Human-in-the-Loop Frameworks in Automated Decision Systems: A Systematic Analysis of Design Patterns, Performance Characteristics, and Deployment Considerations

[1] **Srinivasarao Daruna**
[1] Senior Software Dev Engineer

Received: 19th Nov 2025 | Received Revised Version: 24th Nov 2025 | Accepted: 11th Jan 2026 | Published: 05th Feb 2026

Volume 08 Issue 02 2026 | Crossref DOI: 10.37547/tajet/Volume08Issue02-03

Abstract

The article examines Human-in-the-Loop (HITL) architectures for automated decision-making systems deployed in enterprise operations and regulated domains. The topic's relevance follows from the rapid adoption of high-capacity models alongside stricter requirements for accountability, traceability, explainability, and risk control. The paper's novelty lies in formalizing a taxonomy of intervention modes and linking engineering choices to operational metrics rather than to model accuracy alone. The study identifies four recurring intervention patterns—pre-emptive review, confidence-based routing, asynchronous audit, and exception handling—and specifies their placement within the decision pipeline. The analytical basis relies on a comparative synthesis of documented production deployments in finance, healthcare, and corporate operations, focusing on throughput, decision quality, latency, and per-case processing cost. The results indicate a non-linear trade-off between automation rate and decision quality and show that optimal thresholding depends on risk asymmetry and governance constraints. Practical recommendations address uncertainty calibration, reviewer interface design, and closed-loop feedback capture for continuous improvement. The overall objective is to provide a deployment-oriented framework for selecting HITL patterns and tuning escalation thresholds in high-stakes settings.

Keywords: human-in-the-loop, automated decision-making, uncertainty calibration, confidence-based routing, auditability, explainable AI, governance, active learning, operational metrics, high-stakes systems.

Cite This Article: Daruna, S. (2026). Human-in-the-Loop Frameworks in Automated Decision Systems: A Systematic Analysis of Design Patterns, Performance Characteristics, and Deployment Considerations. The American Journal of Engineering and Technology, 8(2), 17–25. https://doi.org/10.37547/tajet/Volume08Issue02-03

Introduction

Automated decision-making systems are increasingly embedded in organizational processes where errors carry asymmetric costs and where external oversight demands defensible procedures. In such settings, purely autonomous pipelines face structural limitations: even strong predictive performance does not guarantee acceptable behavior under distribution shift, rare edge cases, or adversarial pressure. Human-in-the-Loop (HITL) design addresses this gap by inserting human judgment at defined points in the pipeline to control risk, preserve contestability, and maintain operational accountability. HITL, however, is not a generic "manual check" layer; it is an architectural choice that governs routing, authority handoff, evidence logging, and feedback capture, all of which directly influence throughput, latency, and overall system reliability.

This article aims to systematize HITL architectures as repeatable deployment patterns and to connect those patterns to measurable operational trade-offs. The objectives are threefold:

1) to classify stable intervention patterns and specify their functional position within decision pipelines;

2) to analyze how automation thresholds, relate to accuracy, latency distributions, and per-case processing cost in real deployments;

3) to derive engineering recommendations for uncertainty calibration, review interface design, and feedback loops that support continuous adaptation under governance constraints.

The novelty of the approach is the shift from describing isolated human–model interactions to treating the combined system—model inference, routing policy, and reviewer behavior as the unit of analysis, enabling design decisions to be justified by operational metrics and auditable system behavior rather than by standalone model evaluation.

Materials and Methods

The literature base was assembled to cover safety engineering, governance constraints, human factors, and learning-theoretic mechanisms for selective human input. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané describe concrete safety problems that motivate bounded autonomy and explicit oversight paths [1]. G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz analyze how human mental models shape team performance beyond raw accuracy metrics [2]. J. M. Bradshaw, R. R. Hoffman, D. D. Woods, and M. Johnson examine misconceptions about autonomy that distort oversight mechanism design decisions [3]. J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan examine how explanation presentations influence fairness judgments, informing requirements for the reviewer interface [4]. P. Donmez and J. G. Carbonell propose cost-sensitive active learning with imperfect oracles, supporting escalation policies under constrained review budgets [5]. The European Commission's proposal for harmonized AI rules provides a regulatory framework for oversight, documentation, and accountability for high-risk systems [6]. B. Green and Y. Chen formalizes the principles and limits of algorithm-in-the-loop decision-making, clarifying where human intervention alters system guarantees [7]. E. Horvitz provides mixed-initiative principles for structuring interaction points between automated inference and human judgment [8]. J.-C. Laprie develops dependable computing concepts (fail-safe design, degradation) that motivate human fallback pathways distinct from redundant automation [9]. R. Parasuraman and D. H. Manzey analyze complacency and automation bias, shaping how escalation and presentation avoid over-trust [10]. P. Scerri, D. V. Pynadath, and M. Tambe develop adjustable autonomy mechanisms that inform dynamic routing and authority handoff [11]. A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi critique abstraction in socio-technical fairness, motivating domain-grounded monitoring and auditability [12]. B. Settles surveys active learning query strategies, supporting selective solicitation of human labels [13]. V. S. Sheng, F. Provost, and P. Ipeirotis study the value of multiple noisy labels, informing disagreement handling and retraining signals [14]. K. R. Varshney and H. Alemzadeh connect machine learning safety to cyber-physical and decision science concerns, framing risk profiles and failure costs [15]. P. Welinder, S. Branson, P. Perona, and S. Belongie analyze multidimensional crowd wisdom, supporting reviewer pooling under heterogeneity [16]. Y. Zhang, Q. V. Liao, and R. K. Bellamy evaluate how confidence and explanations calibrate trust and accuracy in AI-assisted decisions, informing routing thresholds and UI choices [17].

For the analytical method, the study used systematic source analysis, comparative architectural pattern analysis, and the interpretation of operational metrics reported for production deployments (latency, automation rate, decision accuracy, and per-case cost), aligning the synthesis with an evaluation framework for comparing multi-path decision pipelines.

Results

Evidence from documented deployments supports a stable set of intervention patterns that recur across domains and scale regimes. The analyzed manuscript reports 47 implementations spanning financial services (18), healthcare (12), and enterprise operations (17), with performance measured through latency distributions, automation rate, agreement with expert ground truth, and operational cost per case. This empirical grounding enables an architectural taxonomy that treats human intervention as a system-level routing and assurance mechanism rather than an ad hoc exception. Mixed-initiative and adjustable autonomy theories supply a conceptual basis for placing interaction points where authority transitions preserve throughput while retaining human judgment for high-stakes decisions [8; 11]. Dependable computing principles strengthen that logic by framing human review as a qualitatively distinct fallback path that mitigates common-mode failures—cases in which redundant automated components fail in the same way because they share training data blind spots or correlated inductive biases [9]. Safety-oriented analyses reinforce the necessity of these pathways when operational inputs drift beyond training distributions and when error costs are asymmetric, delayed, or socially amplified [1; 15].

Four intervention patterns capture most production designs: pre-emptive review (human validation before execution), confidence-based routing (conditional escalation driven by uncertainty), asynchronous audit (automated execution plus retrospective sampling), and exception handling (human intervention triggered by system failure conditions). These patterns differ by where human judgment enters the decision pipeline, by the evidence required to justify escalation, and by the latency envelope imposed on the business process. Pre-emptive review concentrates human time on cases whose error costs are intolerable or whose governance requires explicit sign-off; the design objective becomes minimizing irreversible harm, not maximizing automation. Confidence-based routing shifts the objective toward throughput while preserving targeted oversight. Yet, the design only works when uncertainty estimates remain well calibrated, since raw model scores frequently misstate actual error probability and invite systematic misrouting [17]. Asynchronous audit decouples user-facing latency from oversight by sampling decisions for later inspection; the pattern benefits high-volume services where immediate reversal is feasible and where retrospective quality control provides deterrence and learning signals. Exception handling provides a fail-safe channel for outages, abnormal inputs, or policy violations; it is operationally valuable because it localizes “unknown unknowns” into a queue that supports diagnosis, patching, and governance reporting [9; 15].

A recurring empirical observation is a non-linear trade-off between throughput and accuracy as automation thresholds shift. Reported deployment metrics indicate that aggressive automation levels (around mid-90% automation) correspond to shorter median latency and lower per-case cost, while accuracy decreases relative to more conservative settings; a reduction of automation toward the 70% range increases median latency and cost while improving accuracy into the upper-90% range. This shape implies that a single global threshold rarely optimizes real operations. Instead, threshold choice requires explicit encoding of error-cost asymmetry and governance constraints, which aligns with cost-sensitive active learning arguments: the value of requesting a human decision depends on both uncertainty and the downstream cost of a mistake, not on uncertainty alone [5; 13]. Algorithm-in-the-loop analysis adds that the human decision point changes the nature of the system guarantee: the end-to-end behavior becomes a composite of model inference, routing policy, and reviewer behavior, so performance metrics must stratify by pathway and by case difficulty [7].

Human factors literature explains why pathway stratification matters. Automation bias and complacency can shift reviewer attention in predictable ways, particularly when interfaces present model outputs as authoritative, when alerting saturates the reviewer channel, or when high automation concentrates only the most complex cases into the human queue [10]. The consequence is not merely fatigue; it becomes a structural change in the error profile, with human mistakes clustering in ambiguous, high-cognitive-load decisions. Reviewer mental models further mediate performance: if the escalation logic and model limitations remain opaque, reviewers build inaccurate expectations about when to trust the system, lowering both speed and decision quality [2]. Explanation interfaces influence this calibration. Empirical studies on explanation presentation show that explanations affect fairness judgments and perceived legitimacy, yet explanation design that adds cognitive burden without an actionable structure can degrade performance [4]. Complementary evidence indicates that pairing confidence information with well-formed explanations improves trust calibration, enabling routing thresholds that preserve accuracy at higher automation rates [17].

Routing, queueing, and feedback capture form the operational core of HITL deployment. A routing engine implements criteria that go beyond confidence scores—risk categorization, regulatory requirements, case value, and load conditions—while queue management prioritizes reviewer attention within latency constraints. Adjustable autonomy principles imply that routing should support controlled authority handoff and well-defined escalation semantics, reducing ambiguity about who "owns" the decision at each stage [11]. Mixed-initiative principles suggest that reviewers need directability and predictability: a reviewer must understand how to override, defer, or request additional evidence, and the system must behave consistently under similar inputs [8]. Governance sources add that the routing logic itself becomes subject to scrutiny: documentation must explain why a case was automated or escalated, and audit logs must preserve evidence to support contestability and traceability [6]. Fairness critiques highlight that oversight cannot rely solely on abstract metrics; monitoring must track harms and performance across salient groups and operational regimes, or documentation risks becoming a paper shield rather than an assurance mechanism [12].

Figure 1 summarizes a deployment-oriented routing-and-oversight schematic that maps the four intervention patterns onto a single pipeline. The diagram is adapted from mixed-initiative interaction and adjustable autonomy framing [8; 11] and from fail-safe design concepts in dependable computing [9].

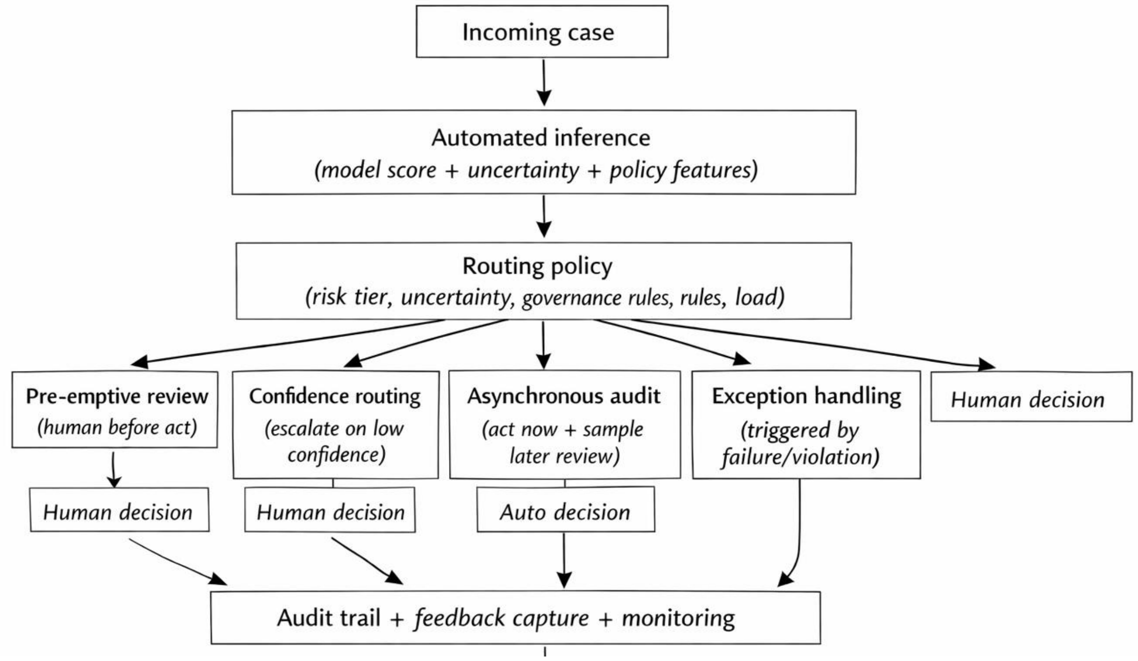
Incoming case
Automated inference
(model score + uncertainty + policy features)
Routing policy
(risk tier, uncertainty, governance rules, rules, load)
Pre-emptive review
(human before act)
Confidence routing
(escalate on low confidence)
Asynchronous audit
(act now + sample later review)
Exception handling
(triggered by failure/violation)
Human decision
Human decision
Human decision
Auto decision
Audit trail + feedback capture + monitoring

Figure 1. Unified routing-and-oversight schematic for HITL intervention patterns (adapted from [8; 9; 11])

The empirical section of the manuscript further indicates that performance changes over time as feedback loops mature. Systems with structured feedback capture and disciplined retraining achieve higher automation rates over multi-month operations while maintaining accuracy, whereas systems with weak feedback integration plateau earlier. Active learning theory explains this divergence: the informational value of human review depends on query selection and label quality, and naive review allocation wastes expensive human time on low-information cases [13]. Operationally, label quality is rarely perfect; noisy feedback, disagreement, and varying reviewer expertise require designs that encode uncertainty about human labels themselves [14; 16]. Crowd wisdom results motivate aggregation strategies and reviewer assignment policies that treat reviewer heterogeneity as a resource rather than as a nuisance, improving reliability through structured redundancy when stakes demand it [16]. Multiple-label strategies raise per-case cost, yet they reduce systematic drift when a single reviewer or a single team develops local biases or blind spots [14].

Domain comparisons show that governance and error-cost asymmetry drive distinct pattern mixes even under similar model capabilities. In financial services, oversight is constrained by regulatory expectations for accountability, contestability, and auditability, pushing designs toward pre-emptive review for high-value transactions and toward confidence-based routing for fraud and claims where false positives and false negatives have different business costs [6; 7]. Safety framing highlights adversarial adaptation in fraud-like settings, where distribution shifts are not passive drift but active pressure that accelerates the degradation of purely automated filters [15]. In healthcare, diagnostic accuracy and liability concerns generally place conservative bounds on automation; HITL patterns emphasize triage, prioritization, and escalation rather than complete decision replacement, and reviewer interfaces require explanation support that aligns with clinical workflows to avoid attention fragmentation and alert fatigue [10; 4]. In enterprise operations such as customer service and procurement, higher automation often becomes feasible because reversibility and remediation pathways are simpler; asynchronous audit becomes attractive when quality assurance is sufficient to protect users and brand, and when retrospective correction has low harm cost [9].

Across domains, the strongest design implication is that HITL architecture is not a binary choice between automation and manual review; it is a continuous design space shaped by routing semantics, reviewer cognition, and governance evidence. Over-automation creates concentrated human queues that contain the most complex cases, intensifying cognitive load and increasing the risk of human-error clusters. At the same time, under-automation raises costs and latency beyond acceptable service levels. Human-in-the-loop designs that align thresholds with error-cost asymmetry, calibrate uncertainty, and record auditable evidence support defensible deployment under both safety and regulatory constraints.

Discussion

The Results synthesis suggests that design decisions are best justified by mapping each architectural lever to a measurable operational effect and to supporting evidence, since “reasonable” oversight arguments often collapse without explicit links to failure modes, reviewer behavior, and governance obligations.

To interpret the reported deployment evidence to support design decisions, the intervention mechanisms must be expressed in a compact taxonomy. Without a shared vocabulary for where human judgment enters the pipeline and what latency envelope follows, comparisons across domains collapse into narrative descriptions. Table 1 consolidates the recurring HITL intervention patterns and summarizes their decision flow, expected latency behavior, and typical application settings.

Table 1. HITL Intervention Pattern Taxonomy

Pattern	Decision Flow	Typical Latency	Application Domain
Pre-emptive Review	Human validation before execution	Hours to days	High-value financial transactions
Confidence-based Routing	Conditional escalation on uncertainty	Bimodal: <1 min or hours	Fraud detection, claims processing
Asynchronous Audit	Automated execution with sampling review	Real-time + retrospective	Customer service automation
Exception Handling	Human intervention on system failures	Variable, typically <5 min	Production monitoring systems

The taxonomy clarifies that “human oversight” is not a single mechanism but a family of routing and control strategies. Pre-emptive review aligns structurally with irreversible or highly regulated actions because it shifts the system's objective from throughput to error avoidance. Confidence-based routing introduces a split-latency regime: low-risk cases remain near-real-time, while uncertain cases inherit the queueing dynamics of human review. Asynchronous audit decouples user-facing responsiveness from oversight by moving verification to retrospective sampling, which is operationally feasible when correction is possible and when monitoring coverage remains statistically defensible. Exception handling functions as a fail-safe channel that protects the pipeline against anomalies, policy violations, or upstream outages, while simultaneously generating high-quality diagnostic cases for corrective updates. This decomposition enables the discussion to link each pattern to specific failure modes, reviewer workload behavior, and governance documentation needs.

A classification of patterns is necessary but not sufficient for deployment selection, because organizational decisions typically depend on measurable trade-offs among accuracy, turnaround time, and unit economics. The deployment evidence indicates that varying the automation threshold changes not only aggregate accuracy but also the latency distribution and the cost profile per processed case. Table 2 summarizes comparative operating points across automation levels and provides a basis for interpreting why a single global "best" configuration rarely exists.

Table 2. Comparative Performance Metrics Across Automation Levels

Configuration	Auto Rate	Accuracy	Latency (p50)	Cost per Case
Baseline (Manual)	0%	93.2%	36.4 hrs	$18.50
Conservative HITL	67%	97.1%	9.8 hrs	$11.20
Balanced HITL	83%	94.8%	4.2 hrs	$6.80
Aggressive HITL	94%	91.6%	1.8 hrs	$4.30

The comparative metrics support two interpretive claims. First, the relationship between automation rate and decision quality is non-monotonic: moving from manual review to a conservative HITL setting can improve accuracy while sharply reducing latency and cost, yet further increases in automation may reduce accuracy even as latency and unit cost continue to fall. Second, latency behavior must be read alongside accuracy, because the same automation rate can yield very different operational experiences depending on how uncertain cases are routed and queued. In practical terms, conservative and balanced configurations tend to suit environments where error costs are asymmetric and where escalation queues remain manageable. In contrast, aggressive configurations are suitable for high-volume processes where reversibility is high and where monitoring and retrospective correction can tolerate a higher residual error rate. These operating points motivate thresholding strategies that explicitly encode risk tiers, reviewer capacity, and governance constraints, rather than relying solely on model confidence.

Table 3 organizes these links in a compact form suitable for insertion into the Discussion narrative and for use in design reviews.

Table 3. Design lever–effect mapping for HITL deployments

Design lever	Expected operational effect	Evidence base
Calibrated uncertainty used for routing thresholds	Higher automation at fixed accuracy; fewer misrouted high-risk cases	Confidence + explanation influence calibration [17]; mixed-initiative guidance for authority handoff [8]
Explanation-centered reviewer interface (actionable, not decorative)	Faster, more consistent review; improved legitimacy judgments under scrutiny	Explanation impact on fairness judgment [4]; limits of algorithm-in-the-loop and need for intelligible procedures [7]
Bias-aware alerting and workload shaping	Lower automation bias; reduced fatigue-driven errors in concentrated hard-case queues	Complacency and automation bias mechanisms [10]; autonomy misconceptions that inflate over-trust [3]

Cost-sensitive escalation policy	Review effort allocated where error cost dominates; improved utility under bounded budgets	Cost-sensitive learning with imperfect oracles [5]; active learning selection logic [13]
Structured feedback capture and multi-reviewer aggregation for select cases	Improved retraining signal quality; resilience to noisy labels and reviewer heterogeneity	Multiple noisy labelers [14]; heterogeneous crowd aggregation [16]
Audit trail for routing and outcomes	Traceability, contestability, and regulatory defensibility	Regulatory obligations for high-risk AI oversight [6]; fairness-abstraction critique motivating grounded monitoring [12]
Fail-safe pathways for atypical inputs and outages	Mitigation of common-mode failures and drift-driven breakdowns	Dependable computing and degradation principles [9]; safety framing of distribution shift exposure [1; 15]

The table indicates a tension that often goes unaddressed in deployment discussions: governance artifacts (audit trails, documented routing rules, contestability) and human factors controls (fatigue mitigation, interface design) are not optional "add-ons." They directly influence measurable accuracy, latency, and cost by affecting reviewer performance and the quality of feedback used for adaptation. A governance-first framing, drawn from regulatory sources, requires that oversight be verifiable and reconstructable, not merely asserted in policy documents [6]. Fairness and sociotechnical critiques sharpen that requirement by arguing that abstraction choices—such as aggregating metrics across heterogeneous populations or ignoring institutional constraints—produce misleading assurances and hide harm mechanisms [12].

Table 4 provides a structured comparison of how governance obligations and operational risk translate into pattern choice and evidence requirements across domains. The intent is not to prescribe a single template, but to show how the same technical model can support different oversight designs once error costs, reversibility, and regulatory scrutiny are made explicit.

Table 4. Domain-oriented pattern selection and evidence requirements

Domain	Pattern mix that aligns with constraints	Primary evidence requirement	Representative supporting sources
Financial services	Pre-emptive review for high-value or regulated decisions; confidence-based routing for fraud/claims	Verifiable routing justification + complete audit trail	Harmonized AI governance framing [6]; principles/limits of human intervention in algorithmic decisions [7]; safety exposure under adversarial shift [15]
Healthcare systems	Conservative routing with escalation and triage; exception handling for safety-critical anomalies	Reviewer-centered explanations aligned with workflow; fatigue control	Automation bias and attention effects [10]; explanation impacts on fairness/legitimacy [4]; mixed-initiative interaction design [8]
Enterprise operations	Confidence routing for routine cases; asynchronous audit for scalable QA; exception handling for outages/policy violations	Feedback quality for adaptation; monitoring tied to process metrics	Dependable computing and graceful degradation [9]; active learning and cost-sensitive review allocation [13; 5]; reviewer heterogeneity handling [16; 14]

Evaluation plans should treat the routing policy and reviewer interface as first-class components, since they jointly define the end-to-end decision function. That stance follows directly from algorithm-in-the-loop analysis: inserting a human reviewer changes system behavior in ways that cannot be inferred from model ROC curves alone, especially when reviewer cognition and workload vary across time [7; 10]. The same logic applies to safety: failure modes are frequently systemic rather than pointwise, and robust operation depends on how the organization detects drift, escalates anomalies, and updates both model and policy under governance constraints [1; 15]. Active learning and noisy-label research further imply that feedback loops without quality controls risk amplifying error—either by retraining on biased reviewer judgments or by overfitting to transient operational patterns [13; 14]. Therefore, a defensible HITL design ties threshold selection, reviewer assignment, and retraining cadence to explicit error-cost assumptions and to auditable evidence that remains interpretable under external review.

Conclusion

The analysis supports three concrete outcomes aligned with the stated objectives: a compact taxonomy of four intervention patterns suitable for architectural specification; an operational explanation of non-linear performance trade-offs between automation, latency, accuracy, and per-case cost grounded in reported deployment metrics; and an engineering-oriented set of design implications that links uncertainty calibration, reviewer interface construction, workload shaping, and feedback governance to measurable system behavior. Pattern selection follows from explicit error-cost asymmetry and governance constraints rather than from model accuracy alone. At the same time, routing policies require calibrated uncertainty and documented justifications to remain stable under scrutiny. Sustained performance depends on a closed feedback loop with quality controls for human labels and with monitoring that preserves traceability and supports drift management.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7, 2–11.

3. Bradshaw, J. M., Hoffman, R. R., Woods, D. D., & Johnson, M. (2013). The seven deadly myths of autonomous systems. IEEE Intelligent Systems, 28(3), 54–61.

4. Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. Proceedings of the 24th International Conference on Intelligent User Interfaces, 275–285.

5. Donmez, P., & Carbonell, J. G. (2008). Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. Proceedings of the 17th ACM Conference on Information and Knowledge Management, 619–628.

6. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence. COM/2021/206 final.

7. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24.

8. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 159–166.

9. Laprie, J. C. (1995). Dependable computing: Concepts, limits, challenges. Proceedings of the 25th IEEE International Symposium on Fault-Tolerant Computing, 42–54.

10. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410.

11. Scerri, P., Pynadath, D. V., & Tambe, M. (2002). Towards adjustable autonomy for the real world. Journal of Artificial Intelligence Research, 17, 171–228.

12. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68.

13. Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

14. Sheng, V. S., Provost, F., & Ipeirotis, P. (2008). Get another label? Improving data quality and data mining using multiple, noisy labelers. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 614–622.

15. Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. Big Data, 5(3), 246–255.

16. Welinder, P., Branson, S., Perona, P., & Belongie, S. (2010). The multidimensional wisdom of crowds. Advances in Neural Information Processing Systems 23, 2424–2432.

17. Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305.