default $W = 500$ decisions); $\text{Escalate}(d_i) = 1$ indicates that the decision was sent for human review; $h_i =$ confirm means that the human reviewer agreed with the AI decision without modification; $\text{Escalate}(d_i) = 0$ indicates that the decision was auto-approved; and $\text{PostReview}(d_i) =$ error indicates that the auto-approved decision was flagged as erroneous during a subsequent periodic audit.

Also, the ensemble detector weights in Equation 2 are updated using a performance-weighted rebalancing mechanism. Detector weights are adjusted proportionally after every feedback cycle to improve the risk assessment performance of each determinant:

$$w_j^{(t+1)} = \frac{w_j^{(t)} \cdot \exp\left(\mu \cdot \text{Acc}_j^{(t)}\right)}{\sum_{j'=1}^{J} w_{j'}^{(t)} \cdot \exp\left(\mu \cdot \text{Acc}_{j'}^{(t)}\right)} \tag{15}$$

where each parameter is defined as follows: $w_j^{(t)}$ is the current weight of detector j at update step t; $\mathrm{Acc}_j^{(t)} \in \{0,1\}$ is the classification accuracy of detector j computed over the most recent W feedback-labeled decisions; $\mu > 0$ is a temperature parameter that controls the aggressiveness of weight rebalancing, with larger μ causing faster convergence toward the best-performing detector (default $\mu = 2.0$); and the softmax-style normalization in the denominator ensures that updated weights $\{w_j^{(t+1)}\}_{j=1}^{J}$ satisfy the unit-sum constraint $\sum_{j=1}^{J} w_j^{(t+1)} = 1$.

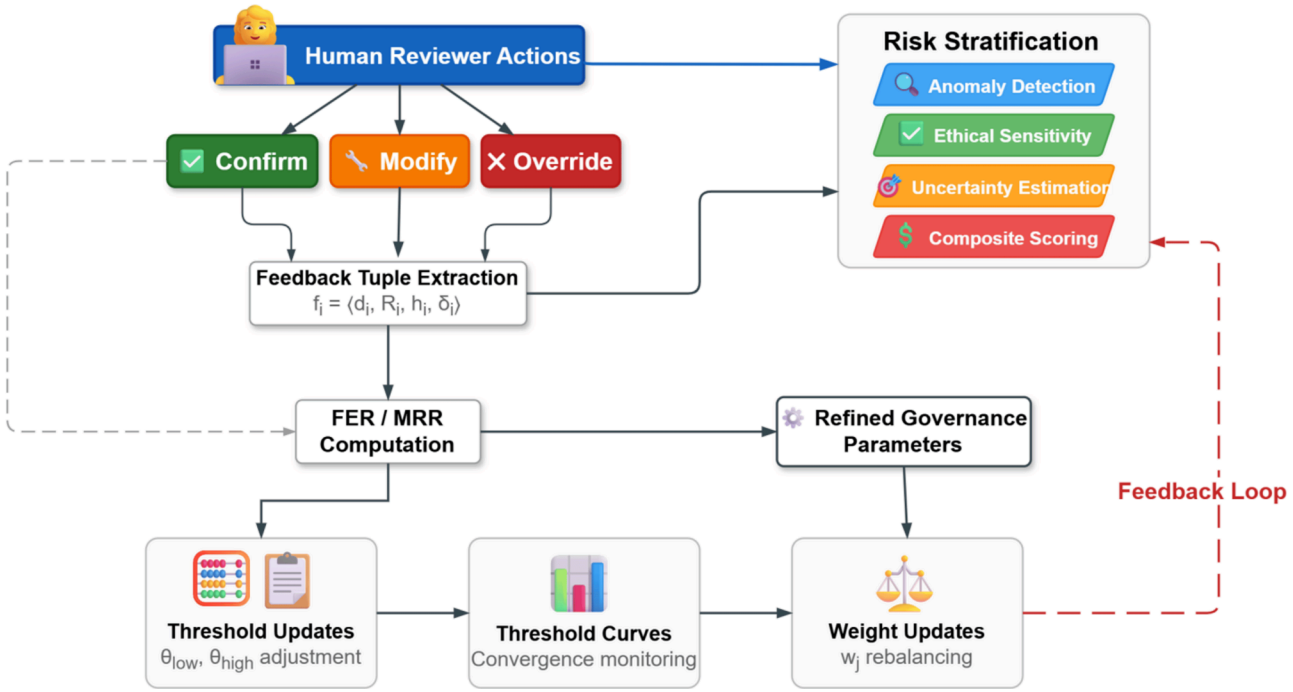

FIGURE 4: Feedback-driven learning module architecture showing the closed-loop governance refinement cycle

Original figure developed by the authors.
FER, False Escalation Rate; MRR, Missed Risk Rate

Figure 4 illustrates the closed-loop governance refinement cycle implemented by the feedback-learning module. Human reviewer actions collected via Equation 10 are processed for each W-decision window to update the three sets of learnable governance parameters: the risk thresholds θ_{low} and θ_{high} (Equations 11 and 12), and the ensemble detector weights $\{w_j\}$ (Equation 15).

End-to-End Governance Pipeline

Algorithm 1 presents the complete governance pipeline integrating all five architectural layers.

Algorithm 1: Cascaded HITL Governance Pipeline

Require: Decision stream $D = \{d_1, d_2, \ldots\}$, thresholds $\theta_{\text{low}}, \theta_{\text{high}}$, weights $\{w_j\}$

Ensure: Governance outcomes G

1. for each decision $d_i \in D$

2. Layer 1: Extract monitoring tuple $m_i = \langle x_i, c_i, s_i, \tau_i \rangle$

3. Layer 2: Compute risk components $r_i^{\text{op}}, r_i^{\text{eth}}, u_i$

4. Compute composite score $R_i = \alpha \cdot r_i^{\text{op}} + \beta \cdot r_i^{\text{eth}} + \gamma \cdot u_i$

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

5. Assign risk tier: $\mathbf{Tier}(d_i)$ via Equation 6

6. Layer 3: Evaluate compliance vector $\boldsymbol{v_i}$

7. Layer 4: Determine escalation via Equation 8

8. if $\mathbf{Escalate}(d_i) = \mathbf{1}$ then

9. Assign reviewer via capability matrix $\boldsymbol{C}$

10. Collect human feedback $\boldsymbol{f_i}$

11. Layer 5: Update $\boldsymbol{\theta_{\text{low}}, \theta_{\text{high}}, \{w_j\}}$ via Equation 11 and Equation 15

12. end if

13. Record governance outcome g_i to G

14. end for

15. Return $\boldsymbol{G}$

Experimental setup

Simulation Environment and Dataset Generation

The experimental evaluation was carried out in a purpose-built simulation environment set up to simulate a range of heterogeneous workforce decision scenarios under controlled and repeatable conditions. In the present study, no organizational dataset was available; therefore, all experimental observations had to be synthetically generated. The simulation was therefore built as a proof of concept to test the internal dynamics of the proposed governance design rather than as a replacement for real-world governance validation within organizations.

A total of 15,000 workforce decisions were simulated and distributed across five representative decision domains: Human Resources (HR) operations, operational safety management, financial compliance, resource allocation, and performance evaluation. Each domain contained 3,000 decision instances. Each decision was encoded as a 48-dimensional feature vector, including operational attributes of the decision, contextual metadata, temporal indicators, and domain-specific parameters. The five domains were chosen to create diversity with respect to operational risk, ethical sensitivity, and regulatory exposure.

The simulation assumes that workforce decisions are not uniformly random but follow structured distributions that differ across individuals. Baseline feature distributions were therefore parameterized to represent normal operating conditions, with domain-specific variations added to represent differences in risk and decision complexity. These assumptions were based on the literature on workforce analytics and organizational decision-making. The distributions obtained are indicative of a range of assumptions about the organizations and not observations of any specific organization.

The ability of the proposed architecture to detect anomalies was tested by artificially perturbing decision instances and introducing anomalies into the test set. The perturbation mechanisms used comprised feature drift, policy-violation patterns, and ethical-boundary violations. The overall anomaly rate was approximately 8% of the simulated sample, with domain-specific anomaly rates shown in Table 2. Anomaly injection was used to develop high-risk cases with which to test the risk stratification and escalation processes.

Compliance violations were randomly generated according to domain-specific compliance conditions. A simulated decision was marked as non-compliant if one or more compliance requirements were not met. Compliance labels were separated to allow the ethical compliance layer to be assessed as a standalone safety feature. This separation also enabled the experiments to evaluate whether escalation could occur even when the numerical risk score was relatively low (as a composite).

A consensus-based labeling protocol with three simulated domain-expert assessors was used to establish reference labels for risk status, compliance status, and the governance action that should be taken. Assessors were asked to independently score the decision cases generated based on the rules of the domain and the governance criteria. Differences were resolved by consensus for cases to arrive at the reference label for evaluation. These labels are based on expert judgment and do not reflect what would likely be applied by operational organizational reviewers.

The resulting dataset was then split and processed using the same protocol of the simulation for all methods. A total of 10 independent random seeds were used to repeat the experiment, thus providing some variation in results due to random generation and evaluation. The entire data set was thus used to compare the proposed architecture with baseline governance strategies under the same conditions.

The simulation does include several domains, multiple heterogeneous risk profiles, controlled anomalies and compliance violations, and expert-style consensus labels, but it does not represent all the features of real workforce decision-making. For example, there can be significant variations between organizational contexts, lack of documentation, disagreement among reviewers, changing regulations, concept drift, and other operational contexts. Therefore, the findings must be considered as evidence of computational feasibility and proof-of-concept effectiveness, and not as direct evidence of organizational superiority in practice.

Table *2* summarizes the dataset characteristics and distribution parameters.

Domain	Decisions	Features	Anomaly %	Compliance Violations %
HR Operations	3,000	48	7.5	4.2
Safety Management	3,000	48	9.3	6.1
Financial Compliance	3,000	48	8.7	5.8
Resource Allocation	3,000	48	6.8	3.1
Performance Evaluation	3,000	48	7.7	4.5
Total	15,000	48	8.0	4.7

TABLE 2: Simulated workforce decision dataset characteristics

HR, Human Resources

Implementation Details

It used the following technology stack to implement the framework in Python version 3.10: scikit-learn 1.3.2 to perform ensemble anomaly detection (Isolation Forest, One-Class SVM, Local Outlier Factor), PyTorch 2.1.0 to perform individual neural decision models (with Monte Carlo dropout-based uncertainty estimation), and custom modules to perform ethical compliance and ensure order-based escalation. The feedback-based learning module was applied on the basis of a sliding window mechanism that allows various window sizes ($W = 500$ decisions) and a learning rate ($\eta = 0.01$). The experiments were run on an Intel Xeon E5-2680 v4 processor, 64 GB of RAM, and an NVIDIA Tesla V100 graphics card. The simulation code and configuration files were developed specifically for this study. At the time of manuscript revision, the implementation is not publicly archived. To support reproducibility, the manuscript therefore reports the principal

simulation assumptions, dataset structure, parameter settings, experimental hardware and software environment, anomaly-generation procedure, labeling protocol, and evaluation protocol. Public release of the implementation can be considered following completion of the publication process.

Baseline Methods and Evaluation Metrics

Five baseline methods were established for comparative evaluation:

1. Fully Autonomous (FA): Everything made without human intervention.

2. Static HITL (S-HITL): Random sampling with 20% human review.

3. Rule-Based Escalation (RBE): Predefined rule alarms to be evaluated by humans depending on constant feature thresholds.

4. Risk-Only Stratification (ROS): No ethical compliance validation, feedback learning: composite risk scoring.

5. Non-Adaptive Cascaded (NAC): Full cascaded architecture without the feedback-based learning module.

The baseline methods used for comparative evaluation are summarized in Table *3*.

Metric	Definition
Decision Accuracy (DA)	Percentage of decisions matching expert-validated ground truth
Compliance Adherence Rate (CAR)	Percentage of decisions satisfying all compliance criteria
Risk Detection Sensitivity (RDS)	True positive rate for identifying high-risk decisions
False Escalation Rate (FER)	Proportion of escalated decisions confirmed without modification
Escalation Efficiency (EE)	Ratio of meaningful escalations to total escalations
Productivity Index (PI)	Normalized throughput relative to fully autonomous baseline

TABLE 3: Evaluation metrics used for comparative assessment of all governance methods. Six quantitative indicators are defined to measure decision quality, compliance adherence, risk detection capability, escalation behavior, and operational productivity

Experimental Protocol

The experiments were structured into three evaluation stages: (i) cross-sectional performance evaluation of all methods using the complete 15,000-decision dataset; (ii) longitudinal feedback learning analysis that monitors the evolution of governance metrics under all the architectural parameters (α, β, γ, η, W); and (iii) sensitivity analysis of the effects of key architectural parameters (α, β, γ, η, W) on system performance. Paired t-tests with Bonferroni correction at $p < 0.05$ were used to determine statistical significance, and Cohen's d was used to calculate effect sizes. For all statistically significant comparisons reported in the Results section, exact p-values and corresponding Cohen's d effect sizes are reported to quantify both statistical and practical significance. Each experiment was repeated using 10 independent random seeds to provide a robust estimation. For reproducibility, all experiments used the same predefined simulation configuration across methods, while ten independent random seeds were used to quantify stochastic variation. The simulation

generated the same five decision domains, 48-dimensional feature representation, anomaly-generation mechanisms, compliance-labeling rules, and consensus-based reference labels for each comparative evaluation. No real employee-level or organization-identifiable data were used in the experiments.

Results And Discussion

Experimental setup summary

For completeness and reader convenience, we briefly reiterate the experimental configuration underpinning all results reported in this section. The evaluation was conducted on a dataset of 15,000 simulated workforce workflow decisions spanning five organizational domains (HR Operations, Safety Management, Financial Compliance, Resource Allocation, and Performance Evaluation), each comprising 3,000 instances with 48-dimensional feature vectors. The anomaly injection rate was set at 8% and the compliance violation rate at 4.7% across the full dataset, as detailed in Table *2*. The proposed cascaded HITL framework was implemented in Python 3.10 with scikit-learn 1.3.2 and PyTorch 2.1.0, executed on an Intel Xeon E5-2680 v4 CPU with 64 GB RAM and an NVIDIA Tesla V100 GPU. Five baseline methods were evaluated: FA, S-HITL, RBE, ROS, and NAC. All results are averaged over 10 independent random seeds, with statistical significance assessed via paired t-tests with Bonferroni correction at $p < 0.05$.

Cross-sectional performance comparison

Table *4* reveals the overall comparison of the performance of all six methods on the entire 15,000-decision evaluation dataset. The proposed cascaded HITL governance structure achieves the best scores on all governance-critical metrics and is competitive in terms of productivity.

Method	DA (%)	CAR (%)	RDS (%)	FER (%)	EE (%)	PI
FA	78.2	71.4	68.2	—	—	1.00
S-HITL	83.6	79.8	64.3	42.7	57.3	0.81
RBE	85.1	82.3	71.6	35.2	64.8	0.84
ROS	88.7	86.5	79.4	28.6	71.4	0.87
NAC	90.3	88.9	84.7	22.1	77.9	0.89
Proposed	**92.5**†	**91.1**†	**90.3**†	**14.8**†	**85.2**†	**0.91**

TABLE 4: Cross-sectional performance comparison across all evaluation metrics

Bold values indicate best performance.
Exact p-values and corresponding Cohen's d effect sizes are as follows: DA (p = 0.003, d = 1.24), CAR (p = 0.002, d = 1.37), RDS (p = 0.001, d = 1.51), FER (p = 0.004, d = 1.18), and EE (p = 0.003, d = 1.29).
† denotes statistically significant improvement over the next-best method based on paired t-tests with Bonferroni correction.
CAR, Compliance Adherence Rate; DA, Decision Accuracy; EE, Escalation Efficiency; FA, Fully Autonomous; FER, False Escalation Rate; NAC, Non-Adaptive Cascaded; PI, Productivity Index; RBE, Rule-Based Escalation; RDS, Risk Detection Sensitivity; ROS, Risk-Only Stratification; S-HITL, Static Human-in-the-Loop

Statistical analysis confirmed that the performance improvements achieved by the proposed cascaded HITL governance framework were statistically significant across all primary governance metrics. Compared with the NAC baseline, DA improved significantly (92.5% vs. 90.3%; p = 0.003, Cohen's d = 1.24), CAR increased significantly (91.1% vs. 88.9%; p = 0.002, d = 1.37), and RDS also showed a significant improvement (90.3% vs. 84.7%; p = 0.001, d = 1.51). Furthermore, the reduction in FER (14.8% vs. 22.1%) was statistically significant (p = 0.004, d = 1.18), while EE increased significantly from 77.9% to 85.2% (p = 0.003, d = 1.29). According to the conventional interpretation of Cohen's d, all observed effects represent large effect sizes, indicating that the improvements are not only statistically significant but also practically meaningful.

Notably, the proposed architecture has a decision accuracy of 92.5%, which is 14.3 percentage points higher than the fully autonomous baseline and 2.2 percentage points higher than the non-adaptive cascaded counterpart. The compliance adherence rate is 91.1%, which is 19.7 percentage points higher than the autonomous baseline, indicating the usefulness of the dedicated ethical compliance validation layer. The highest relative increase (22.1 percentage points over the autonomous baseline) is the risk detection sensitivity of 90.3%, which shows that the probabilistic risk stratification engine with uncertainty-aware anomaly detection greatly increases the capacity of the system to detect truly high-risk decisions that need human intervention.

Notably, the false escalation figure of 14.8% is a decrease of 27.9% compared with the non-adaptive HITL approach, indicating that the adaptive governance approach significantly decreases unnecessary human engagement without any adverse effect on the quality of oversight. The proposed architecture has a productivity index of 0.91, which shows that the productivity of the fully autonomous system has been retained, with a productivity trade-off of 91% in the proposed architecture. This represents a limited and operationally satisfactory productivity trade-off relative to the significant gains in governance.

Figure *5* provides a visual comparison of DA, CAR, and RDS across all six evaluated methods. The proposed architecture consistently achieves the highest values on all three metrics.

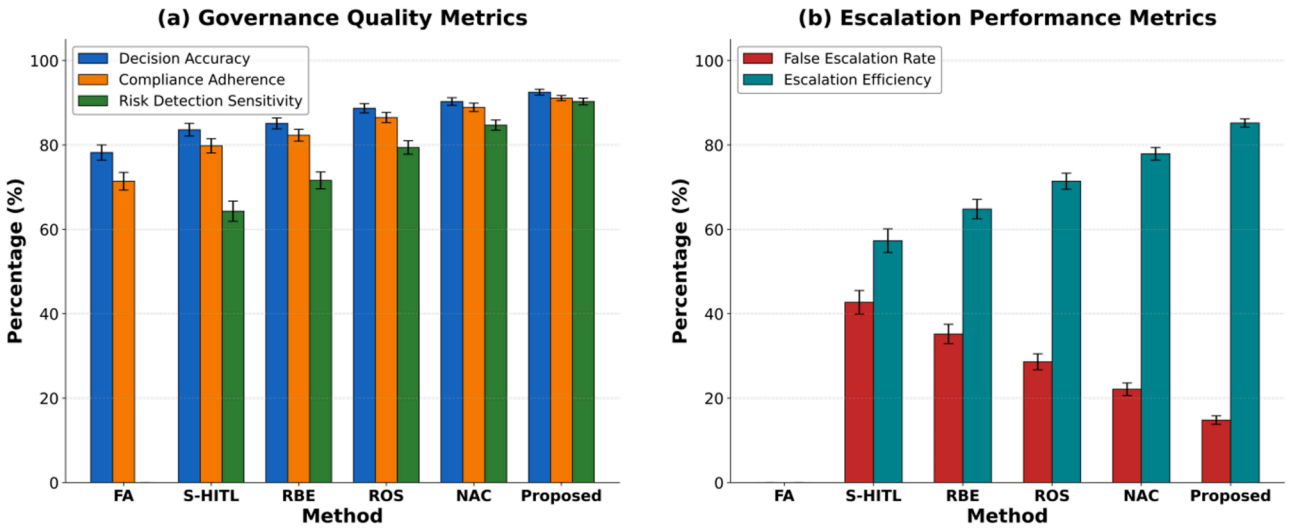

FIGURE 5: Comparative performance visualization across the six evaluation methods for the three primary governance metrics

FA, Fully Autonomous; NAC, Non-Adaptive Cascaded; RBE, Rule-Based Escalation; ROS, Risk-Only Stratification; S-HITL, Static Human-in-the-Loop

Longitudinal feedback learning analysis

As shown in Figure *6*, the governance metrics show a consistent trend of progressive optimization across 30 successive feedback learning cycles. The findings show that there are regular patterns of improvement in all the metrics that have been monitored, with the most significant improvements registered in the initial 10 feedback cycles and then convergence behavior sets in. All four governance metrics improve monotonically across 30 feedback cycles, with the most rapid gains observed in cycles 1 through 10. Beyond cycle 15, the metrics enter a near-plateau region, confirming convergence of the adaptive threshold update rules defined in Equations 11 and 12.

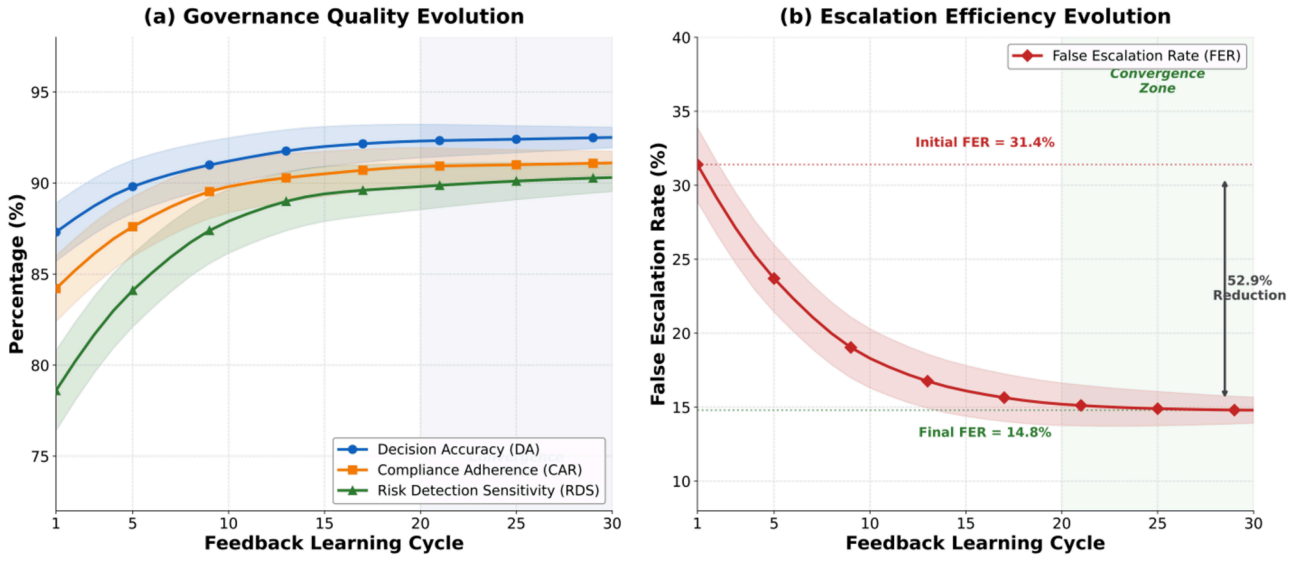

FIGURE 6: Longitudinal evolution of governance metrics across 30 feedback learning cycles, demonstrating progressive governance optimization and convergence behavior

Table *5* reports the metric values at key temporal checkpoints during the longitudinal evaluation.

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

Cycle	DA (%)	CAR (%)	RDS (%)	FER (%)
1 (Initial)	87.3	84.2	78.6	31.4
5	89.8	87.6	84.1	23.7
10	91.2	89.8	87.9	18.3
15	92.0	90.5	89.4	16.1
20	92.3	90.9	89.8	15.2
25	92.4	91.0	90.1	14.9
30 (Final)	92.5	91.1	90.3	14.8

TABLE 5: Governance metric evolution at selected feedback cycle checkpoints

CAR, Compliance Adherence Rate; DA, Decision Accuracy; FER, False Escalation Rate; RDS, Risk Detection Sensitivity

The FER shows the most impressive improvement, reducing the rate by 52.9% on average, from 31.4% at the beginning to 14.8% at the end of the 30 cycles. This incremental optimization attests to the fact that the feedback-based learning module is able to tune escalation rates to distinguish between decisions that actually demand human attention and cases that only constitute routine matters and, as such, do not demand human-level intervention, which addresses the efficiency issues that plague the over-escalating nature of traditional HITL systems.

Domain-specific performance analysis

Table *6* shows disaggregated performance outcomes across the five workforce decision areas and indicates significant differences in governance effectiveness across the operational settings.

Domain	DA (%)	CAR (%)	RDS (%)	FER (%)
HR Operations	93.1	92.3	91.7	13.2
Safety Management	91.4	89.6	88.9	16.8
Financial Compliance	92.8	91.5	90.6	14.1
Resource Allocation	93.5	92.0	91.2	12.7
Performance Evaluation	91.7	90.1	89.1	17.2

TABLE 6: Domain-specific performance of the proposed cascaded governance architecture

CAR, Compliance Adherence Rate; DA, Decision Accuracy; FER, False Escalation Rate; HR, Human Resources; RDS, Risk Detection Sensitivity

Safety Management and Performance Evaluation domains have marginally poorer governance performance, which is explained by their greater complexity, situational variability, and more subtle ethical imperatives. Decision-making in safety management involves multifactorial risk assessment, in which complex interactions among operational parameters may produce non-observable risk situations that are problematic for the anomaly detection ensemble. Decisions in performance evaluation can be characterized by subjective aspects (e.g., qualitative assessment criteria and interpersonal dynamics), which are harder to represent in quantitative risk stratification. These domain-specific results demonstrate the significance of configurable governance parameters that can be adjusted to the unique qualities of any operational circumstance.

Parameter sensitivity analysis

The results of the parameter sensitivity analysis are presented in Table *7*, which shows the impact of key architectural parameters on governance performance.

Parameter Variation	Δ DA	Δ CAR	Δ RDS	Δ FER	Δ PI
α=0.6 (high op. weight)	-0.8	-1.4	+0.6	-1.2	+0.02
β=0.5 (high eth. weight)	-0.3	+1.1	-0.5	+2.1	-0.03
γ=0.4 (high unc. weight)	-1.1	-0.6	+1.3	+3.4	-0.04
η=0.05 (fast learning)	+0.2	-0.4	+0.3	-2.8	+0.01
η=0.001 (slow learning)	-0.5	+0.3	-0.7	+4.1	0.00
W=200 (small window)	+0.4	-0.8	+0.5	-1.7	+0.01
W=1000 (large window)	-0.3	+0.5	-0.4	+2.3	0.00

TABLE 7: Parameter sensitivity analysis results (percentage change from default configuration)

CAR, Compliance Adherence Rate; DA, Decision Accuracy; FER, False Escalation Rate; PI, Productivity Index; RDS, Risk Detection Sensitivity

The analysis indicates that the system is most fragile with respect to the uncertainty weight parameter γ and learning rate η, with superfluous values of either parameter causing significant increases in FERs. The default parameter setting offers a solid balance across all metrics, although domain-specific optimization can provide slight gains for specific governance purposes.

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8

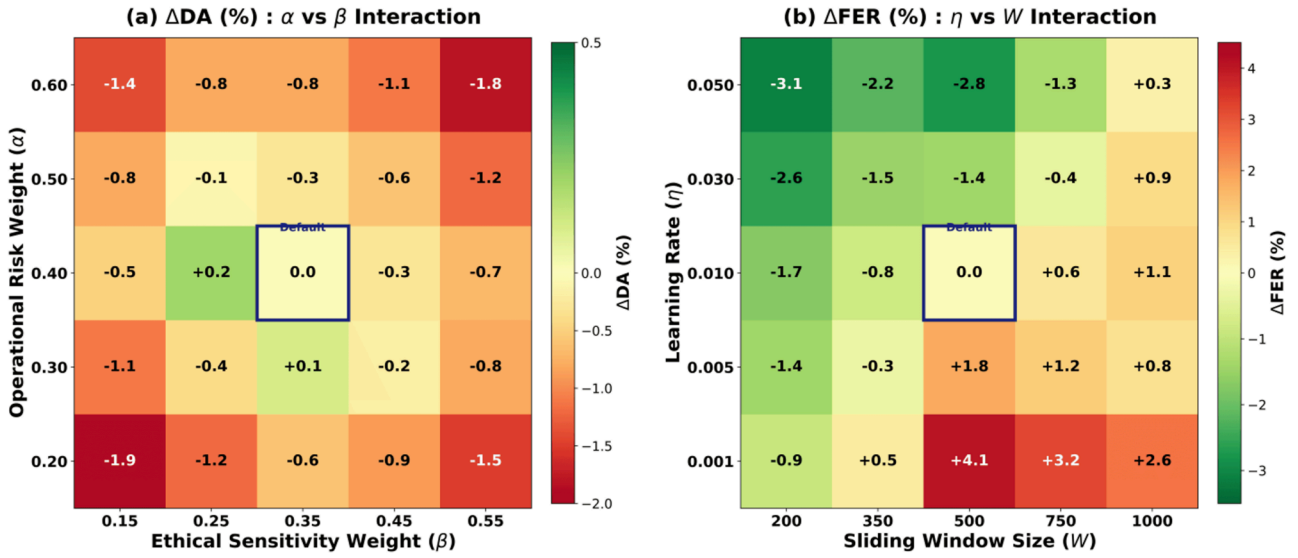

FIGURE 7: Parameter sensitivity heatmap showing interaction effects between key architectural parameters

DA, Decision Accuracy; FER, False Escalation Rate

Figure *7* visualizes the pairwise interaction effects between the seven parameter configurations listed in Table *7*. Darker cells indicate stronger adverse interactions, with the γ-η combination producing the largest joint degradation in FER when both are set to high values simultaneously.

Computational overhead analysis

Table *8* provides a computation of the overhead caused by individual governance layers compared to the baseline of complete autonomy.

Governance Layer	Avg. Latency (ms)	% of Total Overhead
Automation Monitoring	2.3	8.1
Risk Stratification	12.7	44.9
Ethical Compliance	6.4	22.6
Escalation Orchestration	3.8	13.4
Feedback Learning	3.1	11.0
Total Governance Overhead	28.3	100.0

TABLE 8: Computational overhead per decision by governance layer

The overall governance overhead of 28.3 ms per decision is clearly within the tolerable range in the workforce decision environment, with decision latencies on the order of seconds to minutes being the norm. The risk stratification engine has the highest proportion of computational overhead (44.9%) because of ensemble anomaly detection and Monte Carlo Dropout uncertainty estimation, but this cost is justified because it has a significant impact on the accuracy of governance.

Discussion

The empirical findings can be viewed as supporting the main hypothesis that adaptive, cascaded governing frameworks have a high potential to impact the quality, safety, and ethical integrity of AI-enhanced workforce decisions without decreasing operationally viable productivity rates. There are a number of important findings that should be discussed in detail.

The gradual decrease in FERs with the increase in feedback cycles (31.4% to 14.8%) is perhaps the most practical implication of the results, since unnecessary escalations are one of the key operational obstacles to HITL governance implementation in practice [1,5]. The most common argument made by organizations regarding the productivity overhead of human review in governance structures is that the overhead can indeed decrease, while the quality of governance is preserved and even enhanced through feedback-based optimization, which needs to be cited specifically in this regard.

The consistent improvement of the proposed architecture compared to the NAC one across all metrics validates the usefulness of feedback-based learning as an architectural element. The 2.2% DA increment and the 7.3% FER decrease resulting from the feedback module confirm that governance systems can be significantly enhanced by the presence of human correctional signals in the process of policy refinement. This observation is in line with recent developments in reinforcement learning with human feedback on language models [22] and with the application of the principle of human feedback-driven optimization to governance policy areas.

The domain-specific performance differences indicate a useful practical point: governance architectures used within heterogeneous organizational contexts need to be able to be calibrated to domain-specific profiles. The reduced performance in safety management and performance evaluation implies that these contexts can be improved by increasing ethical sensitivity weights (β) or risk thresholds, or by using domain-specific anomaly detectors trained on representative decision distributions [11,13].

The computational overhead analysis indicates that the proposed architecture is computationally feasible within the evaluated simulation environment, with a total governance overhead of 28.3 ms per decision. However, this measurement does not establish end-to-end latency or practical deployability in production environments, where network communication, database access, system integration, queueing, and human review time may contribute additional delays.

Practical implications

Several practical implications of the proposed architecture for organizations interested in implementing an AI-assisted workforce decision system are possible. First, risk-based escalation can enable organizations to allocate human review resources based on the risk of a decision (not every automated decision receives the same level of review). This could apply especially when high-risk decisions are subject to greater human control, while routine decisions are subject to more extensive automation.

Second, the independent ethical compliance layer also offers a way to enforce non-negotiable policy or regulatory limits before a final decision is made. In an actual deployment, the compliance matrix would have to be tailored to the policies, legal requirements, sector-specific regulations, and jurisdictional constraints applicable to the organization. The compliance rules are not necessarily transferable to other domains, as they may differ across jurisdictions.

Third, the reviewer capability-matching mechanism can have organizational workload management implications. There is a need for adequate reviewer capacity, domain expertise, queue management, and measures to prevent automation bias and reviewer overload if it is going to be used on a large scale. Therefore, the reported escalation-efficiency results should not be used to claim a definite reduction in the workload of organizations without operational validation.

Fourth, the scalability of computations is a key deployment concern. The average governance overhead was 28.3 ms for every decision made in the simulated evaluation, but production systems could incur further communication and computational overheads. Optimizing the anomaly detection, uncertainty estimation, compliance checking, data access, and feedback processing will be necessary to scale the framework to high-volume enterprise decision streams.

Lastly, adoption would necessitate governance policies that clarify who has the authority to review decisions, who is held accountable in cases of disagreement between AI recommendations and human decisions, how decisions can be audited, and who has the authority to approve them. The proposed architecture provides a structure that can be implemented to perform these functions, but implementation would need to be carried out in conjunction with existing organizational processes and governance policies.

Limitations

There are a number of significant limitations to this study. First, the experimental data were all constructed. While the simulation allows for a controlled and repeatable test bed for examining the proposed architecture, a synthetic decision-making process does not have the complexity, noise, organizational dependencies, cultural norms, incomplete information, or distributional characteristics of a real decision-making process.

Second, the reference labels used for evaluation were created through a consensus-based simulated expert-labeling protocol. While helpful for establishing an objective benchmark, such labels cannot fully capture the disagreement, subjectivity, context, and variation that can occur among actual reviewers in organizations.

Third, the experimental design is a subset of five workforce decision domains. Other domains might have different distributions of risk, ethical considerations, regulatory requirements, and reviewer practices. Thus, the validity of the observed performance for generalization across industries and organizational settings has yet to be determined.

Fourth, the evaluation of the feedback learning was conducted over a limited number of simulation cycles. In real deployments, concept drift, changes in organizational policies, shifts in workforce behavior, changing regulatory requirements, and modifications to the underlying AI decision-making model would all be potential sources of concept drift. These factors can influence the stability of threshold and detector-weight adaptation.

Fifth, the computational overhead measurement was taken in a controlled hardware and software environment. The latency of production deployment would add a layer of network, database, enterprise software integration, security control, logging, and human review queues.

Conclusions

This paper presented a cascaded HITL governance architecture for ethical AI-augmented workforce systems. The four primary contributions of this work are as follows. First, a formal five-layer governance pipeline was proposed, comprising automation execution monitoring, probabilistic risk stratification via ensemble anomaly detection and Monte Carlo dropout uncertainty estimation, ethical compliance validation through a structured binary compliance matrix, adaptive human escalation orchestration with capability-matched reviewer assignment, and a feedback-driven learning module that refines escalation thresholds and detector weights from human correction signals. Second, governance intensity was formalized as a continuously modulated composite risk score $R_i = \alpha \cdot r_i^{op} + \beta \cdot r_i^{eth} + \gamma \cdot u_i$, enabling context-sensitive and proportionate human oversight that scales with actual decision risk. Third, a feedback-driven escalation optimization mechanism was introduced that reduces false escalation rates by 52.9% over 30 feedback cycles (from 31.4% at initialization to 14.8% at convergence) while simultaneously improving all governance quality metrics. Fourth, comprehensive empirical evaluation across 15,000 simulated workforce decisions demonstrated decision accuracy of

92.5%, compliance adherence of 91.1%, and risk detection sensitivity of 90.3%, each representing statistically significant improvements over fully autonomous and static HITL baselines, while retaining 91% of autonomous system productivity. Taken together, these contributions establish feedback-driven adaptive governance as a practically viable and measurably superior alternative to conventional static oversight models for AI-augmented workforce environments.

Future directions for research are (i) validation with actual workforce decision-related datasets from partner organizations, including longitudinal assessment of workforce decisions over the period of operation; (ii) expansion of the feedback-learning module to multi-stakeholder feedback (e.g., affected employees, compliance officers, and domain experts) instead of single-reviewer signals; (iii) capabilities of large language models to support natural language explanation generation natively coupled with escalation decisions; (iv) the creation of formal safety guarantees to develop feedback-based threshold adaptation with constrained optimization structures; and (v) exploration of federated governance architectures that can support cross-organizational governance learning without violating data privacy. A critical next step is validation using real-world organizational datasets obtained under appropriate privacy, security, and ethical safeguards. Such validation should evaluate not only predictive and governance metrics but also reviewer workload, disagreement rates, decision latency, organizational adoption, fairness across relevant workforce groups, and compliance with applicable regulations. A prospective deployment study would provide stronger evidence regarding whether the improvements observed in simulation transfer to operational workforce environments. The proposed architecture is a step forward in the area of human-centric AI governance by showing that feedback-based, adaptive oversight can simultaneously improve the quality of governance and its efficiency. With the increased dependence of organizations on AI-enhanced workforce systems, it will be necessary to have principled, adaptive governance as a design component rather than a compliance overlay to ensure that intelligent automation is used to leverage the goals of the organization without compromising human control, workforce dignity, and ethical values.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sarat Piridi, Satyanarayana Asundi, Srinivas Kamineni, Chaitanya B. Dadi

Acquisition, analysis, or interpretation of data: Sarat Piridi, Satyanarayana Asundi, Srinivas Kamineni, Chaitanya B. Dadi

Drafting of the manuscript: Sarat Piridi, Satyanarayana Asundi, Srinivas Kamineni, Chaitanya B. Dadi

Critical review of the manuscript for important intellectual content: Sarat Piridi, Satyanarayana Asundi, Srinivas Kamineni, Chaitanya B. Dadi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

Acknowledgements

No Acknowledgement Statement

References

1. Shneiderman B: Human-Centered AI. Oxford University Press, Oxford, UK; 2022. 10.1093/oso/9780192845290.001.0001
2. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascaraán J, Fernández-Leal Á: Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review. 2023, 56:3005-3054. 10.1007/s10462-022-10246-w
3. Green B: The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review. 2022, 45:105681. 10.1016/j.clsr.2022.105681
4. Budhwar P, Chowdhury S, Wood G, et al.: Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. Human Resource Management Journal. 2023, 33:606-659. 10.1111/1748-8583.12524
5. Floridi L, Cowls J: A unified framework of five principles for AI in society. Harvard Data Science Review. 2019, 1:1-14. 10.1162/99608f92.8cd550d1
6. Whyte C: Learning to trust Skynet: Interfacing with artificial intelligence in cyberspace. Contemporary Security Policy. 2023, 44:308-344. 10.1080/13523260.2023.2180882
7. Gomez C, Cho SM, Ke S, Huang C-M, Unberath M: Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. Frontiers in Computer Science. 2025, 6:1521066. 10.3389/fcomp.2024.1521066
8. Falemi A, Akhigbe R, Akin-Oluyomi OT: A conceptual enterprise risk management model that integrates HR strategy with operational analytics. International Journal of Multidisciplinary Research and Growth Evaluation. 2024, 5:1730-1753. 10.54660/.ijmrge.2024.5.6.1730-1753
9. Fasawe O, Akinola AS, Umoren O: Scenario-based resource capacity planning model for optimizing global team performance. International Journal of Advanced Multidisciplinary Research and Studies. 2023, 3:1175-1189. 10.62225/2583049x.2023.3.3.5063
10. Joshi S: AI-augmented human reliability and dependency assessment in financial risk: Implications for systemic stability and regulatory oversight [PREPRINT]. Preprints. 2026, 1-30. 10.20944/preprints202601.1985.v1
11. Xiao X, Zhu H, Liang J, Tong J, Wang H: A comprehensive review of human error in risk-informed decision making: Integrating human reliability assessment, artificial intelligence, and human performance models [PREPRINT]. arXiv. 2025, 10.48550/arXiv.2507.01017
12. Al-Bashrawi MA, Al-Sharafi MA, Elgendy IA, Helal MYI, Anbalagan MK, Chae I, Dwivedi YK: Agentic AI systems and the future of entrepreneurship: a perspective on co-agency, innovation, and ecosystem transformation. International Entrepreneurship and Management Journal. 2026, 22:27. 10.1007/s11365-026-01164-2
13. Hynek N, Gavurova B, Kubak M, Senk M: Expert views on integrating robots, drones, cameras, and AI into critical infrastructure protection and national security: an opportunity for sustainable entrepreneurship. International Entrepreneurship and Management Journal. 2026, 22:37. 10.1007/s11365-026-01177-x
14. Ibitoye JS: Zero-trust cloud security architectures with AI-orchestrated policy enforcement for U.S. critical sectors. International Journal of Science and Engineering Applications. 2023, 12:88-100. 10.7753/ijsea1212.1019
15. Ranjan S, Chembachere D, Lobo L: Introduction to enterprise agentic AI. Agentic AI in Enterprise. Apress, Berkeley, CA; 2025. 1-34. 10.1007/979-8-8688-1542-3_1
16. Pandey B, Patel A: Revolutionizing the Cloud: Generative AI, Security, and Sustainability. Pandey B, Patel A (ed): Springer, Cham; 2026. 10.1007/978-3-032-07479-9
17. Nadler N, Patel Y, Muthukumar A, et al.: Towards human-centric AI regulation: Balancing innovation with societal resilience. SSRN. (2025). Accessed: September 14, 2026: https://www.researchgate.net/publication/398761280_Towards_Human-Centric_AI_Regulation_Balancing_Innovation_with_Soci....

18. Adeyemo AJ: Data-driven corrective and preventive action (CAPA) systems for quality assurance optimization. International Journal of Advanced Multidisciplinary Research and Studies. 2025, 5:907-926. 10.62225/2583049X.2025.5.6.5288

19. Marcellin MC, Pescaroli G, Schmidt M, et al.: A brief introduction to the convergence of artificial intelligence and biotechnology. Biotechnology and AI: Technological Convergence and Information Hazards. NATOARW 2025. NATO Science for Peace and Security Series A: Chemistry and Biology. Cummings CL, Trump BD, Prado V, Ellinport B, Linkov I (ed): Springer, Cham; 2026. 11-55. 10.1007/978-3-032-05246-9_2

20. Barrett JR, Korostelina KV: Resilient systems: AI-mediated communication and frontline public safety. Sustainability. 2026, 18:1071. 10.3390/su18021071

21. Srivastava S, Prajapati A, Tiwari MK, Srivastava AK: Real-time threat detection in AI-integrated zero trust environments. TEJAS Journal of Technologies and Humanitarian Science. 2025, 4:25-37. 10.63920/tjths.44004

22. Wang T, Xu N, Liu F: Revolutionizing biotech and pharmaceutical education with adaptive learning. Frontiers in Education. 2025, 10:1679222. 10.3389/feduc.2025.1679222

23. Pomeroy JK: Generative AI in Healthcare: Strategic, Ethical, and Operational Readiness for Clinical Integration [Dissertation]. Drexel University, Philadelphia, PA; 2025. 10.17918/00010929

24. Dogru O, Xie J, Prakash O, et al.: Reinforcement learning in process industries: Review and perspective. IEEE/CAA Journal of Automatica Sinica. 2024, 11:283-300. 10.1109/jas.2024.124227

25. Veach M: Agentic AI Systems in Professional Domains: A Probabilistic Framework for Role Suitability and Legal Accountability (Thesis). Eastern Michigan University, Ypsilanti, MI; 2025.

26. Taheri Hosseinkhani N: Economic impacts of artificial intelligence integration in Industry 4.0 manufacturing systems. SSRN. (2025). Accessed: September 14, 2026: https://aurora.auburn.edu/bitstream/handle/11200/50712/Economic%20Impacts%20of%20Artificial%20Intelligence%20Integrat....

27. Feng T, Jin C, Liu J: How far are we from AGI: Are LLMs all we need? [PREPRINT]. arXiv. 2024, 10.48550/arXiv.2405.10313

How to cite this article:
Piridi S, Asundi S, Kamineni S, et al. (September 20, 2026) A Cascaded Human-in-the-Loop Governance Architecture With Feedback-Driven Escalation Optimization for Ethical Artificial Intelligence-Augmented Workforce Systems. Cureus J Comput Sci 3 : es44389-026-00284-8. DOI https://doi.org/10.7759/s44389-026-00284-8