Partners Universal Innovative Research Publication (PUIRP)
Volume: 04 Issue: 05 | September - October 2026 | ISSN: 3048-586X | www.puirp.com

Human-in-the-Loop Intelligent Automation: A Review of Trust, Oversight, and Workforce Implications

Dr.A.Shaji George[1], B.Roshini[2]

[1]*Independent Researcher, Chennai, Tamil Nadu, India.*
[2]*Department of Computer Science and Business Systems, KPR Institute of Engineering and Technology, Coimbatore, India.*

Abstract- Intelligent automation is now used to retrieve information, predict, recommend, generate content, route workflows and take physical or administrative action. Human-in-the-loop (HITL) design is often suggested as a possible fail-safe but it does not necessarily mean that a person is always in control. Oversight can be undermined by lack of time, authority, competence, independent evidence, or feedback or by a lack of real accountability despite the title of **oversight**. In this paper we explore HITL intelligent automation from three intertwined perspectives: calibrated trust, operational oversight, and workforce implications. It combines existing human-automation research, guidance for human–AI interaction, risk-management standards, regulation, and evidence of labor. A threat model encompasses technical misfortune, adversarial manipulation, cognitive and organizational failure, and harmful deployment of workforce. The mathematical formulation determines the automation modes based on expected error consequence, uncertainty, review cost, workload and skill preservation. The suggested architecture is that of separating the execution of the model from the risk routing, human decision making, dual authorization, audit, appeal, and learning feedback. A seeded synthetic experiment with 10,000 binary decisions is conducted for the comparisons between automation-only, passive universal review, risk-adaptive HITL, and dual control. With the declared generator, adaptive review achieves a 83.6% to 85.0% accuracy rate, while reviewing 37.1% of cases, and dual control achieves a 87.0% accuracy rate with a 5.2% rate of high impact error. These are indicative values and not measurements taken in the workplace. The review finds that effective HITL systems must have meaningful intervention capability, calibrated reliance, representative evaluation, protected worker voice, continuous competence and evidence of oversight functioning under realistic workload and failure.

Keywords: Human-in-the-loop, Intelligent automation, Human–AI interaction, Trust calibration, Oversight, automation bias, Future of work, Responsible AI.

1. INTRODUCTION

Intelligent automation is a blend of machine learning, rules, optimization, RPA, generative models, and workflow engines to complete or coordinate tasks that used to require humans to make decisions. Applications are document processing, customer service, fraud screening, hiring support, medical triage, industrial operations, software development, finance, logistics and public administration. Automation can help to streamline repetitive tasks, ensure consistency, and enhance access to expertise. It also can amplify mistakes, mask accountability, boost oversight, push for faster work, or eliminate the human factor to identify unusual circumstances.

Human-in-the-loop (HITL) is an attractive solution as it seems to preserve the human judgment and still gain from the speed of computation. The words, however, refer to a number of different, material, arrangements. A person could be someone who labels training data, approves all recommendations, reviews only exceptions, monitors autonomous action, intervenes following an alert, investigates complaints, or governs system design. These jobs vary on the dimensions of information, authority, timing, competence and consequences. A reviewer with five seconds and who can only see the model's preferred answer is not the same as a reviewer who is an accountable professional who can get evidence, override the system and learn from the results.

The classic automation research has proven that automation does not replace, but rather modifies human labor. Parasuraman, Sheridan, and Wickens categorize automation in information acquisition, analysis, decision selection and action implementation, which can vary in degrees of automation [1]. Out-of-the-loop performance issues can arise with high automation monitoring decreases, situation awareness diminishes and manual recovery is challenging. Reliance is very much dependent on trust, especially when users are unable to inspect a complex system completely, safe performance is therefore dependent on the users' level of trust commensurate with the system's capabilities, and not the maximum level of trust [2]. Having a nominal operator doesn't necessarily keep situation awareness or recovery capability [13]. Complacency and automation bias are the result of interacting attentional and task factors, so it is important to consider workload, interface design, feedback and fallback practice during evaluation [14].

These problems are exacerbated by the current AI systems, which can exhibit performance differences across subgroups, contexts, prompts, time and task formulations. Explanations might convince but not allow verification. Generative output may be smooth and not supported. When productivity targets encourage agreement, an individual can turn into a rubber stamp, as well as when disagreement means more work. On the other hand, false alarms and failures that never get resolved can lead to disuse even if automation could be of benefit. NIST's AI Risk Management Framework (AI RMF) views governance as an ongoing lifecycle function and it emphasizes the need for varying human roles and responsibilities in human-AI setups [3]. The European Union AI Act calls for proper human oversight of high-risk AI systems, including the capacity to grasp their limitations and intervene or halt the system's operation [5].

The impact of displacement on the workforce goes beyond that to task composition, autonomy, work intensity, skill, monitoring and learning. For many occupations that are exposed, the ILO believes that transformation is more likely than full automation [10]. The OECD evidence suggests benefits as well as privacy, work intensity and algorithmic-management issues, training and consultation with better reported experiences [11] [12].

There are three questions in this paper. **RQ1**: What is operationally meaningful as opposed to ceremonial when it comes to human oversight. **RQ2**: What is the combination of trust, risk, workload and skill preservation when assigning tasks between humans and automation. **RQ3**: What are the necessary workforce safeguards and evaluation strategies to enable responsible adoption. The paper adds a multidisciplinary review, a sociotechnical threat model, a risk-adaptive formulation, a reference architecture, reproducible evaluation logic and a transparent synthetic study showing trade-offs and organizational data.

II. RELATED WORK

A. *Levels of Automation and Human Control*

The levels-of-automation tradition rejects a binary choice between manual and autonomous operation. Functions can be allocated differently automation may collect information while a human decides, or it may recommend, execute after approval, execute with a veto interval, or act autonomously while reporting. Allocation should reflect reliability, consequence, workload, and human performance [1]. Flexible and adaptive automation can change allocation with context, but dynamic changes can themselves confuse users unless mode, reason, and authority are visible.

Meaningful human control requires more than a user interface. The overseer needs notice that intervention is required, sufficient response time, relevant and independent information, competence to understand limitations, legal and organizational authority, an effective control such as override or safe stop, and protection from retaliation for raising concerns. Post-hoc review remains important for accountability and improvement but cannot prevent an immediate harm. High-impact decisions may therefore require review before action, dual control, or prohibition rather than a general monitoring dashboard.

B. Trust, Reliance, and Human–AI Interaction

Trust is an attitude that an agent will help achieve a goal under uncertainty, whereas reliance is observable behavior. Appropriate reliance occurs when people accept useful automation and reject it when it is wrong. Lee and See argue that trust should correspond to automation performance, process, and purpose [2]. Mistrust can cause disuse; overtrust can cause misuse. Calibration is dynamic because users update beliefs from experience, organizational reputation, interface cues, explanations, and observed failures.

Design guidance for human–AI interaction recommends communicating what a system can do, setting expectations about likely quality, timing services to context, showing relevant contextual information, supporting correction and dismissal, and learning cautiously from user behavior [7]. These practices improve usability but do not guarantee careful judgment. Bucinca et al. found that cognitive forcing interventions can reduce overreliance by requiring users to engage before seeing or accepting AI advice, although such friction has time and usability costs [8]. Explanations should therefore help a person test a claim, locate evidence, and identify uncertainty; a polished narrative alone may strengthen unjustified trust.

NIST identifies automation bias and overreliance as risks in generative-AI configurations and recommends empirical evaluation of human–AI interaction [4]. Appropriate measures include decision accuracy, error detection, override quality, response time, confidence calibration, workload, comprehension, and subgroup outcomes. Satisfaction or stated trust alone is insufficient because a user may like an interface that encourages unsafe acceptance.

C. Oversight, Governance, and Accountability

The NIST AI RMF organizes risk management around Govern, Map, Measure, and Manage, and explicitly recognizes lifecycle roles for people who design, deploy, evaluate, operate, and are affected by AI [3]. ISO/IEC 42001 establishes requirements for an organizational AI management system, including policy, risk processes, competence, operational control, performance evaluation, and improvement [6]. These frameworks support repeatability, but an organization must still define decision rights, evidence, escalation, and remedies for each application.

The EU AI Act provides a legal example of risk-based oversight. Article 14 requires high-risk AI systems to be designed so natural persons can effectively oversee them during use, with measures proportionate to risk, autonomy, and context [5]. Oversight includes understanding capacities and limitations, remaining alert to automation bias, correctly interpreting output, deciding not to use it, overriding it, and stopping operation

where appropriate. Compliance obligations vary by role and implementation date this paper uses the Act as an oversight design reference rather than legal advice.

Accountability can fail through a **moral crumple zone,** in which a nearby human receives blame for a complex automated system despite having limited control [9]. Responsibility should track actual decision authority, system design, resource allocation, and ability to prevent harm. Logs must capture model version, input, output, uncertainty, human action, available evidence, time budget, and downstream outcome, while respecting privacy and avoiding surveillance unrelated to the legitimate purpose.

D. Workforce Implications

Labor effects are task-specific and institution-dependent. Technology may automate portions of a job while creating verification, coordination, relationship, and exception-handling tasks. ILO analysis estimates widespread exposure to generative AI but distinguishes exposure from displacement and identifies transformation as the more likely near-term effect in many occupations [10]. OECD surveys and case studies suggest that AI can improve performance and autonomy for some workers, while algorithmic management, pervasive data collection, reduced human interaction, privacy concern, and higher work intensity can harm job quality [11], [12].

Worker participation matters because frontline employees understand exceptions, hidden work, customer needs, and unsafe targets. Consultation should begin before procurement and continue through monitoring, redesign, and grievance handling. Training must cover domain judgment and system limitations rather than only button operation. Organizations should also protect skill through deliberate practice, rotation, simulations, and manual fallback exercises. Table I summarizes common HITL patterns.

III. SOCIO-TECHNICAL THREAT MODEL

The system contains a data pipeline, model or rule engine, workflow orchestrator, user interface, identity and access services, human operators, managers, auditors, affected persons, and downstream action systems. Assets include decision quality, safety, rights, confidential data, model integrity, worker autonomy and skill, audit evidence, and organizational legitimacy.

Threat actors may be external attackers, malicious insiders, users seeking to game a decision, suppliers, or managers optimizing narrow targets. Unintentional threats are equally important distribution shift, flawed labels, interface ambiguity, fatigue, inadequate training, conflicting incentives, and poorly designed handoffs can generate harm without an adversary. The model assumes that automation is imperfect and that humans have bounded attention, knowledge, and time.

Technical threats include poisoned or unrepresentative data, prompt injection, model extraction, manipulated confidence, insecure tools, unauthorized workflow changes, outage, and hidden model substitution. Decision threats include biased recommendations, confident error, omission of relevant evidence, and inconsistent treatment. Human-factor threats include automation bias, complacency, under trust, alert fatigue, confirmation bias, skill decay, and failure to recognize mode changes. Organizational threats include understaffed review, productivity quotas that discourage escalation, responsibility without authority, inaccessible appeals, retaliatory monitoring, and using worker data for purposes beyond those communicated.

The protection goals are G1, correct and equitable task outcomes within declared bounds G2, intervention that is timely, informed, authorized, and effective G3, calibrated reliance rather than maximum acceptance

G4, traceable accountability across design and operation G5, privacy, contestability, and remedy for affected persons G6, workload and competence sufficient for oversight and G7, workforce transition that includes consultation, training, and monitoring of job quality. HITL cannot guarantee these goals if the organization withholds resources or if both human and automation share the same missing evidence.

IV. MATHEMATICAL FORMULATION

For case i, automation produces action $\hat{y}_i$, confidence pi, uncertainty u_i, explanation evidence e_i, and estimated consequence c_i. Context includes task type z_i, protected or vulnerable-group indicators where lawful and necessary, time budget τ_i, operator workload w_t, and competence k_t. An allocation policy chooses

$$a_i \in \{A, H, R, D, X\}, \tag{1}$$

where A is autonomous action, H is human decision with AI advice, R is mandatory single review, D is independent dual control, and X is prohibited or escalated to a specialist process.

The policy minimizes expected socio-technical loss

$$a_i^* = \arg\min_a \big[\Pr(\text{error} \mid i, a)c_i + \lambda_t T(i, a) + \lambda_w W(w_t, a) + \lambda_f F(i, a) + \lambda_d D(k_t, a) + \lambda_p P(i, a)\big], \tag{2}$$

where T is delay, W is workload, F is fairness disparity or constraint violation, D is expected skill depreciation, and P is privacy or monitoring cost. Hard constraints require

$$\Pr(\text{harm} > h_{\max} \mid i, a) \le \epsilon_i, \tag{3}$$
$$w_t + \Delta w(a) \le w_{\max}, \tag{4}$$
$$\tau_i \ge \tau_{\text{review}}(a), \tag{5}$$
$$\text{authority}(a) \supseteq \text{requiredControl}(i). \tag{6}$$

If meaningful review cannot occur within the available time, the answer is not automatically to accept automation the safe design may reduce scope, delay the action, or prohibit the use case.

Reliance calibration can be evaluated using confidence bins b ∈ B. Let r_b be the fraction of recommendations accepted and q_b their observed correctness. A reliance-calibration error

is

$$\text{RCE} = \sum_{b \in B} \frac{n_b}{N} |r_b - q_b|. \tag{7}$$

Low RCE does not by itself imply good outcomes because people might correctly distrust a poor system, accuracy, consequence, and subgroup performance remain necessary.

Operator competence is modeled illustratively as

$$k_{t+1} = k_t + \eta r_t (1 - k_t) - \delta (1 - r_t) k_t, \tag{8}$$

where r_t is effective deliberate practice, η is learning, and δ is decay. Passive exposure is not assumed to equal practice. The equation is a planning model, not a universal law of human learning.

Table -I: HUMAN-IN-THE-LOOP PATTERNS, BENEFITS, AND FAILURE MODES

Pattern	Human role and timing	Appropriate use	Common failure and safeguard
Manual with AI advice	Human decides after optional recommendation	Novel, ambiguous, or high-consequence tasks where expertise is available	Anchoring on advice; require an initial judgment, evidence access, and recorded rationale for critical cases.
Approval before action	Automation proposes; human approves, edits, or rejects	Moderate/high-impact actions with adequate response time	Rubber stamping under volume targets; sample review quality and cap workload rather than measuring approvals alone.
Exception handling	Automation acts on routine cases; human receives uncertain or policy-sensitive cases	High-volume work with reliable routing and well-defined exceptions	Selective-label blind spots and difficult case concentration; audit non-escalated samples and staff for complexity.
Supervisory control	Automation acts continuously; human monitors and can intervene	Time-sensitive systems with tested alerts and safe fallback	Out-of-the-loop performance and alert fatigue; use meaningful alarms, drills, state visibility, and bounded autonomy.
Dual control	Independent reviewers or human plus protected automated control must agree	Irreversible, regulated, safety-critical, or high-value actions	Shared bias or nominal second approval; preserve independence and define adjudication.
Post-action audit	Human reviews samples, outcomes, complaints, and drift	Low-immediacy harm or as a second layer for all patterns	Harm occurs before review; pair with appeal, remedy, stop triggers, and prospective controls.
Participatory governance	Workers and affected groups influence design, deployment, and evaluation	Organizational and societal impacts extending beyond individual transactions	Consultation without decision influence; document responses, representation, and enforceable escalation rights.

V. REFERENCE ARCHITECTURE

Figure 1 separates business execution from oversight and appeal. Requests enter a policy and risk router that checks permitted purpose, data, identity, impact, confidence, and workload. Low-risk in-scope cases may proceed automatically within limits. Other cases enter a review workspace that presents source evidence, uncertainty, alternatives, and known limitations. High-impact cases require an independent second approval or specialist escalation. Every path creates an audit event. Outcomes, overrides, complaints, and sampled non escalated cases feed evaluation and controlled improvement, not unreviewed self-modification.

Operational controls include separate reviewer authentication, independent high-impact review, confidence shown as uncertainty rather than certainty, and explanations linked to evidence. Queues expose age, impact, and capacity. A stop mechanism freezes affected versions appeals correct outcomes and reveal defects. Workforce governance reviews monitoring purpose, staffing, training, job quality, and redeployment.

VI. WORKFLOW AND ALGORITHM

Figure 2 presents the decision workflow. The person makes an initial assessment before seeing AI advice where anchoring risk is high. The system then shows recommendation, uncertainty, relevant evidence, and known limitations. Agreement does not eliminate verification for high-impact cases. Disagreement triggers reason capture and may require independent adjudication. Outcomes are later linked to the case so both

Partners Universal Innovative Research Publication (PUIRP)
Volume: 04 Issue: 05 | September - October 2026 | ISSN: 3048-586X | www.puirp.com

model and oversight quality can be evaluated. Algorithm 1 implements conservative allocation. Thresholds are tied to action consequence and validated performance, not selected only to minimize labor. Queue overload raises escalation or reduces automated scope it does not silently lower review quality. Protected attributes are used only were lawful and necessary for fairness evaluation or accommodation, with access controls.

VII. EXPERIMENTAL DESIGN

The experiment is a transparent simulation for comparing oversight policies, not a study of employees. Seed 1001 generates 10,000 binary decisions. Twenty percent are labeled high impact. Difficulty $d_i \sim \text{Beta}(2.2, 4.0)$. Automation correctness probability is

$$p_i^A = \text{clip}(0.96 - 0.30d_i - 0.06I_i, 0.50, 0.99), \quad (9)$$

where I_i indicates high impact. Human correctness is $p_i^H =$ clip $(0.90 - 0.18d_i + 0.03I_i, 0.50, 0.97)$. Conditional correctness draws are independent, an intentionally simplifying assumption. Displayed automation confidence equals $p_i^A +_+ \mathcal{N}(0, 0.07)$ clipped to [0.50, 0.99].

Four policies are evaluated. P1, automation-only, always uses the automated answer. P2, passive review, sends every case to a reviewer but models automation bias by selecting the human answer in only 35% of disagreements. P3, adaptive HITL, reviews every high-impact case and cases with confidence below 0.78; after an independent-first workflow, the human answer is selected in 72% of disagreements. P4, dual control, uses P3 and sends high-impact disagreements to an independent expert with correctness clip $(0.96-0.12d_i, 0.75, 0.98)$.

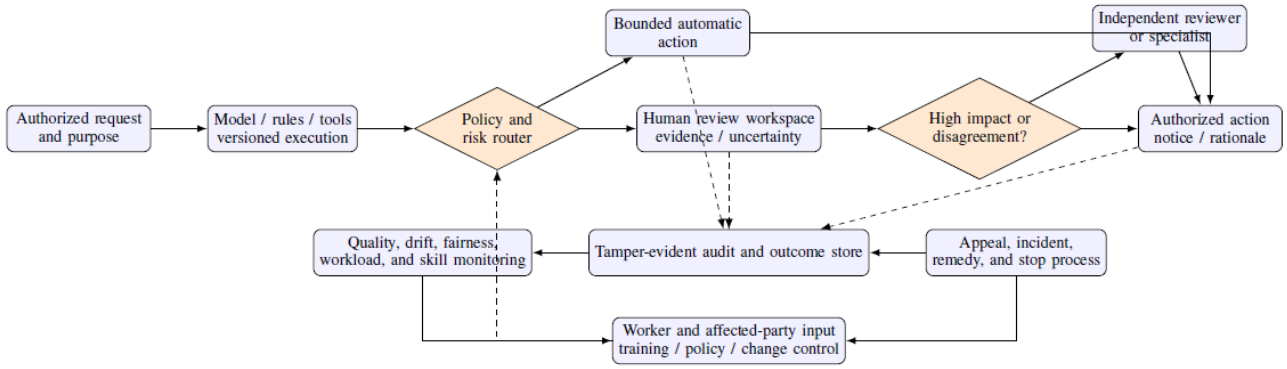


Fig -1: Risk-adaptive HITL architecture with independent oversight, appeal, and workforce feedback.

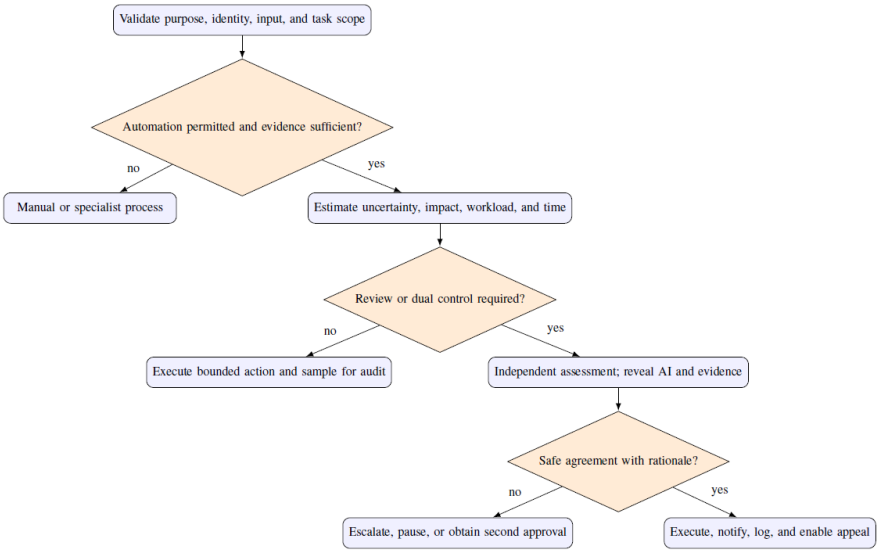


Fig -2: Risk-adaptive decision and escalation workflow

Metrics are total accuracy, error rate among high-impact cases, fraction reviewed, modeled mean cycle time, and accuracy on automation–human disagreements. Automatic processing takes 0.7 time units and primary review adds 4.3; expert adjudication adds 3.2. These are declared costs, not measured labor times. Workforce skill is illustrated with Eq. (8), $k_0 = 0.8$, $\eta = 0.008$, $\delta = 0.003$, and effective deliberate-practice rates 0, 0.10, 0.37, and 0.45 for P1–P4. The lower passive-review rate represents shallow approval rather than genuine practice.

No statistical inference to a workforce is valid. A real study would require preregistered hypotheses, representative tasks and participants, informed consent, privacy protections, compensation, accessible interfaces, subgroup analysis, longitudinal measurement, qualitative worker evidence, and organizational context. It should compare productivity and accuracy with fatigue, autonomy, job quality, learning, appeal, and downstream harm.

VIII. RESULTS

Table II reports the exact seeded output. Automation-only achieves 83.6% accuracy and a 21.5% high-impact error rate. Passive review of every case produces little improvement because the simulated reviewer usually follows automation during disagreement; it also imposes the largest routine workload. Adaptive HITL reviews 37.1% of cases, reaches

Partners Universal Innovative Research Publication (PUIRP)
Volume: 04 Issue: 05 | September - October 2026 | ISSN: 3048-586X | www.puirp.com

Algorithm 1 Risk-adaptive human oversight allocation

Require: Case i, model output o, policy Π, workload w, competence k
Ensure: Allocation a, decision y, evidence record E

if purpose invalid, input unsupported, or prohibited use **then**
  return $(X, \text{pause/escalate}, E)$
end if
Evaluate uncertainty u, consequence c, time τ, and coverage
if c is critical or action is irreversible **then**
  $a \leftarrow D$
else if $u > \theta_u$ or confidence $< \theta_p$ or case is sampled **then**
  $a \leftarrow R$
else
  $a \leftarrow A$ with bounded authority
end if
if $a \in \{R, D\}$ and $(w > w_{\max}$ or $k < k_{\min})$ **then**
  Pause, reassign, or use specialist/manual process
end if
Present evidence, limits, alternatives, and effective controls
Record recommendation, human judgment, timing, and rationale
if unsafe disagreement or invariant failure **then**
  Stop action and escalate independently
end if
Link outcome, complaint, and remedy to immutable case record
return (a, y, E)

85.0% accuracy, and reduces high-impact error to 15.1%. Dual control uses the same primary review fraction plus expert adjudication for high-impact disagreements, reaching 87.0% accuracy and 5.2% high-impact error.

Figure 3 shows the central trade-off. Universal review is not automatically safer than selective review because oversight quality, independence, and attention matter. Dual control performs best here because the generator gives the expert a

Table -II: SEEDED SYNTHETIC RESULTS (NOT WORKFORCE MEASUREMENTS)

Policy	Acc. (%)	HI err. (%)	Review (%)	Time units	Disagr. acc. (%)
Automation	83.6	21.5	0.0	0.7	48.7
Passive review	83.8	18.7	100.0	5.0	49.2
Adaptive HITL	85.0	15.1	37.1	2.3	53.8
Dual control	87.0	5.2	37.1	2.5	61.2

Partners Universal Innovative Research Publication (PUIRP)
Volume: 04 Issue: 05 | September - October 2026 | ISSN: 3048-586X | www.puirp.com

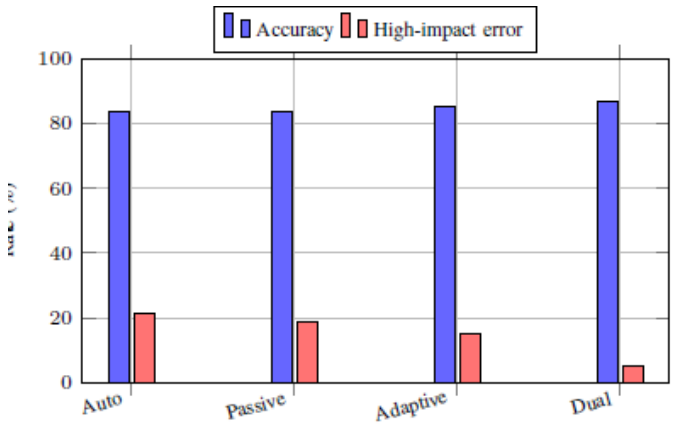


Fig -3: Accuracy and high-impact error in the declared simulation

favorable error profile and focuses that resource on consequential disagreement. Different correlations between human and model errors could reverse the ranking.

The skill model produces end values of 0.592, 0.631, 0.727, and 0.753 for P1–P4 after 100 periods. Figure 4 visualizes the assumed trajectories. It does not show that review necessarily preserves real skill; it demonstrates how workforce outcomes can be included explicitly rather than treated as an externality.

Passive review provides little modeled learning because routine agreement is not deliberate practice.

IX. DISCUSSION

A. Interpretation

The simulation supports a limited conceptual result: allocating scarce attention according to uncertainty and consequence can outperform both no review and indiscriminate low-quality review. It does not establish the numerical value of HITL in any organization. Human and automation errors are often correlated because both depend on the same data, procedures, or stereotypes. Review may also change behavior people learn the model's patterns, adversaries adapt to thresholds, and managers alter staffing after observing throughput.

Trust should be an outcome of trustworthy evidence, not a target optimized through persuasive design. Interfaces should reveal limitations, data provenance, uncertainty, and recourse. Calibration requires periodic exposure to system successes and failures, feedback about outcomes, and an ability to compare advice with independent evidence. Explanations should be tested for whether they improve error detection, not only whether users report satisfaction.

Oversight must be designed for the actual queue. If exceptional cases are harder than routine ones, routing 20%

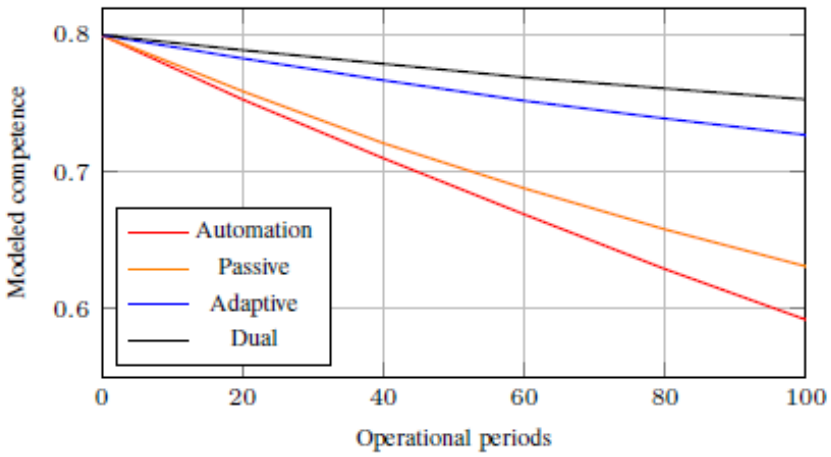


Fig -4: Illustrative competence trajectories under the declared skill model.

of cases may consume more than 20% of cognitive effort. Staffing models should use arrival rates, complexity, interruption, recovery, and required independence. Queue latency and reviewer agreement can be misleading: a fast, unanimous team may be rubber stamping. Random audit of non-escalated cases and observation of override outcomes are needed to detect routing blind spots.

B. Workforce and Organizational Implications

Intelligent automation redistributes tasks. Workers may lose repetitive burdens and gain analytical or interpersonal work, but they may also inherit only ambiguous exceptions, emotional conflict, and responsibility for failures. Exception work should be recognized in job design, staffing, compensation, and career pathways. Training should include model limits, data quality, bias, security, decision rationale, appeal, and manual fallback. Periodic practice should preserve capabilities that are critical during outage or novelty.

Monitoring presents a governance dilemma. Interaction logs are necessary to evaluate oversight, but granular productivity surveillance can reduce autonomy, chill disagreement, and be repurposed for discipline. Data minimization, clear purpose, access controls, retention limits, worker representation, and aggregate evaluation can reduce this risk. Individual overrides should not be treated as poor performance merely because they disagree with the system; their subsequent outcomes should be investigated fairly.

Transition policy should include worker consultation, impact assessment, reskilling time, redeployment, accessible grievances, and attention to vulnerable groups. Productivity alone does not demonstrate responsible implementation; organizations should also monitor job control, pace, psychosocial risk, learning, fairness, and distribution of benefits.

C. Research Directions

First, researchers need stronger measures of *effective oversight*. Binary presence of a human should be replaced by measures of available time, evidence access, competence, authority, intervention success, independence, workload, and remedy. Audits should test these properties under peak load and ambiguous failure.

Second, *appropriate reliance* requires longitudinal study. Trust changes with repeated use, organizational messaging, performance drift, and rare failures. Future experiments should include delayed outcomes, changing reliability, strategic behavior, and recovery after trust violation rather than short, static tasks.

Third, *human–AI team evaluation* should model correlated error and diversity. A second reviewer adds value only if their information or reasoning is sufficiently independent. Research should compare allocation policies using consequence weighted error, subgroup impact, review capacity, and adversarial robustness.

Fourth, *oversight interfaces* should be optimized for verification rather than persuasion. Useful work includes contrastive evidence, provenance, counterfactual tests, uncertainty decomposition, and designs that elicit an initial judgment. Evaluation should measure whether users identify model errors without imposing unacceptable delay.

Fifth, *skill dynamics and job quality* need longitudinal field evidence. Researchers should examine when automation causes deskilling, reskilling, or new expertise, and how task rotation, simulation, mentoring, and feedback alter trajectories. Worker-reported autonomy and stress should be paired with observational and organizational measures.

Sixth, *participatory governance* needs testable methods reporting who participates, what they influence, and whether design or outcomes change. Representation must include people who bear risk. Finally, technical assurance must connect to remedy: affected persons need explanation, challenge, timely correction, and prevention of recurrence. Privacy-protected appeal data can support evaluation.

D. Limitations and Ethical Transparency

This structured narrative review spans safety engineering, psychology, HCI, governance, law, and labor economics; definitions and evidence quality therefore vary. Legal duties depend on jurisdiction, actor, use case, and implementation date. The synthetic experiment uses disclosed artificial distributions and assumes conditionally independent human and automation correctness. It contains no employee, customer, or proprietary data. The skill curve is conceptual, and no numerical result represents measured productivity, safety, employment, or return on investment.

X. CONCLUSION

Human-in-the-loop intelligent automation is not a compromise located halfway between manual work and autonomy. It is a socio-technical control system whose effectiveness depends on allocation, information, timing, competence, authority, incentives, and learning. A person cannot meaningfully oversee an action that they cannot understand, pause, change, or contest. Nor should an organization transfer responsibility to a reviewer while retaining design and staffing choices that make intervention impossible.

The proposed architecture and formulation place consequence, uncertainty, workload, fairness, privacy, and skill alongside accuracy and speed. The synthetic study illustrates why selective, independent, risk-adaptive review can be preferable to universal passive approval, while clearly limiting the claim to the declared model. Responsible implementation requires calibrated reliance, protected escalation, independent high-impact review, appeals and remedies, continuous evaluation, and worker participation.

The central design objective is not to keep a human nominally **in the loop.** It is to create a loop in which human and automated capabilities are complementary, limitations are observable, intervention is effective, responsibility follows control, and workforce benefits and burdens are evaluated as part of system performance.

REFERENCES

[1] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A Model for Types and Levels of Human Interaction with Automation," IEEE Trans. Syst., Man, Cybern. A, vol. 30, no. 3, pp. 286–297, May 2000, doi: 10.1109/3468.844354.

[2] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," Human Factors, vol. 46, no. 1, pp. 50–80, 2004, doi: 10.1518/hfes.46.1.50.30392.

[3] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," National Institute of Standards and Technology, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[4] C. Autio et al., "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," National Institute of Standards and Technology, NIST AI 600-1, July 2024, doi: 10.6028/NIST.AI.600-1.

[5] European Parliament and Council, "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence," Official Journal of the European Union, July 2024, Art. 14.

[6] International Organization for Standardization, ISO/IEC 42001:2023 Information Technology—Artificial Intelligence—Management System, 2023.

[7] S. Amershi et al., "Guidelines for Human–AI Interaction," in Proc. CHI Conf. Human Factors in Computing Systems, 2019, Art. no. 3, pp. 1–13, doi: 10.1145/3290605.3300233.

[8] Z. Buc, inca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AIAssisted Decision-Making," Proc. ACM Hum.-Comput. Interact., vol. 5, no. CSCW1, Art. no. 188, 2021, doi: 10.1145/3449287.

[9] M. C. Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," Engaging Sci., Technol., Soc., vol. 5, pp. 40–60, 2019, doi:10.17351/ests2019.260.

[10] P. Gmyrek et al., "Generative AI and Jobs: A Refined Global Index of Occupational Exposure," International Labour Organization, Working Paper 140, 2025, doi: 10.54394/HETP0387.

[11] OECD, OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market. Paris, France: OECD Publishing, 2023, doi:10.1787/08785bba-en.

[12] M. Lane, M. Williams, and S. Broecke, "The Impact of AI on the Workplace: Main Findings from the OECD AI Surveys of Employers and Workers," OECD Social, Employment and Migration Working Papers, no. 288, 2023, doi: 10.1787/ea0a0fe1-en.

[13] M. R. Endsley, "From Here to Autonomy: Lessons Learned from Human–Automation Research," Human Factors, vol. 59, no. 1, pp. 5–27, 2017, doi: 10.1177/0018720816681350.

[14] R. Parasuraman and D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," Human Factors, vol. 52, no. 3, pp. 381–410, 2010, doi: 10.1177/0018720810376055.