The Fiduciary's Footprint:

How Legal AI Agents Can Prove They Were Loyal

Ken Priore
Deputy General Counsel, Docusign
ken@kenpriore.com · www.kenpriore.com

Discussion Draft

NYU Law Workshop: "Fiduciary Duties and AI: Legal Frameworks, Technical Implementation, and Governance"

Information Law Institute, NYU School of Law + GliaNet Alliance
June 4–5, 2026 · Vanderbilt Hall, New York

May 2026

Abstract

AI agents are already performing tasks in legal workflows that implicate fiduciary duties: reviewing contracts, analyzing risk, routing compliance decisions, surfacing research. The lawyers and institutions deploying these agents remain fiduciaries to their clients. But when a client or regulator asks "prove your AI acted loyally and with care," there is currently no answer. Not because the duty has disappeared, but because no infrastructure exists to document its fulfillment in a form that survives adversarial scrutiny.

This paper proposes the Agent Attestation as the technical operationalization of fiduciary duty for autonomous legal AI agents. The Agent Attestation is a tamper-evident, cryptographically sealed record that captures what an agent did, under what authority it operated, what diligence it exercised, and whether a human reviewed the outcome, structured specifically for legal evidentiary admissibility. The four layers of the Agent Attestation architecture map onto the four core fiduciary duties: the duty of care maps to consensus verification, the duty of loyalty maps to policy binding, the duty of confidentiality maps to the privacy-preserving verification model, and the duty of disclosure maps to the human review record.

Crucially, this infrastructure exists today, is deployable at modest cost, and requires no novel technology; only the deliberate application of widely available cryptographic tools to a problem the legal AI community has not yet systematically addressed.

Contents

Introduction: The Lawyer's Dilemma

A managing partner at a mid-size firm is defending the firm's use of an AI agent to conduct contract review for a private equity client. The client has raised a concern: they believe the AI recommended approving a clause that a reasonably careful attorney would have flagged. The managing partner knows, intuitively, that the review process was aligned to the firm's policies, the agent was well-configured, a senior associate reviewed the output, and the recommendation was made in good faith. But when pressed to *prove* it, she finds she cannot.

The AI platform provides logs. They show that the agent ran, that inputs were processed, that outputs were generated. They do not show what policy the agent was operating under, whether it exercised any form of independent verification, what confidence it attached to the high-risk clauses, or whether the associate who reviewed the output saw the agent's actual reasoning or just its recommendation. The logs were designed for debugging, not for fiduciary accountability. They answer the engineering question, "did the system run correctly?", but not the legal question: did the agent act loyally and with care toward this client?

This is the fiduciary gap in legal AI. Legal AI tools have become genuinely useful for contract review, due diligence, research, and compliance triage. The gap is in accountability infrastructure — the systems and records that make fiduciary duty not just a professional obligation but a demonstrable fact.

This paper argues that closing this gap is attainable today with a new kind of record: an Agent Attestation.

The Fiduciary Gap

Fiduciary Duty Persists Through Delegation

The threshold point is simple but frequently obscured: delegating a task to an AI agent does not delegate the fiduciary duty. A lawyer who uses AI to review a contract still owes their client the duty of care that a reasonable attorney would exercise. An investment advisor whose AI system generates portfolio recommendations still owes the client suitability obligations. A physician whose clinical AI triages patient records still owes the patient a duty grounded in their medical relationship.

American Bar Association (2024) makes this explicit for legal practice: competence and supervision duties under Model Rule 5.3 apply to AI tools in the same way they apply to non-lawyer staff. A supervising attorney is responsible for the work of the AI agent they deploy. The relevant question is not whether an AI was used but whether the supervising professional exercised adequate oversight.

FIGURE 1

Figure 1. *The Agent Attestation captures four layers of evidence — authority, diligence, human oversight, and execution — sealed in a cryptographic chain. Each layer maps to a core fiduciary duty. The full record stays with the certifying platform; only a fingerprint is registered publicly.*

This creates a specific evidentiary requirement. When oversight is later questioned — by a client, a disciplinary authority, or a court — the professional must be able to demonstrate what oversight actually occurred. Today, they frequently cannot.

What Existing Systems Cannot Prove

Current AI platforms offer three categories of documentation, none of which satisfies the fiduciary accountability requirement.

Execution logs. Agent orchestration frameworks maintain records of what ran: which tools were called, what inputs were provided, what outputs were generated. These records are non-standardized across platforms, optimized for debugging rather than legal production, and not tamper-evident. A log entry can be altered without detection. More fundamentally, logs answer *what happened* but not *under whose authority* and *with what diligence* — the questions that fiduciary accountability requires.

Observability dashboards. Specialized monitoring tools provide real-time visibility into agent performance: latency, error rates, reasoning traces, flagged outputs. These are valuable operational tools. They do not produce a sealed artifact. They cannot answer

"was this decision made with reasonable care?" for a specific client transaction six months after the fact. They are watch-only systems; they observe what agents do but cannot certify it.

Manual review records. Some firms maintain separate records of attorney review: notes, emails, approval workflows. These are not integrated with the AI system's actual output. The gap between "what the agent produced" and "what the reviewer saw and approved" is where fiduciary accountability lives, and it is what these systems leave undocumented.

None of these approaches produces what fiduciary accountability requires: a single, sealed record linking the agent's specific reasoning to the policy it was authorized to apply, to the diligence it exercised, to the human oversight that occurred, and to the executed outcome — with each of these elements tamper-evident and independently verifiable.

TABLE 1

THE ACCOUNTABILITY GAP

What fiduciary law requires an AI agent to prove — and whether existing infrastructure can do it

ACCOUNTABILITY QUESTION	CURRENT INFRASTRUCTURE	AGENT ATTESTATION
What policy governed this decision?	GAP *Logs may exist but are mutable and provider-controlled. No tamper-evident record binds the decision to the principal's authorized instruction set at the moment of execution.*	LAYER 1 Policy version is sealed into the Agent Decision Chain at task initiation. The hash chain prevents retroactive alteration.
Did the agent exercise reasonable diligence?	GAP *Single-pass inference leaves no record of whether the agent checked its own reasoning. There is nothing to audit for diligence.*	LAYER 2 Consensus Verification documents k independent passes, agreement rate, and dissenting analysis. The diligence record is part of the sealed chain.
Was a qualified human reviewer in the loop?	GAP *Human involvement may have occurred but is undocumentable. No verifiable record of who reviewed what, when, and on what basis.*	LAYER 3 Human Review Record seals reviewer identity, timestamp, what was presented, and decision rationale. Absence of review is also documented.
Is privileged content protected from disclosure?	GAP *Clients must trust the provider's assurances. No technical mechanism prevents privileged content from crossing into a public log or third-party audit trail.*	LAYER 4 Only the root hash enrolls in the public log. Privileged content never leaves the certifying platform. Verified without being exposed.
Can this be audited retroactively?	GAP *Provider-controlled logs are mutable and require the provider's cooperation. No independent audit path exists. No chain of custody from decision to outcome.*	ALL LAYERS Each layer is hash-chained to the next. A verifier can confirm the full attestation without the provider's cooperation or consent.
Is the record admissible as evidence?	GAP *Provider-controlled logs face authenticity challenges in litigation. No equivalent to a notarized record or court-recognized timestamp under ESIGN or eIDAS.*	SEAL Trusted timestamp, digital signature, and verifiable log enrollment produce a legally recognized record analogous to a notarized instrument.

Table 1. *Six accountability questions that fiduciary law requires AI agents to answer, against what current infrastructure can and cannot document. The gap is not a failure of intent — it is the structural absence of the evidentiary infrastructure that fiduciary accountability requires.*

So What?

The deployment of legal AI is accelerating at the same moment that regulatory and professional accountability frameworks are sputtering. The EU AI Act (European Parliament and Council of the European Union, 2024) imposes logging and traceability requirements for high-risk AI systems that will apply to many legal AI deployments. The SEC has

begun scrutinizing AI use in investment advice. State bar disciplinary committees are developing guidance on AI supervision. Courts are beginning to demand disclosure of AI use in litigation support.

In each of these contexts, the question will eventually be asked: "Prove your AI acted appropriately." Firms that have invested only in capability — better models, faster processing, wider coverage — without investing in accountability infrastructure will find themselves unable to answer.

The Agent Attestation

What It Is

The Agent Attestation is a tamper-evident, cryptographically sealed record that certifies an autonomous AI agent workflow. For each consequential agent workflow execution, the Agent Attestation captures four layers of evidence:

Layer 1: The Agent Decision Chain. What every agent in the workflow did, why, under what policy it operated, what evidence it considered, and how confident it was. Each agent's record is linked to the next in a cryptographic chain: if any record is modified after the fact, the modification is detectable. The chain cannot be selectively edited.

Layer 2: Consensus Verification. When the workflow employs multiple independent reasoning passes over the same question — a form of built-in quality control — the Agent Attestation records the agreement rate across passes and captures any dissenting analysis. An agent that checks its own reasoning five times and reaches the same conclusion each time has demonstrably exercised more diligence than one that does not. This record is not optional: the dissenting passes are captured even when the majority prevails, preventing selective disclosure.

Layer 3: The Human Review Record. When the workflow policy requires, or when confidence falls below a threshold, the decision escalates to human review. The Agent Attestation captures who reviewed, when, what they decided, and their stated rationale — cryptographically chained to the agent's decision so the two cannot be separated. Human oversight is not just noted; it is sealed into the record as an integral element.

Layer 4: The Execution Record. When the workflow produces an actionable output — a contract redline, a compliance determination, a risk assessment — the executed artifact and the agent governance record are sealed together. The Agent Attestation links the decision process to its outcome in a single artifact.

The completed Agent Attestation is sealed using widely available cryptographic tools: a tamper-evident hash chain links all four layers, a trusted timestamp provides legally recognized proof of creation time, and a digital signature binds the certifying platform's identity to the record. The full attestation is stored by the certifying platform; only a

fingerprint is registered with a public verifiable log, preserving confidentiality of privileged content while enabling independent verification.

How It Already Exists

This is not a proposal for new technology. Every component of the Agent Attestation is built on infrastructure that has been in production use for decades.

The tamper-evident hash chain is the same mechanism used in blockchain consensus and code signing. The trusted timestamp is the same mechanism used by qualified electronic signatures under eIDAS (European Parliament and Council of the European Union, 2014) in Europe and relied upon in commercial practice globally. The digital signature infrastructure is the same public key infrastructure that secures every HTTPS connection on the internet. The verifiable log architecture is the same model that has been used since 2013 to make TLS certificate issuance publicly auditable.

What is new is the deliberate application of these tools to AI agent governance records, structured specifically to produce an artifact suitable for legal evidentiary use. The deployment path mirrors how electronic signature infrastructure was adopted in the early 2000s: a layer of trust infrastructure built on open standards, deployable by any platform willing to implement it, creating a market for verified compliance rather than requiring a single centralized authority.

Fiduciary Duties, Layer by Layer

The Agent Attestation's four-layer architecture maps onto the four core fiduciary duties with a precision that is not coincidental — it reflects that the accountability structure fiduciary law demands and the structure required for cryptographic certification are solving the same underlying problem.

Duty of loyalty → Layer 1 (Authority Binding). The duty of loyalty requires that a fiduciary act in the client's interest, not the agent's own or the platform's. For an AI agent, this means operating under the principal's authorized instructions rather than the platform's default configuration. Every agent record in the decision chain includes a reference to the specific policy version under which the agent operated — a versioned instruction set representing the client's authorized mandate. The sealed record proves what instructions the agent was following at the time of the decision. If the platform's defaults diverge from the client's mandate, the policy reference exposes that divergence.

Duty of care → Layer 2 (Consensus Verification). The duty of care requires a fiduciary to exercise the diligence that a reasonable professional would exercise. For an AI agent, this requires some form of independent verification of reasoning, not simply accepting the first output. The consensus mechanism operationalizes this: multiple independent reasoning passes, an agreement rate, and captured dissent. An attestation showing five of

TABLE 2

FIDUCIARY DUTY	AGENT ATTESTATION LAYER	MECHANISM	WHAT IT PROVES
LOYALTY *Authority binding*	LAYER 1 **Agent Decision Chain**	Policy version sealed at task initiation. Every decision traces back to the principal's authorized instruction set — not the platform's defaults. Evidence considered, confidence level, and reasoning trace are recorded.	*The agent acted under the principal's authority, not a third party's interest.*
CARE *Diligence record*	LAYER 2 **Consensus Verification**	k independent reasoning passes are documented. Agreement rate and dissenting analysis are captured. Threshold is configurable to domain risk level (k ≥ 2 recommended for high-stakes decisions).	*The agent exercised reasonable diligence — it checked its own reasoning before acting.*
DISCLOSURE *Oversight proof*	LAYER 3 **Human Review Record**	Reviewer identity, timestamp, what was presented to the reviewer, and decision rationale are sealed into the chain. Where escalation did not occur, the record documents its absence.	*A qualified human reviewed the output at a defined gate — or the record shows the omission.*
CONFIDENTIALITY *Hash-only log*	LAYER 4 **Execution Record**	Output artifact sealed and linked by hash. Only the root hash enrolls in the public verifiable log. Privileged content never leaves the certifying platform or crosses a disclosure threshold.	*Privileged content was handled correctly — verified without being exposed.*

Table 2. *Each fiduciary duty maps to a distinct layer of the Agent Attestation. The mapping is not decorative: each layer produces a tamper-evident record of exactly the evidence that duty requires the fiduciary to preserve.*

five passes reaching the same conclusion on a high-risk clause provides stronger evidence of care than a single-pass output. An attestation showing two of five passes in agreement is itself material information — it documents that the uncertainty was recognized and (per Layer 3) escalated.

Duty of disclosure → Layer 3 (Human Review Record). The duty of disclosure requires that a fiduciary inform the principal of material facts and reasoning. When an AI agent makes a determination that a principal should know about, the fiduciary obligation to disclose runs to the human who can explain the determination to the client, not to the AI. The Human Review Record captures the full escalation chain: who received the agent's output, when they reviewed it, what they decided, and what they communicated. Chained cryptographically to the agent decision record, it is impossible to later claim the human reviewed information they did not see, or approved reasoning they did not have access to.

Duty of confidentiality → Layer 4 (Privacy-Preserving Verification Model). A fiduciary's confidentiality obligations extend to the reasoning the agent engaged in on the client's behalf. A verification model that exposes privileged reasoning to a public log would be incompatible with attorney-client confidentiality and its equivalents in other fiduciary relationships. The Agent Attestation's hash-only enrollment model addresses this directly: only the record's cryptographic fingerprint is registered with the public verifiable log. The reasoning traces, client data, and decision rationale stay with the certifying platform. A verifier — regulator, counterparty, court — can confirm the attestation's authenticity and integrity without seeing its contents, unless the holder chooses to produce it.

The Human-in-the-Loop Question

The workshop poses directly: "What does law require of fiduciaries when delegating decision-making to AI systems? Must a human remain in the loop?"

The honest answer is that the law does not yet give a single answer to this question. The requirements vary by domain, by jurisdiction, and by the nature of the fiduciary relationship. ABA Model Rule 5.1 requires supervising attorneys to take reasonable steps to ensure non-lawyer staff comply with professional obligations, but what "reasonable steps" means for AI agents is currently being worked out. SEC guidance on AI in investment advice is still evolving. Medical boards are at early stages of addressing AI in clinical decision support.

The Agent Attestation does not resolve this doctrinal question. But it makes whatever answer the law provides *enforceable and auditable*.

Today, the "human in the loop" question is addressed primarily through internal policy: firms assert that attorneys review AI outputs, and there is generally no way to verify or falsify this claim. The Agent Attestation turns an assertion into a verifiable fact, and does so for the regulator, counterparty, and court as much as for the firm itself.

If regulators determine that human review is required before a legal AI agent's output can be relied upon in a specific context, the Agent Attestation's Human Review Record layer makes that requirement verifiable. A firm can prove, for any given decision, that a human reviewed it — when, who, what they saw, and what they decided. If regulators determine that human review at defined checkpoints is sufficient, the Agent Attestation documents those checkpoints in a tamper-evident record that cannot be reconstructed after the fact.

The firm that has adopted Agent Attestation infrastructure is not simply compliant — it can *prove* compliance on demand. That proof is not contingent on the regulatory answer to the normative question. It is available regardless of what the final rule turns out to be, because the record was sealed at the time of the decision.

What This Means for Legal AI in Practice

The Supervising Attorney Gets an Audit Trail

Under current practice, the supervising attorney's review of AI output is largely undocumented. The attorney sees the output, exercises judgment, and either adopts, modifies, or rejects it. If the judgment is later questioned, the attorney's account of what they did is the primary evidence.

The Agent Attestation changes this. The supervising attorney's review is recorded — including what they actually saw (the agent's reasoning, not just its output), when they

reviewed it, and what they decided. The chain from AI reasoning to human decision to client communication is sealed. The attorney can demonstrate not just that they reviewed the AI's output but that they engaged with the underlying analysis.

A file memo has always served this purpose, documenting the attorney's exercise of professional judgment. The Agent Attestation is the AI-era equivalent, automated and cryptographically verifiable.

The Client Gets a Record They Can Actually Examine

When a client questions an AI-assisted legal determination, they currently have no direct access to the agent's reasoning. They receive the lawyer's explanation of what the AI found and why the lawyer agreed. The Agent Attestation creates the possibility — at the lawyer's discretion and consistent with privilege — of producing a structured record showing exactly what the agent considered, what confidence it assigned, and what the reviewing attorney decided.

This is analogous to what audit firms have begun doing with AI-assisted audit procedures: documenting the AI's analysis as a distinct element of the audit record, separate from but connected to the auditor's professional judgment. The Agent Attestation provides the infrastructure for legal practice to make the same move.

The Trust Market for Legal AI

There is a practical business dimension here that deserves acknowledgment in a workshop with practitioners and entrepreneurs. Fiduciary accountability infrastructure creates a trust market.

Today, legal AI tools compete primarily on capability: which tool reviews contracts faster, identifies risk more accurately, surfaces relevant precedent most reliably. This is a market that rewards the best model.

A market with Agent Attestation infrastructure rewards something different: the most verifiably trustworthy workflow. A legal AI platform that can produce an Agent Attestation for every consequential decision — and have it independently verified by a client's auditor or a regulator — is offering something that a platform without that infrastructure cannot match, regardless of raw model performance.

Electronic signature platforms followed the same logic. What made them defensible was not the signature itself but the trust record — a tamper-evident artifact proving the signing event occurred, independently verifiable, built to survive adversarial scrutiny.

Limitations

The Agent Attestation proves process, not outcome. A perfectly sealed Agent Attestation can certify a bad legal decision. If the AI agent's reasoning was flawed, or the supervising attorney's judgment was poor, the Agent Attestation faithfully documents that flawed process. It does not independently verify whether the legal analysis was correct. Fiduciary accountability infrastructure is a floor, not a ceiling — it proves that a documented process occurred, not that the process was substantively sound. The evaluation of AI reasoning quality is a separate and harder problem that the legal AI research community is actively working on.

Adoption requires bilateral trust. An Agent Attestation is only valuable if the party receiving it trusts the entity that produced it. A firm's self-issued attestation provides weaker assurance than one issued by an accredited certification platform, just as a self-signed SSL certificate provides weaker assurance than one issued by a trusted certificate authority. Building the institutional credibility of attestation issuers will take time and probably requires accreditation frameworks analogous to those that govern qualified electronic signature providers under eIDAS. The electronic signature precedent is instructive: the ESIGN Act (ESIGN) was enacted in 2000, but widespread legal professional acceptance took roughly fifteen years.

The autonomy threshold problem. There is a theoretical boundary to what any certification framework can achieve for complex multi-agent AI systems. Tibebu (2026) proves that beyond a certain autonomy threshold, no accountability framework can simultaneously satisfy attributability, foreseeability, and completeness for human-agent collectives. The Agent Attestation operates within the tractable regime below this threshold — it is designed for workflows where agents operate under defined policies with configurable human oversight gates. For more open-ended agentic systems with complex multi-agent delegation, the accountability problem becomes formally intractable, and the Agent Attestation's value as a fiduciary accountability tool diminishes accordingly. Practitioners should be aware of this boundary when scoping AI agent deployment for fiduciary contexts.

From Assertion to Evidence

The fiduciary's footprint is the record that proves the duty was fulfilled. For human professionals, this record has always existed in some form: file memos, engagement letters, review notes, correspondence. For AI agents acting in fiduciary contexts, it does not yet exist in a form suitable for the accountability demands that fiduciary law, regulatory scrutiny, and adversarial proceedings will place on it.

The Agent Attestation is a proposal for what that record should look like, built on

technology that already exists, deployable at costs legal organizations can absorb, and structured to map onto the core duties that fiduciary law has already defined. It does not require waiting for new regulation or new technology. It requires only the deliberate application of existing tools to a problem that the legal AI community has deferred too long.

The question is not whether AI agents will occupy fiduciary roles in legal practice — they already do. The question is whether the profession will have the infrastructure to prove they discharged those roles faithfully.

References

American Bar Association. Formal opinion 512: Generative artificial intelligence tools, 2024. ABA Ethics Opinion 512.

ESIGN. Electronic signatures in global and national commerce act (ESIGN act). Pub. L. 106-229, 114 Stat. 464, codified at 15 U.S.C. § 7001 *et seq.*, 2000.

European Parliament and Council of the European Union. Regulation (EU) no 910/2014 on electronic identification and trust services for electronic transactions in the internal market. Regulation L 257, Official Journal of the European Union, 2014.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act). Regulation L 2024/1689, Official Journal of the European Union, 2024.

Hana Tibebu. The accountability horizon: An impossibility theorem for governing human-agent collectives, 2026. Submitted April 9, 2026.