Certificate of Action: A Trust Primitive for Verifiable Autonomous Legal AI Agent Workflows

Kenneth Priore [1]

Abstract

Autonomous AI agents are increasingly deployed in consequential legal workflows, yet no standardized mechanism exists to certify what an agent did, why, under what authority, and whether a human reviewed the outcome. We propose the *Certificate of Action* (CoA), a tamper-evident, cryptographically sealed record that certifies an autonomous AI agent workflow as a single verifiable artifact. The CoA captures four evidence layers built on existing open standards for tamper evidence, temporal integrity, identity binding, and independent verification: agent decision chains, consensus verification, human review records, and execution lifecycle. We provide a systematic regulatory compliance mapping demonstrating that the CoA satisfies traceability, documentation, and oversight requirements under the EU AI Act and NIST AI RMF, and fills a tamper-evidence gap in the emerging ISO/IEC 24970 logging standard. The Certificate of Action operates at the harness level of contemporary agent infrastructure, providing a legal-evidentiary primitive complementary to the runtime governance mechanisms that recent surveys identify as required but underoperationalized: review gates, rollback, and provenance tracking (Zhou et al., 2026). We distinguish the configurations under which a certificate is issued—self-hosted, independent, or regulated—and state explicitly what the CoA does and does not certify: the integrity of what was sealed, not that everything was sealed. We demonstrate feasibility through a proof-of-concept implementation and argue that trust infrastructure for legal AI is as critical as model capability.

[1]Docusign, Inc., San Francisco, CA, USA. Correspondence to: Kenneth Priore <ken@kenpriore.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

AI agents are making consequential decisions in legal and regulated domains—reviewing contracts, triaging compliance issues, underwriting loans, routing insurance claims—at a pace that outstrips existing oversight infrastructure. Unlike traditional AI systems that respond to discrete queries for human review, agentic AI systems chain multiple decisions together autonomously, determining their own execution paths with limited human oversight at each step (Pandey, 2026). This autonomy creates a fundamental accountability gap: when a regulator asks "show me the decision record for this agent-driven determination," the answer is typically scattered across log files, model outputs, and internal databases that were never designed to serve as evidence (Arbel et al., 2026).

This gap is structurally analogous to a problem that was solved for human agreements. Before the widespread adoption of electronic signature platforms, the question "was this document signed?" required producing a physical paper trail. The industry resolved this through standardized, tamper-evident certificates of completion—single documents, cryptographically sealed, that captured the full signing lifecycle and could be independently verified by courts, regulators, and counterparties. That certificate became the trust primitive for human agreements.

AI agent decisions in legal contexts require the same kind of trust primitive.

Recent work on agent infrastructure frames the shift from model-centric to system-centric capability through the lens of externalization: the relocation of cognitive burdens out of the model and into runtime modules—memory, skills, and protocols—coordinated by what Zhou et al. (2026) term the *harness*, a unification layer organized around three operational surfaces (Permission, Control, and Observability). Within that framework, observability is defined as the surface that supports "debugging, compliance auditing, and post-incident analysis" (Zhou et al., 2026, §6.2.4). We argue that this characterization is necessary but insufficient for autonomous legal AI workflows. Compliance auditing is an internal-control function; post-incident analysis is an engineering function. Legal accountability requires a record that

survives adversarial third-party scrutiny—by counterparties, arbitrators, regulators, and courts. The Certificate of Action proposes an evidentiary primitive layered over the harness's observability surface: a cryptographically sealed, tamper-evident record of agent decisions, the policies they operated under, and the human review they received, designed for admissibility rather than introspection.

This work responds directly to the disciplinary disconnect identified by Mahari et al. (2023): legal NLP research has focused heavily on benchmark performance while underinvesting in the infrastructure required for trustworthy deployment. As a practitioner-authored contribution, we present governance infrastructure that the legal AI community needs before the evaluation research this workshop advances can be responsibly deployed in production legal workflows.

We propose the **Certificate of Action (CoA)**: a formal, tamper-evident document that certifies an autonomous AI agent workflow—what every agent did, why, under what authority, with what controls, and whether a human reviewed the outcome. The CoA is to AI agent decisions what the certificate of completion is to human signatures: a single, verifiable record that can be produced for regulators, shared with counterparties, and defended in court.

Our contributions are fourfold:

1. We introduce the Certificate of Action as a new trust primitive for legal AI systems, providing a formal architectural specification (Section 3).

2. We ground the CoA in existing open standards at varying maturity levels—from battle-tested cryptographic primitives to emerging industry conventions for AI agent trace capture—demonstrating that no novel protocol infrastructure is required (Section 3.2).

3. We demonstrate feasibility through a working proof-of-concept implementation in a consumer lending workflow with five autonomous agents, hash chain construction, and automated escalation to human review (Section 3.5).

4. We provide a systematic regulatory compliance mapping showing how the CoA satisfies specific traceability, documentation, transparency, and human oversight requirements under the EU AI Act and the NIST AI Risk Management Framework, and identify how the CoA fills a tamper-evidence gap in the emerging ISO/IEC 24970 logging standard (Section 4).

2. Background and Related Work

2.1. The Accountability Gap in Agentic AI

Traditional AI liability frameworks assume a clear chain of human decision-making: a developer builds a model, a deployer integrates it, a user reviews its output, and responsibility flows along that chain (Surden, 2019). Agentic AI breaks this model. When an agent makes five sequential decisions and the third creates a problem, existing frameworks provide no clear answer to who owns that outcome (Arbel et al., 2026).

Kolt (2026) argues that AI agents are becoming not merely tools but *participants* in legal systems—subjects of law, consumers of law, and even producers and enforcers of law. This transformation demands governance infrastructure that treats agent decisions as first-class objects requiring certification, not merely outputs requiring logging. O'Keefe & Frazier (2026) raise a complementary concern: when AI systems themselves perform compliance tasks, the question of "who audits the auditor" becomes structural rather than procedural.

The legal AI research community has recognized this gap. Mahari et al. (2023) identify a persistent disconnect between legal NLP research and actual practitioner needs, noting that the field has focused heavily on benchmark performance while underinvesting in the infrastructure required for trustworthy deployment. Guha et al. (2024) advance evaluation methodology for legal reasoning, but evaluation alone is insufficient—practitioners also need mechanisms to certify that deployed systems operated within documented parameters. Chan et al. (2025) reframe this as a systems-architecture problem, arguing that accountability requires *infrastructure* external to agents that mediates their interactions with human institutions—the CoA is one such infrastructure component.

2.2. Existing Audit and Traceability Approaches

Current approaches to AI agent audit trails fall into three categories, none of which produce the standardized, independently verifiable artifact that legal and regulatory contexts demand.

Platform-specific audit logs. Agent orchestration platforms (e.g., LangChain, AutoGen, CrewAI) maintain internal execution logs capturing tool calls, model responses, and workflow state. These logs are non-standardized across platforms, not independently verifiable by external parties, and typically stored in formats optimized for debugging rather than regulatory production.

Agent observability tools. Specialized observability platforms provide monitoring dashboards, trace visualization, and performance analytics for deployed agents. While valu-

able for operational insight, these tools are watch-only: they observe what agents do but do not *certify* it. They cannot produce a single sealed artifact suitable for regulatory examination or counterparty due diligence. Industry surveys indicate that 62% of production agent teams identify observability as their most urgent investment priority (Cleanlab, 2025), but observability is an input; certification is the missing output.

Blockchain-based approaches. Several proposals anchor AI decision records to public blockchains for immutability and independent verification. While conceptually appealing, blockchain-based approaches face several commonly cited practical barriers in legal contexts: transaction costs at scale, latency incompatible with real-time agent workflows, storage limitations for rich reasoning traces, and poor fit for confidential legal work that cannot be recorded on public ledgers. Architectures that commit only a hash on-chain while keeping content off-chain mitigate the storage and confidentiality concerns; the CoA adopts exactly this commit-only pattern (Section 3.3) without requiring a public blockchain, relying instead on an append-only transparency log.

Verifiable execution kernels. Zhang (2026) propose a runtime sovereignty kernel using RFC 6962 Merkle trees for real-time agent execution control, achieving sub-millisecond verification latency. The CoA addresses the complementary *certification* layer—producing a sealed evidentiary artifact for regulatory production rather than enforcing runtime constraints.

Resource governance frameworks. Ye & Tan (2026) propose Agent Contracts, a formal framework specifying execution bounds for autonomous agents—token budgets, API call limits, temporal deadlines, and success criteria—with empirically validated conservation laws for multi-agent delegation. Where Agent Contracts bound what agents may consume before execution, the CoA certifies what agents did after execution: the frameworks are complementary layers of a complete governance lifecycle.

Externalization and harness engineering. Zhou et al. (2026) review agent infrastructure organized around memory, skills, protocols, and harness engineering, arguing that governance must be "co-designed" with externalization rather than added after the fact (Zhou et al., 2026, §8.4). Their framework positions governance as a runtime-safety property of the harness. The CoA operates one level deeper: rather than treating governance as an internal safety property, we treat the harness as a substrate for legal accountability, operationalizing provenance tracking with the additional requirement of evidentiary admissibility—producing a record designed for third-party scrutiny, not engineering introspection.

Table 1. CoA vs. existing agent audit approaches. "—" = absent; "P" = partial or platform-specific.

Capability	Logs	Obs.	Gov.	CoA
Agent identity	✓	P	✓	✓
Decision chain	✓	✓	✓	✓
Consensus / pass^k	—	—	—	✓
Human review (chained)	P	P	P	✓
Tamper evidence	—	—	P	✓
3rd-party verifiable	—	—	—	✓
Regulatory mapping	—	—	P	✓
Legal temporal proof	—	—	—	✓

Table 1 compares the CoA with representative existing approaches across capabilities relevant to legal and regulatory contexts.

No existing platform combines cryptographic tamper evidence with third-party verifiability and consensus attestation. The CoA is complementary to these tools: it operates as a cryptographic attestation layer that consumes trace data and produces a sealed certificate, rather than replacing operational monitoring.

2.3. Regulatory and Governance Landscape

Three regulatory developments create direct demand for the infrastructure the CoA provides. The **EU AI Act** (European Parliament and Council of the European Union, 2024) (Articles 11–14, 26) imposes documentation, logging, transparency, and human oversight requirements for high-risk AI systems, with the emerging **ISO/IEC 24970** standard defining what to log but not how to ensure integrity. The **NIST AI RMF** (National Institute of Standards and Technology, 2023) and its new **AI Agent Standards Initiative** (CAISI, 2026) explicitly call for tamper-proof agent logging. Singapore's **IMDA framework** (Infocomm Media Development Authority, 2026) introduces agent identity mechanisms for prospective agent governance. We map the CoA to these frameworks in detail in Section 4.

3. The Certificate of Action

Definition 3.1 (Certificate of Action)**.** A Certificate of Action (CoA) is a tamper-evident, cryptographically sealed record that certifies an autonomous AI agent workflow by capturing: (i) what every agent in the workflow did and why, (ii) under what policy authority each agent operated, (iii) what controls were applied, and (iv) whether and how a human reviewed the outcome. The CoA is designed to be independently verifiable by third parties without requiring access to the originating platform.

Unit of certification. A CoA covers a single bounded workflow: the sequence of agent and human actions initiated by one principal instruction and terminating at a completion event, both sealed at Layer 1. It is not a certificate of an

isolated model call, nor of an entire engagement or matter. The workflow boundary is fixed at capture and cannot be redrawn after sealing—a definitional point that matters for the completeness analysis in Section 4.3.

3.1. Architecture

The CoA captures four layers of evidence, each building on the previous to create a complete governance record.

Layer 1: Agent Decision Chain. Each agent in the workflow is recorded with its model identifier and version, the policy or instruction set it operated under, the evidence or data it considered, its confidence level or calibration metrics, and its reasoning output. Entries are linked in a cryptographic chain: each entry includes a cryptographic fingerprint of the preceding entry, creating a tamper-evident sequence. Modifying any element invalidates all subsequent entries.

A contract review pipeline illustrates the structure: entity extraction, clause risk assessment, precedent comparison, and compliance verification each produce a discrete, auditable record linked to the full execution context.

Layer 2: Consensus Verification. When agent workflows employ multiple independent reasoning passes over the same evidence (a technique sometimes called pass^k verification), the CoA records the number of passes, the agreement rate, and any divergences. High consensus (e.g., five of five passes agree) indicates strong confidence. Low consensus (e.g., two of five) is itself critical information for reviewers and is recorded as such. This layer is particularly relevant for legal applications where independent verification of reasoning is a professional norm.

Layer 3: Human Review Record. When agent decisions are escalated for human review—whether by policy or by confidence threshold—the CoA captures who reviewed the decision, when, what they decided, and their stated rationale. The human review record is cryptographically chained onto the agent decision chain: the two cannot be separated. This ensures that the governance record is complete from autonomous decision through human oversight.

Layer 4: Execution Record. For workflows that produce actionable outputs (a signed agreement, a compliance determination, a lending decision), the CoA captures the execution lifecycle: document creation, routing, delivery, review, execution, and completion. The executed artifact and the agent governance record are sealed together, creating a unified record linking the decision process to its outcome.

Figure 1 illustrates the four-layer structure.

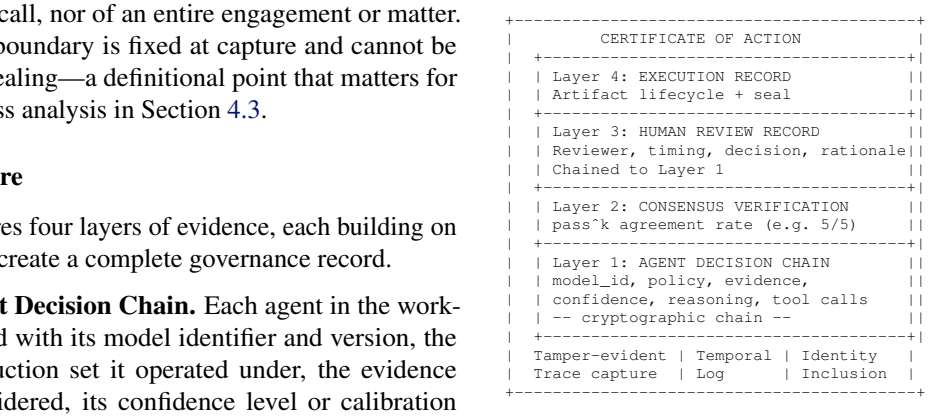

Figure 1. CoA architecture: four evidence layers sealed by open-standard cryptographic primitives, linked by tamper-evident hash chaining. Only the root hash is committed to the verifiable log.

3.2. Standards Foundation

A critical design principle of the CoA is that it requires no novel cryptographic or protocol infrastructure. The architecture draws on existing open standards organized around four required properties, spanning two tiers of maturity.

The first tier covers battle-tested standards with decades of production deployment:

Tamper-evident sequencing—a cryptographic chaining mechanism that links each decision record to the previous one such that modifying any entry invalidates all subsequent records. The underlying primitive is the same one used in code signing and distributed consensus systems.

Temporal integrity—a trusted timestamping mechanism that provides legally recognized proof of the time of certificate creation, compatible with qualified electronic signatures under eIDAS and equivalent frameworks in other jurisdictions.

Identity binding—a certificate-based mechanism that cryptographically binds the certifying platform's identity to the record, using the same public-key infrastructure that secures authenticated communications globally.

Independent verification—an append-only public logging architecture that allows any third party to confirm that a specific record existed and was publicly committed at the time of issuance, without contacting or trusting the originating platform.

The second tier covers emerging industry conventions with broad vendor participation but experimental stability:

Trace capture—a vendor-neutral standard for capturing AI agent behavior—prompts, responses, tool calls, reasoning chains—currently at experimental stability, with active development of agent-specific conventions. The CoA speci-

fication aligns with the direction the observability community is converging toward while maintaining architectural independence: the four integrity properties above do not depend on the trace capture format and can accommodate alternative or future conventions.

This tiered approach ensures that the CoA's core security properties rest on well-established foundations while its interoperability layer adapts as the GenAI observability specification matures.

3.3. Verification Model

The CoA is designed for third-party verification without requiring trust in or access to the originating platform. The verification architecture follows a hash-only commitment design: CoA records may contain variable-length reasoning traces and confidential material, so only the certificate's root hash is committed to the verifiable log, with the full record maintained by the certifying platform. This design enables independent verification without public disclosure of sensitive content—an important property for legal workflows.

The verification model provides four independent assurance guarantees, each confirmable by any third party without contacting the originating platform: that the document has not been altered since sealing (tamper-evidence, confirmed by verifying the internal hash chain); that it existed at the stated time as confirmed by an authority independent of the certifying platform (temporal integrity); that it was issued by the stated platform (identity binding, confirmed by verifying the platform's cryptographic signature); and that it was publicly committed at the time of issuance (log inclusion, confirmed against a publicly auditable append-only log).

A certificate that passes all four checks cannot have been created retroactively, backdated, impersonated, or privately circulated without public commitment.

Legal enforceability is supported by existing frameworks: the ESIGN Act and UETA in the United States and eIDAS (European Parliament and Council of the European Union, 2014) in the European Union already provide the legal basis for electronic records and signatures meeting requirements of intent, association, authentication, and retention. The CoA's structure aligns with Federal Rules of Evidence 902(13) and 902(14), which provide for self-authentication of records generated by electronic processes with certification from a qualified person. The emerging "deference to audit" doctrine proposed for judicial review of AI-assisted decisions (Caputo, 2026) further motivates standardized certification infrastructure: courts deferring to audit require an auditable artifact to defer *to*.

3.4. Who Controls the Certificate

The guarantees above are cryptographic; on their own they say nothing about who operates the certifying platform relative to the parties in dispute. We distinguish three configurations. In the *self-hosted* configuration, the deploying firm issues its own CoAs: the hash chain remains tamper-evident, but the firm also controls which workflows are sealed, so a self-issued CoA carries the evidentiary weight of a self-signed certificate—useful for internal governance, weaker under adversarial scrutiny. In the *independent-certifier* configuration, a third party analogous to a notary or an RFC 3161 time-stamping authority seals workflows without holding their confidential content; the hash-only commitment of Section 3.3 is what makes third-party certification possible without disclosure. In the *regulated-infrastructure* configuration, a supervisory regime designates or accredits issuers. The independent and regulated configurations are the design targets for high-stakes legal use; the self-hosted configuration is an adoption ramp. The locus of control is orthogonal to the cryptography—the primitives are identical across all three—but it determines how much a given CoA is worth to a hostile reader.

3.5. Interactive Demonstration

To illustrate the CoA architecture and its tamper-evidence property, we developed an interactive demonstration. The demonstration implements the hash-chain layer described in Section 3.1 using real cryptographic computation; the additional integrity primitives described in Section 3.2 (trusted timestamping, identity binding, verifiable log enrollment) remain out of scope and are discussed in Section 5.3.

Tamper-evidence demonstration. The first demonstration implements the five-agent consumer lending pipeline described above, with each agent recording its model identifier, policy reference, confidence score, pass^k consensus metric, reasoning trace, and tool calls. The hash chain uses real cryptographic computation via the Web Crypto API. Users can modify any field in any agent's record and observe the tamper-evidence property in real time—the modified agent's hash changes and all downstream hashes cascade to invalid. Three guided scenarios illustrate domain-specific tampering: hiding a dissenting reasoning pass, inflating a credit score, and softening a compliance finding. The interactive demonstration makes this cascade directly observable: modifying Agent 3's consensus metric (e.g., 4/5 $\rightarrow$ 5/5) invalidates Agent 3's hash and causes all downstream agents (4–5) to fail verification while upstream agents (1–2) remain unaffected—confirming that "modifying any element invalidates all subsequent entries" (Section 3.1) is a verifiable behavior, not merely a stated property.

Regulatory compliance mapping. The second demonstration provides an interactive exploration of the regula-

tory mappings in Tables 2 and 3, allowing users to navigate EU AI Act articles and NIST AI RMF functions with corresponding CoA components highlighted. Article 12 (Record-keeping) receives expanded treatment, showing how ISO/IEC 24970 defines the information model while the CoA provides the integrity layer—together forming a complete compliance stack. A governance stack view visualizes the three-pillar argument: ISO/IEC 24970 (what to log), Singapore IMDA (who agents are), and CoA (what agents did).

Implementation scope and limitations. The demonstration implements hash chain construction with real cryptographic computation but does not implement trusted timestamping, identity binding, trace formatting, or verifiable log enrollment. Agent data is static rather than LLM-generated. Closing this gap requires integrating four additional components with mature open-source implementations, estimated at 750–1,400 lines of integration code.

The interactive demonstration and source code are available as supplementary material.

4. Regulatory Compliance Mapping

The following tables map CoA components to specific requirements under the EU AI Act and the NIST AI Risk Management Framework.

4.1. EU AI Act Compliance

Table 2 maps each relevant EU AI Act requirement to the CoA component that satisfies it.

Article 12 (Record-keeping) requires that logging "shall be appropriate to the intended purpose of the high-risk AI system" and enable "monitoring of the operation" and "post-market monitoring." The emerging ISO/IEC 24970 standard provides an information model for AI system event logging designed to operationalize these requirements—defining event types, risk-driven logging triggers, and auditability criteria. However, ISO/IEC 24970, in its current Draft International Standard form, focuses on *what* to log rather than *how* to ensure log integrity. The CoA addresses this complementary concern: its cryptographic hash chain provides the tamper-evidence guarantee that Article 12 requires but that ISO/IEC 24970 does not specify. The CoA architecture is designed to be compatible with the standard's information model while adding the integrity layer. Together, they form the components of a complete Article 12 compliance stack.

Article 14 (Human oversight) requires measures enabling individuals to understand system capacities and interpret outputs. The CoA's Human Review Record captures the full chain from agent decision through human intervention, connecting autonomous action to human judgment in a single artifact.

Article 26 (Deployer obligations) requires deployers to maintain logs under their control. The CoA's third-party verification model ensures that a regulator need not rely solely on the deployer's own log retention.

4.2. NIST AI RMF Compliance

Table 3 maps CoA components to the four core functions of the NIST AI Risk Management Framework (National Institute of Standards and Technology, 2023) and relevant subcategories from the Generative AI Profile (National Institute of Standards and Technology, 2024).

The NIST framework's emphasis on provenance[1] is well-served by the CoA architecture. The Generative AI Profile (National Institute of Standards and Technology, 2024) calls for documenting "the lineage of AI-generated content"; the CoA's agent decision chain extends this lineage from model development into operational deployment. NIST's AI Agent Standards Initiative calls for records of "what agents were allowed to do, what decisions they made, and whether a human approved or overrode actions" (National Cybersecurity Center of Excellence, 2026)—precisely what the CoA captures.

4.3. Gap Analysis

The regulatory mapping reveals a clear perimeter. No existing approach combines standardized format, independent third-party verifiability, bilateral trust, and legally recognized temporal and identity proofs in a single artifact—that gap is what the CoA fills.

What the CoA does not fill is equally important. The CoA certifies that a process occurred as documented, but does not independently verify the *correctness* of the underlying AI reasoning. A CoA can record that an agent identified a clause as high-risk with 92% confidence, but cannot independently confirm that the risk assessment was substantively correct. Similarly, the CoA does not address training data governance, model bias detection, or real-time monitoring—it is complementary to, not a replacement for, broader AI governance frameworks. This boundary has theoretical grounding: Tibebu (2026) proves that beyond a computable autonomy threshold, no accountability framework can simultaneously satisfy attributability, foreseeability, and completeness for human-agent collectives—the CoA operates within the tractable regime below this threshold, where certification remains feasible.

Completeness versus integrity. Tamper-evidence estab-

[1]We use "provenance" in the sense of Zhou et al. (2026, §8.4): tracking decision-chain modifications across time, with the additional requirement of legal-evidentiary admissibility.

Table 2. Mapping of Certificate of Action components to EU AI Act requirements for high-risk AI systems.

Requirement	Article	Regulatory Obligation	CoA Component
Technical documentation	Art. 11(1)	Provide documentation of system capabilities, limitations, intended purpose, and risk assessment prior to deployment	**Agent Decision Chain**: model identifiers, version, policy set, and operating parameters recorded per agent
Record-keeping	Art. 12(1)–(2)	Enable automatic recording of events (logs) for traceability throughout the system lifecycle	**Full CoA**: tamper-evident, timestamped, immutable event sequence across all four layers
Transparency	Art. 13(1)	Design systems to enable deployers to interpret output and use the system appropriately	**Consensus Verification + Human Review Record**: agreement rates, confidence levels, escalation rationale, and human decision reasoning
Human oversight	Art. 14(1), (4)	Enable effective human oversight, including ability to intervene, override, or halt the system	**Human Review Record**: cryptographically chained record of human intervention, including reviewer identity, timing, decision, and rationale
Deployer obligations	Art. 26(5)–(6)	Monitor operation, maintain logs, document incidents, report serious incidents to authorities	**Verification API + Log Registry**: independently verifiable certificate archive; incident-linked CoAs enable regulatory production
Risk management	Art. 9(1)–(4)	Establish and maintain a continuous risk management system throughout the AI lifecycle	**CoA corpus as audit artifact**: aggregate analysis of CoAs enables pattern detection, risk trending, and compliance evidence

lishes that a sealed record was not altered; it does not establish that all relevant records were sealed. A party running a workflow repeatedly could seal and submit only the most favorable run, and each submitted CoA would verify perfectly while the selection among runs went uncertified. Integrity at the level of a single certificate is therefore distinct from completeness or representativeness at capture. The self-hosted configuration (Section 3.4) cannot close this gap on its own; an independent or regulated certifier that seals every workflow initiated—not only those a deployer later chooses to produce—is what converts per-certificate integrity into corpus-level completeness. Within a single sealed workflow, the Layer 2 consensus record narrows a related risk: a dissenting reasoning pass captured before sealing cannot later be hidden, though omission *before* capture remains outside the certificate's guarantee. In short, the CoA certifies the integrity of what was sealed; it does not certify that everything was sealed.

Certification, not reproduction. The CoA is a certificate of what occurred, not a reproducibility instrument. Given the non-deterministic nature of large language model inference, re-running a workflow under identical parameters may produce different outputs; the CoA certifies the record of the actual run rather than a claim that the run reproduces.

This boundary is intentional: the CoA solves the *certification* problem ("prove what your agents did") while leaving the *evaluation* problem ("prove your agents were right") to the benchmarking and evaluation research that this workshop community is actively advancing (Guha et al., 2024; Katz et al., 2024). The CoA also complements *identity* frameworks such as Singapore's agent identity mechanisms (Infocomm Media Development Authority, 2026): identity attestation describes what agents are authorized to do, while the CoA certifies what they actually did. Both are necessary components of a complete governance stack.

5. Discussion

5.1. Implications for Legal AI Deployment

The CoA changes the deployment calculus for legal AI in a specific way: it decouples the question of agent capability from the question of agent governance. Currently, organizations considering agentic AI for legal workflows face a compound risk assessment: is the model capable enough *and* can we govern its deployment? Without standardized governance infrastructure, the second question frequently blocks deployment even when the first is answered affirmatively. Recent practitioner surveys suggest that while broad AI access is widespread in legal organizations, high trust in AI outputs remains uncommon, and trust gaps correlate with lower reported return on investment.

The CoA provides a governance artifact that enables organizations to deploy with confidence. The confidence is procedural, not substantive: every decision is documented, timestamped, independently verifiable, and linked to human oversight. The risk profile shifts from "we cannot prove

Table 3. Mapping of Certificate of Action components to NIST AI Risk Management Framework functions.

Function	Subcategory	Framework Guidance	CoA Component
GOVERN	1.1, 1.2	Legal and regulatory compliance documentation; organizational AI policies and procedures	Agent Decision Chain: policy identifiers link each agent action to the organizational policy it operated under
MAP	3.4, 3.5	AI system provenance tracking; documentation of system limitations and assumptions	Full CoA: captures provenance chain from input data through agent reasoning to output; records model versions and known limitations
MEASURE	2.5, 2.6	Effectiveness evaluation; assessment of AI system trustworthiness characteristics	Consensus Verification: pass^k agreement rates provide a built-in measure of reasoning reliability per decision
MANAGE	1.1, 1.3	Risk response and incident documentation; third-party risk management	Verification API: incident-linked CoAs enable systematic response; third-party verification enables cross-organizational risk assessment

what happened" to "we can prove exactly what happened, and external parties can verify our proof."

The adoption timeline faces a structural tension. Under the 2026 Digital Omnibus, the EU AI Act's high-risk obligations—including the Article 12 logging requirements—take effect on 2 December 2027 for standalone Annex III systems (2 August 2028 for AI embedded in Annex I regulated products), while the harmonized standard (ISO/IEC 24970) remains at Draft International Standard stage and a large share of enterprises have not taken meaningful compliance steps (Vision Compliance, 2026). This standards gap creates an opportunity: the CoA specification offers a reference architecture built on existing open standards that could inform the emerging landscape while providing compliance value ahead of the deadline.

We treat the CoA as a technical specification rather than a regulatory instrument. Hash chains, trusted timestamps, and digital signatures were each developed as voluntary standards before being incorporated into legal regimes such as eIDAS and ESIGN; we expect a comparable path for the CoA—voluntary adoption, then industry standardization, then regulatory incorporation—with the EU AI Act's Article 12 obligation supplying the compliance demand.

This approach aligns with practitioner experience that structured workflows with governance, auditability, and human review outperform maximally autonomous agents in regulated contexts. The most effective legal AI systems constrain autonomy rather than maximize it, building governance into the workflow itself rather than treating it as an afterthought.

5.2. Access to Justice Implications

The CoA has implications for AI and access to justice. Legal aid organizations face resource constraints that make AI-assisted legal work attractive, yet they are among the most risk-averse adopters due to confidentiality and liability concerns. A CoA provides a verifiable governance record that reduces the informational asymmetry between deploying organizations and their oversight bodies.

Trust bootstrapping takes time—the ESIGN Act (United States Congress, 2000) (2000) preceded widespread practitioner acceptance of electronic signatures by approximately fifteen years. The CoA accelerates this by aligning with existing evidentiary frameworks, particularly FRE 902(13)/(14) for self-authentication of electronic records. ABA Formal Opinion 512 (2024), establishing competence and supervision duties for AI use in legal practice, further supports the case for standardized governance artifacts.

5.3. Limitations

This work has several limitations that we state explicitly.

First, this is a position paper presenting an architectural proposal without empirical validation. Preliminary cost analysis suggests infrastructure costs of $50–800 per month for 10,000 daily certificates, depending on whether per-certificate or batch timestamping is used, but these estimates have not been benchmarked at scale.

Second, the CoA addresses documentation and certification but not the correctness of underlying AI reasoning. A perfectly sealed CoA can certify a flawed decision chain. The CoA is necessary but not sufficient for trustworthy legal AI—it must be deployed alongside robust evaluation frameworks.

Third, adoption requires bilateral trust that takes time to build. A CoA is only valuable if the party receiving it trusts the entity that produced it. The electronic signature prece-

dent suggests a timeline of approximately five years from legal recognition to broad practitioner adoption, potentially compressed by the regulatory urgency of the EU AI Act. Institutional credibility, analogous to that built by electronic signature platforms over decades, cannot be manufactured overnight.

Fourth, the industry conventions for AI agent trace capture remain at experimental stability (Section 3.2), introducing migration risk for early implementations.

Fifth, this work reflects a practitioner perspective from a single jurisdiction and organizational context. Cross-jurisdictional applicability—particularly the interaction between CoA and varying legal frameworks for electronic evidence—requires further analysis.

5.4. Open Questions

Three questions remain for the research community. Attorney-client privilege presents the most immediate design tension: what level of reasoning abstraction preserves audit value without exposing protected analysis? A related procedural question is distinct from the technical one the CoA resolves: the certificate settles what can be proven, not who may compel production of the full record. Discovery rules, regulatory examination authority, and privilege govern that access; the hash-only public log is designed to be compatible with those regimes—preserving privilege while still allowing any third party to verify a record's existence and integrity—but a full treatment is future work. Standards convergence is the longer-horizon challenge, particularly how the CoA should interoperate with ISO/IEC 24970 and the forthcoming SP 800-53 agent control overlays—formal interoperability testing would accelerate adoption across both. The empirical gap is harder to specify: what methodology would actually test whether CoA adoption improves practitioner trust, regulatory compliance outcomes, or deployment confidence?

Governance quality as a measurable harness property. Zhou et al. (2026) (§8.6) propose extending agent evaluation to include governance quality but do not specify how to measure it. The CoA provides one candidate benchmark: a harness can be evaluated against four criteria the certificate makes verifiable on demand—hash-chain integrity of the decision sequence; capture of the operating policy version; recorded human review at designated approval gates; and admissibility properties consistent with the relevant evidentiary regime (e.g., FRE 803(6) and 902(13)–(14) for U.S. courts). We propose this four-pronged test as a starting point for governance-quality benchmarks, directly responsive to the open challenge Zhou et al. (2026) identify.

6. Conclusion

We have introduced the Certificate of Action, a trust primitive for autonomous legal AI agent workflows. The CoA provides a standardized, tamper-evident, independently verifiable record of agent decisions, built on existing open standards and mapped to specific requirements of the EU AI Act and NIST AI Risk Management Framework. We have shown that the CoA fills a specific gap in the emerging AI governance landscape: ISO/IEC 24970 defines what to log, Singapore's agent identity mechanisms define who agents are, and the CoA certifies what agents did—with cryptographic integrity that existing observability platforms do not provide.

The central argument of this paper is that trust infrastructure for legal AI is as critical as model capability. The legal AI research community has made significant progress on benchmarking, evaluation, and reasoning quality (Guha et al., 2024; Katz et al., 2024; Chalkidis et al., 2022). These advances enable us to build more capable legal AI systems. But capability without accountability is insufficient for a domain where errors carry direct legal and economic consequences for real people.

The question this workshop poses—"what does it mean for an AI system to be truly competent in law?"—cannot be answered by model performance alone. Competence in law requires not just correct outputs but verifiable, trustworthy processes that can be examined by regulators, counterparties, and courts. The Certificate of Action is a proposal for one piece of that infrastructure. We invite the ML and legal research communities to formalize, evaluate, and extend it.

References

Arbel, Y., Goldstein, I., and Salib, P. How to count AIs: Individuation and liability for AI agents. *arXiv preprint arXiv:2603.10028*, 2026.

Caputo, N. Administrative law's fourth settlement: AI and the capability-accountability trap. *SSRN Electronic Journal*, 2026. SSRN: 6049814; arXiv:2602.09678.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., and Anderljung, M. Infrastructure for AI agents. *arXiv preprint arXiv:2501.10114*, 2025.

Cleanlab. AI agents in production 2025: Enterprise

trends and best practices. https://cleanlab.ai/ai-agents-in-production-2025/, 2025.

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 (eIDAS). Official Journal of the European Union, 2014.

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

Guha, N., Nyarko, J., Ho, D. E., Ré, C., et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Infocomm Media Development Authority. Model AI governance framework for agentic AI. Singapore Infocomm Media Development Authority, 2026.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270), 2024.

Kolt, N. Superintelligence and law. *Harvard Journal of Law & Technology*, 2026. Forthcoming. Available at SSRN: https://ssrn.com/abstract=6302179.

Mahari, R., Stammbach, D., Ash, E., Ho, D. E., et al. The law and NLP: Bridging disciplinary disconnects. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

National Cybersecurity Center of Excellence. Accelerating the adoption of software and AI agent identity and authorization. Concept Paper, National Cybersecurity Center of Excellence (NCCoE), NIST, 2026.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023.

National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative AI profile. Technical Report NIST AI 600-1, U.S. Department of Commerce, 2024.

O'Keefe, C. and Frazier, T. Automated compliance and the regulation of AI. *SSRN Electronic Journal*, 2026. SSRN: 6017756.

Pandey, R. The agentic AI governance framework: A universal model for risk, accountability, and compliance in autonomous systems. *SSRN Electronic Journal*, 2026. SSRN: 5652350.

Surden, H. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35(4), 2019.

Tibebu, H. The accountability horizon: An impossibility theorem for governing human-agent collectives. *arXiv preprint arXiv:2604.07778*, 2026.

United States Congress. Electronic signatures in global and national commerce act (ESIGN). 15 U.S.C. §7001 et seq., 2000.

Vision Compliance. EU AI act readiness report. Vision Compliance, 2026.

Ye, Q. and Tan, J. Agent contracts: A formal framework for resource-bounded autonomous AI systems. In *Proceedings of the 16th International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE 2026), co-located with AAMAS 2026*, 2026. arXiv:2601.08815.

Zhang, J. Right to history: A sovereignty kernel for verifiable AI agent execution. *arXiv preprint arXiv:2602.20214*, 2026.

Zhou, C., Chai, H., Chen, W., Guo, Z., Shan, R., Song, Y., Xu, T., Yang, Y., Yu, A., Zhang, W., Zheng, C., Zhu, J., Zheng, Z., Zhang, Z., Lou, X., Zhang, C., Fu, Z., Wang, J., Liu, W., Lin, J., and Zhang, W. Externalization in LLM agents: A unified review of memory, skills, protocols and harness engineering. *arXiv preprint arXiv:2604.08224*, 2026.