

Human-in-the-Loop Control for High-Risk Legal Agents

Shatrunjay Kumar Singh¹

¹Bloomberg LP

Publication Date: 2026/10/05

Abstract: Agentic artificial intelligence can decompose legal objectives, retrieve authorities, invoke external tools, draft documents, communicate with stakeholders, and execute multi-step workflows. In high-risk legal settings, a final approval click is not an adequate control: errors can propagate across a trajectory, evidence can be withheld from the reviewer, confidential data can cross an unauthorized boundary, and a human may be asked to approve an artifact that no longer matches the version later executed. This paper proposes Human-in-the-Loop Control for Legal Agents (HILC-Legal), a risk-adaptive governance architecture that separates language generation from authorization and separates authorization from execution. The architecture combines hard policy constraints, a calibrated action-risk score, evidence packets, competence-aware reviewer routing, version-bound authorization tokens, a least-privilege execution broker, provenance logging, and post-action monitoring. We define six control invariants for legally consequential agent behavior and a four-tier oversight policy ranging from bounded read-only assistance to human-in-command decision making. We also specify an evaluation protocol covering unsafe action escape, authority and evidence fidelity, confidentiality preservation, approval-version consistency, reviewer correction yield, automation bias, workload, calibration, and recovery. The proposal is a design and evaluation framework; it does not claim deployment results. Its central argument is that meaningful human oversight is an engineered control property - requiring information, competence, authority, time, and effective intervention mechanisms - rather than the mere presence of a person in a workflow.

Keywords: Human-in-the-Loop, Legal Agents, Agentic AI, High-Risk Legal Work, Human Oversight, Legal Ethics, Risk-Adaptive Control, Auditability, Privilege, Authorization.

How to Cite: Shatrunjay Kumar Singh (2026) Human-in-the-Loop Control for High-Risk Legal Agents. *International Journal of Innovative Science and Research Technology*, 11(8), 4229-4237. <https://doi.org/10.38124/ijisrt/26aug422>

I. INTRODUCTION

Large language model systems are evolving from conversational assistants into agents that plan, remember, retrieve, call application programming interfaces, manipulate documents, and act in external systems. A legal agent may identify issues, assemble case law, classify discovery material, revise a contract, prepare a filing, schedule a deadline, or draft a communication to a client or counterparty. These capabilities can improve access and productivity, but they also change the safety problem. A mistaken sentence can become a mistaken tool call; an incorrect classification can propagate through an entire review; and an unauthorized transmission can create consequences that cannot be undone.

Legal work is particularly sensitive to this shift because correctness is contextual and responsibility remains human. A proposition may be accurate in one jurisdiction but wrong in another; a document may be discoverable yet privileged; a draft may be harmless until sent, filed, signed, or relied upon. Empirical work also indicates that retrieval-grounded legal research tools can still produce incorrect or misleading responses, so the existence of citations or retrieval does not remove the need for verification (Magesh et al., 2025). For

agentic systems, the relevant question is therefore not only whether the generated text is plausible, but whether each proposed action is authorized, evidenced, reversible, and attributable.

Many proposed safeguards reduce human oversight to a single endpoint: the model prepares an output and a person clicks approve. This nominal pattern can fail when the reviewer sees only a polished summary, lacks the relevant subject-matter competence, faces severe time pressure, cannot inspect the source record, or has no practical way to stop or reverse the system. Human-1 factors research further cautions that explanations can increase acceptance of an AI recommendation even when the recommendation is wrong (Bansal et al., 2021). Effective oversight must therefore be designed as a control system rather than treated as an organizational slogan.

➤ *This Paper Makes Four Contributions:*

- *Control Invariants:*

We define non-negotiable properties for external legal effect, evidence traceability, protected-data containment, version fidelity, competent authorization, and recovery.

- *Risk-Adaptive Architecture:*

We propose HILC-Legal, which places a policy and authorization control plane between an agent and the legal environment.

- *Formal Oversight Policy:*

We combine a calibrated action-risk score with hard constraints that cannot be lowered by model confidence or convenience.

- *Evaluation Protocol:*

We specify system, legal-quality, security, and human-factor metrics, together with baselines, ablations, adversarial scenarios, and acceptance gates. The term high-risk is used here operationally to denote legal tasks with material consequences, irreversibility, sensitive information, or effects on a person or institution. It is not a conclusion that every such deployment falls within any particular statutory category. The framework is jurisdiction-agnostic by design and must be localized by qualified counsel, security personnel, and domain owners before use.

II. BACKGROUND AND RELATED WORK

➤ *From Decision Support to Legal Agents*

Traditional legal technology assists a user at a bounded step: search, proofreading, citation checking, or document assembly. An agent instead selects intermediate goals and tools across a sequence. This creates trajectory risk. An early error in issue framing can distort retrieval; distorted retrieval can produce a false legal rule; and that rule can drive drafting or execution. Sinha et al. (2026) describe a related propagation problem in AI-assisted legal discovery, where an early misclassification can contaminate later privilege-review steps. The control object is thus the complete action trajectory, not merely the final paragraph.

A useful distinction is among human-in-the-loop (a person must act before a step proceeds), human-on-the-loop (a person monitors and can intervene), and human-in-command (a person retains the substantive decision and defines whether, when, and how automation is used). Levels of-automation research shows that allocation should be matched to task consequences and human capabilities rather than maximized by default (Parasuraman, Sheridan, & Wickens, 2000). HILC Legal uses all three patterns but assigns them to different risk tiers.

➤ *Conditions for Effective Human Oversight*

A human checkpoint is effective only when the reviewer has five properties: sufficient information about the proposed action and its evidence; competence for the issue and jurisdiction; authority to reject, revise, stop, or escalate; adequate time and interface support; and an intervention mechanism that reliably changes system behavior. Interface guidelines for human-AI interaction similarly emphasize showing system status, supporting efficient correction, communicating uncertainty, and enabling users to control the system (Amershi et al., 2019). Chiodo et al. (2025) further argue that different human-in-the-loop structures have materially different technical and responsibility properties;

placing a person at an endpoint does not automatically make a system safe or accountable.

Oversight also has an epistemic problem. A polished explanation can create unwarranted confidence, while an undifferentiated dump of logs can overwhelm the reviewer. The design objective is not maximum explanation volume. It is decision-relevant evidence: what action is proposed, which legal and factual assumptions matter, what sources support each material claim, what contrary information exists, what changed, and what consequence follows from approval.

➤ *Professional and Regulatory Expectations*

Professional obligations supply strong reasons for engineered controls. American Bar Association Formal Opinion 512 addresses competence, confidentiality, communication, supervision, candor, and reasonable fees when lawyers use generative AI. Its practical implication for agentic systems is that a lawyer cannot delegate professional judgment merely because a tool generated a well-formed answer. Review depth should be matched to the tool, task, and consequence (American Bar Association, 2024). HILC-Legal operationalizes this principle through risk-based routing and evidence requirements.

Article 14 of the European Union Artificial Intelligence Act requires human-oversight measures for high-risk AI systems to be commensurate with risk, autonomy, and context, and contemplates the ability to understand limitations, avoid automation bias, interpret outputs, disregard or override them, intervene, and stop operation where appropriate (European Parliament and Council, 2024). Not every legal agent is necessarily classified as a high-risk AI system under that regulation; the provision is used here as a design benchmark for effective oversight. NIST AI RMF 1.0 and its Generative AI Profile similarly support lifecycle governance, mapping, measurement, management, and documented accountability (Tabassi, 2023; Autio et al., 2024).

➤ *Reliability, Grounding, and Human Factors*

Retrieval-augmented generation can reduce some hallucinations, but it does not guarantee legally valid analysis. In a preregistered evaluation, Magesh et al. (2025) reported hallucination rates between 17 percent and 33 percent for the commercial legal research systems they tested at that time. Their typology separates correctness from groundedness: an answer may be factually wrong, may cite an inapplicable source, or may assert that a source supports a proposition it does not support. A control system must therefore expose claim-to-source mappings and jurisdictional context rather than treat retrieval as a binary safety certificate.

The human reviewer can also fail. Reviewers may defer to confident language, miss changes between versions, approve under deadline pressure, or assume that another reviewer checked the evidence. These are predictable system conditions rather than individual moral defects. The architecture should make correct review easier, inappropriate approval harder, and responsibility traceable without using the human as a ceremonial liability shield.

➤ Research Gap

Existing guidance generally states that a human should review AI output, while technical agent frameworks often focus on planning, tool use, and task completion. The missing layer is an enforceable control plane that decides when autonomy is permitted, assembles the evidence needed for review, binds approval to the exact artifact, restricts execution, and monitors consequences. HILC-Legal addresses that gap.

III. PROBLEM DEFINITION AND THREAT MODEL

➤ System Model and Legally Consequential Actions

Let m denote a legal matter with jurisdiction, parties, deadlines, data classifications, assigned roles, and policy constraints. At state s , an agent proposes action a , which contains an artifact, a target system, requested permissions, supporting evidence, and an expected effect. The action may be internal and reversible, such as generating a research outline, or externally consequential, such as filing a document, sending advice, disclosing records, accepting a term, waiving an objection, or changing a rights-affecting classification.

A legally consequential action is any action that can materially affect a client, tribunal, counterparty, employee, data subject, or institutional obligation. Consequence is determined by context, not by the surface form of the action. A draft email is low consequence while stored internally, but the same text may become high consequence when addressed to a client and transmitted. HILC-Legal therefore evaluates proposed execution, not merely content type.

➤ Failure Classes

• Epistemic Failure:

Fabricated facts or authorities, unsupported legal propositions, omitted adverse authority, false confidence, or incorrect synthesis.

• Context and Authority Failure:

Wrong jurisdiction, date, procedural posture, client objective, governing law, or treatment of precedent.

• Agency and Tool Failure:

Incorrect action sequencing, excessive permissions, duplicate execution, wrong recipient, stale form, or unintended commitment.

• Confidentiality and Security Failure:

Privilege leakage, cross-matter contamination, prompt injection, secret exposure, or unauthorized data transfer.

• Human-Factor Failure:

Automation bias, inadequate reviewer competence, alert fatigue, ambiguous responsibility, or approval under insufficient time.

• Governance and Change Failure:

Unrecorded model, prompt, corpus, or policy changes; missing provenance; or inability to reconstruct why an action occurred.

The threat model includes ordinary error, malicious or compromised content in retrieved material, misuse by an authorized user, and failures created by system updates. It assumes that a base model can be wrong and that a human can be wrong. Safety is therefore obtained through independent controls, constrained authority, and recoverability rather than trust in a single component.

➤ Control Invariants

HILC-Legal defines six invariants that must hold regardless of the model score. They are expressed as auditable system properties rather than aspirational principles.

Table 1 HILC-Legal Control Invariants and their Enforcement Mechanisms. 4. HILC-Legal Control Architecture

Invariant	Required Property	Primary Enforcement Mechanism
External-effect authorization	No externally consequential action executes without a scoped authorization appropriate to its risk tier.	Policy gate, human review, authorization token, execution broker.
Evidence traceability	Every material legal or factual proposition presented for approval is linked to inspectable evidence or explicitly marked unsupported.	Claim-evidence map, source snapshot, contradiction and citation checks.
Protected-data containment	Privileged, confidential, personal, or restricted data stays within authorized matter and system boundaries.	Matter isolation, data labels, egress rules, least-privilege tools.
Version fidelity	Approval applies only to the exact artifact, instructions, recipients, and permissions reviewed.	Cryptographic hash, immutable approval record, invalidation on change.
Competent and empowered review	The reviewer has the role, subject competence, authority, and time needed to make the required decision.	Role and competence routing, workload limits, stop and escalation controls.
Recovery and accountability	Actions, overrides, outcomes, and changes are reconstructable; reversible actions can be rolled back or contained.	Provenance ledger, runtime monitor, incident workflow, prepared rollback.

➤ Overview

HILC-Legal places a control plane between the legal agent and any environment in which an action can have legal effect. The agent remains responsible for proposing and revising work, but it does not possess unrestricted execution authority. The control plane classifies the matter and action, applies hard rules, estimates risk, assembles a review record, routes the proposal to an appropriate human, and issues a narrow authorization only for the approved version. A secure broker then performs the action and reports the outcome.

The separation creates two independent boundaries. First, generation is separated from authorization: a model can propose but cannot self-approve. Second, authorization is separated from execution: a reviewer approves a defined action, while a broker verifies that the action, artifact, target, timing, and permissions match the authorization. This two-phase pattern reduces accidental drift between what a person saw and what the system ultimately did.

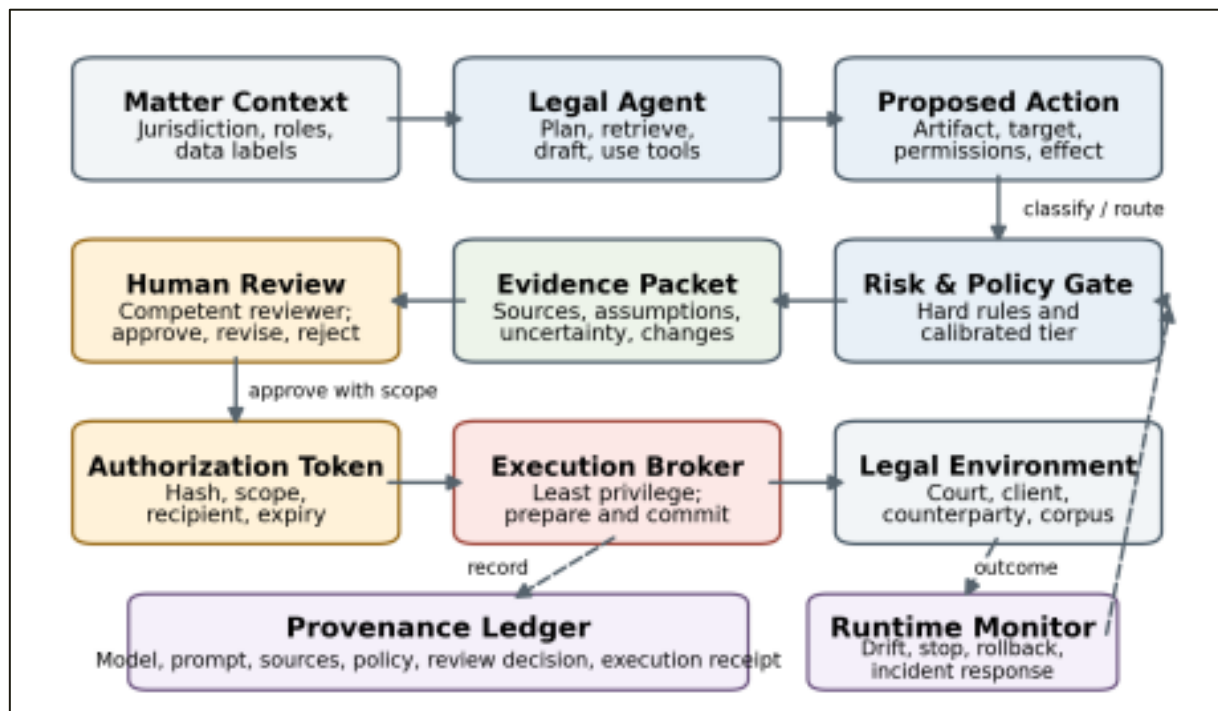


Fig 1 HILC-Legal Control Architecture Separating Agent Proposal, Risk and Evidence Controls, Competent Human Review, Version-Bound Authorization, Least-Privilege Execution, and Post-Action Monitoring.

➤ Functional Components

Table 2 Functional Components of the HILC-Legal Architecture.

Component	Inputs	Outputs / Responsibility
Matter and Role Classifier	Matter metadata, user identity, jurisdiction, data labels.	Establishes matter boundaries, reviewer roles, applicable policies, and authorized systems.
Risk and Policy Engine	Proposed action, matter context, requested permissions, uncertainty signals.	Applies deny and non-delegability rules; calculates the oversight tier; records reasons.
Legal Agent and Tool Planner	User objective, approved tools, scoped memory and corpus.	Produces a plan, draft, retrieval set, tool proposal, alternatives, and explicit assumptions.
Evidence Packet Service	Draft, sources, tool trace, model and policy metadata.	Builds claim-evidence links, source snapshots, contrary authority, changes, data labels, and unresolved issues.
Human Review Console	Proposed action, evidence packet, risk explanation, reviewer profile.	Supports approve, revise, reject, narrow, escalate, request evidence, or stop; captures rationale.
Authorization Token Service	Approved artifact, action scope, reviewer identity, conditions.	Issues a short-lived token bound to artifact hash, target, tool, permissions, and expiry.
Secure Execution Broker	Authorization token, exact artifact, requested external action.	Verifies all bindings, uses least privilege, stages the action where possible, and commits once.
Monitor, Ledger, and Incident Service	Events from all components and the legal environment.	Maintains provenance, detects drift or policy breach, supports rollback, notice, investigation, and revalidation.

➤ Evidence Packets and Provenance

The evidence packet is the unit of review. It contains the exact proposed artifact; intended recipient, system, and effect; matter and jurisdiction assumptions; material factual and legal claims; supporting source spans; subsequent history or adverse authority where relevant; tool calls and intermediate transformations; privilege and confidentiality labels; uncertainty and disagreement signals; alternatives considered; and a version diff from the last approved state. Sources must be inspectable in context rather than represented only by model-generated summaries.

A provenance ledger records the model and prompt versions, retrieval index and date, source identifiers, policy version, tool schemas, reviewer actions, authorization token, execution result, and subsequent corrections. The ledger should support reconstruction while respecting minimization and retention rules; indiscriminate logging can itself create confidentiality, security, and discovery risk.

➤ Competence-Aware Human Review

Routing is based on the decision required, not merely reviewer availability. The system checks whether the reviewer is assigned to the matter, has authority for the action, and possesses the necessary legal or operational competence. A privilege decision may require a lawyer or supervised legal team; a court filing may require the responsible attorney; a data-export exception may require security or privacy approval in addition to legal review.

The interface is designed against rubber-stamping. For the highest tier, the reviewer can be required to identify the dispositive issue or expected consequence before seeing the agent recommendation, inspect contrary evidence, and answer targeted checks. The console displays material changes and unresolved assumptions prominently, avoids default approval, and makes rejection and escalation as easy as acceptance. These features seek appropriate reliance rather than maximum trust.

➤ Version-Bound Authorization and Secure Execution

An approval produces an authorization token, not a general permission. The token binds the approved artifact hash, action type, target, recipient, matter, tool, allowed parameters, expiry, reviewer identity and role, and any conditions. Any post-approval modification invalidates the token and returns the action to review. The execution broker verifies the token and uses the smallest permission set required for that one action.

Where technically feasible, execution follows prepare and commit stages. A filing can be assembled and validated without submission; an email can be rendered with final recipients and attachments; a production set can be staged and checked for privilege and scope. The human authorizes the prepared state, after which the broker commits exactly that state. Idempotency keys prevent duplicate actions, and high-impact connectors remain unavailable to the agent itself.

➤ Runtime Monitoring and Recovery

Post-action controls detect delivery failure, unexpected recipient changes, policy violations, downstream rejection, late-discovered privilege, or drift in an ongoing workflow. The monitor can suspend subsequent actions, revoke tokens, quarantine outputs, trigger notice, and initiate rollback or corrective filing where available. Monitoring does not replace pre-commit review; it limits harm when prevention fails.

IV. RISK-ADAPTIVE OVERSIGHT POLICY

➤ Risk Vector and Score

For a proposed action a in matter m , HILC-Legal evaluates a context vector containing legal impact I , irreversibility V , data sensitivity D , epistemic uncertainty U , externality E , and affected-person vulnerability A . Safeguards G include validated sources, sandboxing, reversibility, narrow permissions, independent checks, and prior performance on the same task class. A calibrated score can be written as:

$$r(a,m) = \text{sigma}(w_I I + w_V V + w_D D + w_U U + w_E E + w_A A - \text{gamma}^T G)$$

The score is not a universal legal formula. Feature definitions, weights, and thresholds must be learned or set for a specific organization and validated on representative scenarios. The score explains routing among otherwise permitted actions; it cannot authorize an action blocked by a hard rule.

➤ Hard Constraints

Hard constraints implement prohibited and non-delegable conditions. The policy denies an action when the requester, agent, or tool lacks authority; when data would cross a forbidden boundary; when a required source or matter context is unavailable; or when the action violates a client, court, security, retention, or organizational rule. Human-in-command review is required for decisions such as final court filing, settlement or waiver, release of privileged information, material client advice, rights-affecting determinations, or other acts that policy assigns to an accountable professional. A low model uncertainty score cannot override these constraints.

➤ Oversight Tiers

- Tier 0 - bounded assistance: Read-only retrieval, formatting, or internal drafting within a scoped corpus; no external legal effect. Logging and automated checks remain active.
- Tier 1 - human-on-the-loop: Low-to-moderate, reversible workflows with monitoring, sampling, confidence and drift triggers, and an immediate stop mechanism.
- Tier 2 - human-in-the-loop pre-commit: A competent reviewer must inspect the evidence packet and approve the exact prepared action before any external effect.
- Tier 3 - human-in-command: The human makes the substantive legal decision, may use the agent as an advisory tool, and personally authorizes the final action within professional and organizational authority.

Risk-Adaptive Oversight Matrix					
Potential legal impact	Limited	T0 Read-only / bounded	T0 Read-only / bounded	T1 Human-on-loop	T2 Pre-commit review
	Material	T0 Read-only / bounded	T1 Human-on-loop	T2 Pre-commit review	T2 Pre-commit review
	Serious	T1 Human-on-loop	T2 Pre-commit review	T2 Pre-commit review	T3 Human-in-command
	Critical	T2 Pre-commit review	T2 Pre-commit review	T3 Human-in-command	T3 Human-in-command
		Fully reversible	Reversible with effort	Hard to reverse	Irreversible
Reversibility of the proposed action					
Modifiers: confidentiality, uncertainty, externality, and affected-person vulnerability may raise the tier. Hard constraints override the matrix for prohibited actions and non-delegable legal decisions.					

Fig 2 Graduated Oversight Matrix Based on Legal Impact and Reversibility. Sensitivity, Uncertainty, Externality, and Vulnerability Can Raise the tier, while Hard Constraints Override Score Based Routing.

➤ Escalation and Abstention

Escalation is triggered by conflicting or missing authority, unsupported material claims, ambiguous jurisdiction, sensitive data, novelty, material model disagreement, suspicious retrieved instructions, unexpected tool output, policy change, reviewer unavailability, or an action outside validated scope. The system preserves the state and evidence, explains why it stopped, identifies the missing decision or permission, and routes to a qualified reviewer. Abstention is an expected safe outcome, not an availability defect.

For ongoing workflows, the tier can rise as context changes. A discovery classifier may operate under Tier 1 on a validated batch, but a detected privilege pattern, new custodian, or anomalous error rate can pause the process and require Tier 2 review. Conversely, a policy may lower review frequency only after documented evidence of stable performance and only for actions that remain reversible and non-consequential.

➤ Control Policies for Representative Legal Workflows

• Legal Research and Memoranda

The agent may frame issues, search an approved corpus, summarize authority, and prepare an internal memorandum. The evidence packet must expose jurisdiction, court level, date, treatment, quotations or supporting spans, and contrary authority. An unsupported citation, false premise, or unresolved conflict triggers targeted retrieval or human review. Internal exploratory work may remain Tier 0 or Tier 1; a material conclusion used for advice, filing, or strategy rises to Tier 2 or Tier 3 according to consequence.

• Discovery and Privilege Review

Batch classification can use human-on-the-loop oversight when the corpus is isolated, the taxonomy is validated, and sampling is designed to detect both false releases and excessive withholding. Privilege indicators, family relationships, communications context, and model uncertainty should inform review. Any release decision, privilege waiver, or production outside the authorized set requires pre-commit review. The system must preserve source documents, classification history, reviewer changes, and defensible quality-control statistics.

• Drafting, Filing, and Deadline Actions

Draft generation is separable from filing. The agent may assemble text, forms, exhibits, and citation tables, but the filing package is staged with the final court, case number, parties, deadline, signature, service list, attachments, and payment details. The responsible professional reviews that exact package under Tier 3. Submission uses a single-purpose authorization token, and the receipt is returned to the ledger. Calendar changes that could waive a right or miss a deadline receive equivalent treatment.

• Client and Counterparty Communications

Internal drafting may be low risk, while transmission can create reliance, admissions, commitments, or waiver. The agent must not infer authority to send, negotiate, settle, or alter a legal position from its ability to draft. The review packet identifies recipients, attachments, represented parties, confidentiality status, requested response, and any language that could be construed as advice or commitment. High-impact communications require human-in-command review.

- *Rights-Affecting and Public-Law Decisions*

When an output may materially affect eligibility, employment, housing, benefits, enforcement, liberty, or access to a service, the human must make the substantive decision and be able to depart from the recommendation. Review should include the governing criteria, complete record, protected or sensitive attributes, plausible alternative explanations, consistency checks, reasons, and available contest or appeal routes. The interface must not reduce the human role to confirming a preselected outcome.

V. IMPLEMENTATION CONSIDERATIONS AND CHALLENGES

➤ *Policy as Code and Matter Isolation*

Policies should be versioned, testable, and traceable to an owner and rationale. They define allowed tools, data classes, review roles, hard triggers, thresholds, retention, and incident actions. Matter isolation applies to memory, retrieval indexes, caches, tool credentials, and logs; a useful response from another matter is still an unauthorized disclosure. Policy changes require review and regression testing against prior high-risk scenarios.

➤ *Tool Security and Prompt-Injection Resistance*

Retrieved documents and web content are untrusted data, not instructions. The agent should use allowlisted tools, typed schemas, parameter validation, output encoding, secret isolation, and narrow credentials. Data labels should propagate through transformations so that privileged or restricted content cannot be laundered into an unrestricted output. External actions are brokered and auditable; the model never receives standing credentials for high-impact systems.

➤ *Reviewer Workload and Service Levels*

A control that sends every action to a person can create alert fatigue and become a de facto approval factory. Risk routing should reserve intensive review for material decisions, while bounded actions use automated checks, statistical sampling, and periodic audit. Work queues expose deadlines and risk, limit concurrent high-risk reviews, and reassign when the reviewer lacks time or competence. Organizations should measure both delay and review quality rather than optimizing only throughput.

➤ *Auditability, Change Management, and Model Drift*

Agent behavior depends on the base model, system prompt, retrieval corpus, embeddings, ranking, tool descriptions, memory, and policy. An update to any of these can change the distribution of proposed actions. Deployment records should therefore identify each component version and trigger revalidation after material change. Periodic replay of a fixed high-risk test suite, live error monitoring, and sampled human audit help detect drift that aggregate task completion metrics may hide.

➤ *Accountability and Incident Response*

Operational ownership should be explicit: a legal owner defines non-delegable decisions; a system owner maintains the control plane; security and privacy owners define data

boundaries; and business owners accept residual risk. Incident plans cover token revocation, workflow suspension, preservation of evidence, client or court notification where required, correction, root cause analysis, and policy updates. Raji et al. (2020) emphasize lifecycle audit artifacts; HILC Legal extends that principle to action-level authorization and execution evidence.

VI. EVALUATION METHODOLOGY

➤ *Research Questions*

- RQ1 - Safety: Does HILC-Legal reduce unauthorized or legally significant erroneous actions compared with direct-agent and final-review baselines?
- RQ2 - Legal quality: Does the architecture improve authority selection, evidence support, confidentiality preservation, and action correctness?
- RQ3 - Human performance: Do evidence packets and competence routing improve reviewer correction without causing unacceptable workload, delay, or automation bias?
- RQ4 - Robustness: Do controls remain effective under false premises, prompt injection, source conflict, model update, stale approvals, tool failure, and time pressure?

➤ *Tasks and Scenario Corpus*

Evaluation should use de-identified, synthetic, or appropriately authorized matters across legal research, discovery, privilege, contract revision, filing preparation, client communication, and rights-affecting decisions. Scenarios are stratified by jurisdiction, impact, reversibility, sensitivity, uncertainty, and affected-person vulnerability. Each scenario contains a gold or expert-adjudicated set of acceptable actions, required evidence, prohibited actions, and escalation conditions. LegalBench can inform reasoning task design, while legal-RAG evaluations such as Magesh et al. (2025) motivate false-premise, jurisdictional, temporal, and grounding probes (Guha et al., 2023).

Adversarial scenarios include conflicting precedent, a superseded authority, a forged citation, a malicious instruction inside a retrieved document, a mislabeled privileged attachment, an unauthorized recipient, expired permissions, a last-minute artifact modification, duplicate submission, an unavailable reviewer, and a model or policy version change between approval and execution.

➤ *Baselines and Ablations*

Baselines include: (i) a direct legal agent with unrestricted tools; (ii) an agent with a final human approval step but no structured evidence or version binding; and (iii) a fixed gate that requires approval for every external action. The full HILC-Legal system is compared with ablations that remove the risk engine, hard constraints, evidence packet, competence routing, version-bound token, secure broker, or runtime monitor. This design tests whether each control adds measurable protection rather than merely architectural complexity.

➤ Metrics

- Unsafe-action escape rate: Fraction of prohibited or materially erroneous actions that reach external execution; report severity-weighted and critical-event counts.
- Decision-relevant error rate: Fraction of completed workflows containing an error capable of changing legal advice, rights, filing status, privilege, or obligation.
- Authority and evidence fidelity: Validity of cited authority, supported-claim rate, source entailment accuracy, adverse-authority coverage, and jurisdiction/date correctness.
- Confidentiality and authorization integrity: Protected-data egress incidents, cross-matter access, excessive permissions, and unauthorized recipient or tool use.
- Version fidelity: Rate at which the executed artifact, target, and parameters exactly match the reviewed and authorized state.
- Human correction yield: Share of reviews that identify and correct a material issue, together with false-alarm and unnecessary-escalation rates.
- Human factors: Inappropriate agreement with wrong recommendations, review time, evidence-inspection behavior, workload, confidence calibration, and inter-reviewer agreement.
- Operational resilience: Intervention latency, stop success, duplicate-action prevention, rollback success, trace completeness, and time to contain an incident.

➤ Human Evaluation and Statistical Analysis

Qualified reviewers complete a balanced set of scenarios under baseline and HILC-Legal conditions. The study records decision accuracy, corrections, evidence use, time, confidence, and reasons. Case order and interface condition should be counterbalanced, with sufficient washout to reduce learning effects. Mixed-effects analysis can account for reviewer and matter variation; bootstrap confidence intervals and severity-stratified results should accompany aggregate means. Critical failures are reported individually rather than hidden by an average.

➤ Acceptance Gates and Reporting

Thresholds must be prespecified for the deployment context. At minimum, the test environment should demonstrate no unauthorized execution of designated critical actions, complete approval to-version matching, containment of protected test data, traceability of every high-risk decision, and successful stop behavior under the defined fault scenarios. These are proposed acceptance gates, not claimed results. Reporting should disclose the scenario distribution, reviewer population, model and policy versions, confidence intervals, residual risks, and representative failures.

➤ Ethical, Regulatory, and Professional Considerations

• Responsibility Without Human Scapegoating

Human oversight should preserve professional judgment, but it should not be used to transfer all responsibility to a reviewer who lacked information or control. Responsibility follows design and authority:

developers and deployers remain responsible for validation, permissions, interfaces, monitoring, staffing, and truthful capability claims; reviewers remain responsible for decisions within their competence and assigned role. Logs should reveal whether the human had a real opportunity to understand and intervene.

• Fairness, Contestability, and Affected Persons

Risk assessment includes the vulnerability and interests of affected persons, not only organizational loss. For rights-affecting uses, affected parties may need understandable reasons, a route to correct data, and meaningful human reconsideration. Reviewers should examine disparate impacts and proxy variables where relevant. A system that is accurate on average can still be unacceptable if its errors are concentrated on a protected or vulnerable group.

• Confidentiality, Minimization, and Transparency

Evidence and auditability must be balanced with confidentiality and data minimization. The system should retain only what is necessary for review, accountability, security, and applicable obligations; restrict access; and separate operational telemetry from substantive client content where possible. Users and reviewers should know when an agent is active, which sources and tools it uses, and the limits of its authority. External disclosure requirements vary by jurisdiction and context and must be configured accordingly.

• Regulatory Localization

Nannini et al. (2026) emphasize that agent deployments can trigger overlapping EU legal regimes because agents combine data flows, external actions, and changing runtime behavior. Nguyen et al. (2026) likewise demonstrate the value of specialized agents coupled with human verification for legal formalization. HILC-Legal does not replace jurisdiction-specific legal analysis. Its policies must map local professional rules, court requirements, privacy and employment law, sector obligations, contractual restrictions, and organizational authority.

VII. LIMITATIONS AND FUTURE WORK

HILC-Legal is a conceptual architecture and evaluation plan, not evidence that a particular deployment is safe. Risk features and thresholds are context-dependent; policy authors can misclassify an action; evidence extraction can omit a dispositive source; and reviewers can still err. Some consequences cannot be reliably reversed. A secure broker reduces tool risk but does not cure an incorrect professional decision.

The framework also introduces cost and latency. Detailed evidence packets may burden reviewers, while aggressive routing may create queues and delay. Logging can expose sensitive information, and cryptographic version binding does not prove that the reviewer understood the artifact. Cross-system workflows, delegated organizations, and rapidly changing agent plans complicate attribution. These limitations require empirical study rather than assuming that more controls are always better.

Future work includes practitioner studies across jurisdictions and specialties; formal verification of hard policy rules and authorization-token semantics; adversarial benchmarks for prompt injection, privilege, and action drift; learning risk thresholds from expert feedback without weakening hard constraints; adaptive sampling for high-volume review; privacy-preserving audit methods; and standardized reporting of human-agent incidents. The most important empirical question is whether the complete sociotechnical system produces fewer severe legal failures than either the agent or the human process alone.

VIII. CONCLUSION

High-risk legal agents require more than accurate language generation and more than a final approval button. Meaningful human oversight must be implemented as an enforceable control property. HILC-Legal separates proposal, authorization, and execution; combines hard constraints with calibrated risk routing; presents decision-relevant evidence to a competent reviewer; binds approval to the exact artifact; restricts tool authority; and monitors outcomes. The framework does not eliminate error or transfer professional responsibility to software. It makes the conditions for safe delegation explicit, testable, and auditable, enabling organizations to use agentic assistance while preserving judgment, confidentiality, accountability, and the ability to stop.

REFERENCES

- [1.] American Bar Association Standing Committee on Ethics and Professional Responsibility. (2024). Formal Opinion 512: Generative Artificial Intelligence Tools.
- [2.] Amershi, S., Weld, D., Vorvoreanu, M., Fournay, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-13.
- [3.] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
- [4.] Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1-16.
- [5.] Chiodo, M., Muller, D., Siewert, P., Wetherall, J.-L., Yasmine, Z., & Burden, J. (2025). Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility. arXiv:2505.10426. European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 14.
- [6.] Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 36.
- [7.] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Journal of Empirical Legal Studies, 22, 216-242. Nannini, L., Smith, A. L., Maggini, M. J., Panai, E., Feliciano, S., Tiulcanov, A., Maran, E., Gealy, J., & Bisconti, P. (2026). AI Agents Under EU Law. arXiv:2604.04604.
- [8.] Nguyen, H. T., Fungwacharakorn, W., Wehnert, S., Zin, M. M., Kong, Y., Xue, J., Araszkievicz, M., Goebel, R., & Satoh, K. (2026). GDPR Auto-Formalization with AI Agents and Human Verification. arXiv:2604.14607. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30(3), 286-297. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.
- [9.] Sinha, A., Ranganathan, S., Dharmaratnakar, A., & Das, D. (2026). Human-on-the-Loop Orchestration for AI-Assisted Legal Discovery. arXiv:2606.19812.
- [10.] Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1.