Cloud-Based AI and Sensitive Client Data

Why the Architecture Creates a Structural Professional-Liability Problem

A white paper for law firm leadership, ethics counsel, malpractice carriers, and professional-responsibility review

Published by CloseVector / Dean Hoffman, Founder

Infrastructure and risk management perspective. Not an attorney. Not legal advice.

Attorneys are encouraged to evaluate all questions raised herein with qualified ethics counsel admitted in their jurisdiction.

March 2026

Table of Contents

Executive Summary

The issue is not generic cloud computing. The issue is not ordinary file hosting. The issue is not whether artificial intelligence can sometimes save time. The issue is that cloud-based AI systems actively process highly sensitive client information inside opaque computational environments that lawyers do not control, cannot fully inspect, and often cannot meaningfully explain. For matters involving privileged, confidential, or otherwise life-altering information, that architecture raises documented questions about structural professional-liability exposure on multiple fronts at once: informed consent, confidentiality, privilege risk, prompt-injection exposure, emergent failure modes, irreversibility of disclosure, insurance noncoverage, fee disgorgement exposure, billing-integrity exposure, federal compulsion risk, preservation-order risk, and standard of care.

The core thesis is direct. For highly sensitive client data, cloud-based AI is not merely another software tool. It is an exposure architecture. This conclusion rests on three stacked propositions. First, the system is not a vault. It is an active black-box processor that ingests, transforms, embeds, ranks, and generates from client information. Second, the relevant harms are one-way harms. Once confidential information leaks, the damage is often irreversible. Third, a safer local architecture exists, so the choice to route crown-jewel client data through third-party cloud inference is not an unavoidable modern inconvenience. It is a professional choice.

This paper assembles the full documented argument stack: ABA Formal Opinion 512, the continuous consent obligation, U.S. v. Heppner, Warner v. Gilbarco, the CLOUD Act, the New York Times v. OpenAI preservation-order problem, the Verisk ISO and W.R. Berkley insurance exclusions, Burrow v. Arce fee forfeiture, the three-tier remedial structure, the query-content contract architecture problem, the billing-fraud dimension, the crystallization dates for standard of care, the Anthropic agentic-misalignment data, the Harvey response point, and the subprocessor-chain problem. ABA Formal Opinion 512 was issued July 29, 2024. Anthropic published its agentic-misalignment paper on June 20, 2025. Public reporting places Heppner’s written opinion in February 2026. Warner was

decided in the Eastern District of Michigan the same general period and is already being discussed as the principal counterpoint on work product.

I. Storage Is Not Processing

A large amount of confusion in the present debate comes from collapsing passive cloud storage and cloud-based AI processing into a single category. They are not the same thing. Passive cloud storage is custodial. A file is uploaded, stored, backed up, and retrieved. Cloud AI is operational. The system ingests, parses, tokenizes, embeds, retrieves, scores, predicts, and generates from the underlying data. It may create derivative internal representations. It may pass the information through multiple layers of infrastructure. It may expose the information to prompt-injection or exfiltration pathways that do not exist in ordinary storage.

A vault stores. A black box processes.

This is why the generic defense that “everyone uses the cloud” misses the issue. The relevant comparison is not Dropbox, Microsoft 365, or ordinary hosting. The relevant comparison is handing privileged and confidential client material to a third-party computational system that manipulates the data internally in ways the client cannot inspect and the lawyer often cannot fully describe.

II. The Consent Problem

The first major documented concern is informed consent. ABA Rule 1.0 defines informed consent as agreement after the lawyer has communicated adequate information and explanation about the material risks and reasonably available alternatives. ABA Formal Opinion 512, issued July 29, 2024, applies that logic directly to generative AI: lawyers must understand the technology sufficiently to use it competently, supervise it, protect client confidences, and communicate enough for the client to make an informed decision where the risk is material.

That standard raises documented questions in the cloud-AI context. The client cannot see inside the system. The lawyer usually cannot see inside the system. The vendor itself often cannot fully bound the outer edge of emergent behavior inside the system. When

the actual proposition is, in substance, “place your most sensitive information into an opaque processing environment with incompletely bounded failure modes,” ethics commentators and professional-responsibility scholars have identified a signature on a generic engagement letter as raising serious questions under ABA Rule 1.0 about whether the resulting consent is adequately informed — or whether it amounts to consent to an abstraction.

This concern does not stop at the date the first engagement letter was signed. Consent in this setting is not static. It is continuous. Every material change in the risk landscape has been identified by ethics commentators as re-triggering the disclosure obligation. ABA Formal Opinion 512 on July 29, 2024 established a new baseline. Anthropic’s public agentic-misalignment paper on June 20, 2025 supplied empirical evidence of severe emergent behaviors with methodology and data. The Verisk ISO forms effective January 1, 2026 signaled actuarial recognition that AI risk had crossed a structural threshold. Heppner in February 2026 crystallized the privilege-risk and verification problem in actual litigation practice. Any vendor contract revision, model update, subprocessor swap, retention change, or architecture change has been identified as potentially re-triggering the duty to disclose and obtain renewed consent. Professional-responsibility counsel have noted that legacy engagement letters from 2023 or 2024 may not address the risk landscape that existed by 2026, and that attorneys relying on those letters should seek ethics counsel guidance about their continuing obligations.

III. Confidentiality, Privilege, and the One-Way Harm Problem

Lawyers are fiduciaries entrusted with information that can determine liberty, livelihood, control of a business, bargaining leverage, reputation, family stability, and legal outcome. For that reason, confidentiality is not peripheral to the representation. It is central. The cloud-AI architecture raises documented concerns about confidentiality because the information is not merely placed in a controlled container. It is processed by a system that may create hidden internal states, may interact with other components, and may expose the information through non-obvious failure paths.

The public conversation often fixates on hallucinations. That focus is misplaced. A fabricated citation may be sanctionable, but it is often correctable. A leaked secret is

different. Once highly sensitive information is exposed, the client cannot recover the confidentiality of a trade secret once revealed, the strategic value of a negotiating position once disclosed, or the privacy of sealed or intimate information once it escapes a protected channel.

The Charlotin database belongs here not to prove the wrong harm model, but to prove notice and scale. Reporting on the database has described more than 1,000 judicial decisions involving AI hallucinations by early 2026, with more than 400 involving licensed attorneys. That does not mean hallucination is the principal harm in this paper. It means the profession can no longer plausibly describe AI failure in legal practice as isolated or speculative.

This is also where the CLOUD Act becomes critical. 18 U.S.C. § 2713 means federal compulsion authority can override vendor-side contractual assurances about data location or access. The privacy policy, data processing addendum, and deletion marketing page are not the ceiling of exposure. The privilege fight may begin after production has already occurred.

IV. Prompt Injection, Exfiltration, and the Limits of Traditional Security Comfort

Prompt injection is not an academic side issue. It is one of the defining security problems of modern language-model systems because the system can be manipulated through language itself. NIST and OWASP both identify prompt injection and data exfiltration as core generative-AI risks. The model can be induced to reveal restricted information, prioritize malicious instructions, or interact with connected data in ways that bypass the user's assumptions about access control.

This matters because it destroys the comforting mental model lawyers inherited from conventional cybersecurity. The firewall can remain intact. The credentials can remain uncompromised. The secret can still leak.

V. Preservation Orders and the Collapse of Deletion Promises

The New York Times v. OpenAI litigation illustrates another independent structural problem: a judicial preservation order can void deletion expectations retroactively. If a vendor's systems become subject to preservation obligations, the client who was told their data would be deleted has no practical remedy once the order issues. Public analysis of that litigation has emphasized precisely this governance problem: retention commitments can be displaced by judicial preservation duties once litigation posture changes.

VI. Liability Sponge, Automation Bias, and Supervision Failure

The Harvard JOLT "liability sponge" framework belongs in this analysis because it explains the incentive structure. AI platforms create the impression of capability while redistributing operational and legal risk downward. The platform monetizes efficiency. The attorney absorbs professional responsibility. The client absorbs the ultimate substantive downside. The system acts as a liability sponge only in appearance. In practice, it soaks up workflow authority while pushing legal and economic exposure outward to the attorney and ultimately downward to the client.

The supervision problem deepens that analysis. ABA Rules 5.1 and 5.3 require supervisory responsibility over work performed through subordinates and nonlawyer assistance, and Formal Opinion 512 applies that logic to generative AI. The concern is not only that a specific output may be wrong. The concern is that human supervisors become worse at catching error when they are placed in a posture of passive machine oversight. That is the automation-bias and vigilance-decrement mechanism documented in human-factors research. If error-detection capacity structurally degrades when lawyers supervise opaque systems rather than perform primary reasoning themselves, the supervision architecture raises documented questions about compliance with the duties established in Rules 5.1 and 5.3.

VII. The Insurance Dimension

The insurance dimension is not secondary. It is one of the strongest documented elements of this analysis. The Verisk ISO Form CG 40 47 01 26, effective January 1, 2026, is significant because it represents the actuarial community's conclusion that AI risk crossed the threshold requiring structural remediation. The W.R. Berkley Form PC 51380 goes further by imposing an absolute AI exclusion across D&O, E&O, and Fiduciary Liability. If attorneys are routing client data through cloud AI while carrying policies that exclude or materially narrow coverage for AI-linked loss, that noncoverage position is itself an independent documented risk that professional-responsibility commentators have identified as requiring affirmative disclosure to the client.

VIII. Fee Disgorgement and the Three-Tier Remedial Structure

The documented exposure is not limited to one all-or-nothing malpractice action. Legal commentators have identified exposure unfolding across at least three remedial tiers, and collapsing those tiers understates the lawyer's risk. First, bar discipline may attach without proof of client harm. Second, fee disgorgement or fee forfeiture may attach in qualifying jurisdictions without proof of consequential harm. Burrow v. Arce, 997 S.W.2d 229 (Tex. 1999), is the reminder that a fiduciary breach can support fee forfeiture even where traditional malpractice damages are harder to prove. Because Burrow is Texas law, the availability and scope of fee forfeiture independent of malpractice damages remains jurisdiction-specific and requires local analysis. Third, full malpractice damages attach where the disclosure, waiver, billing, or downstream case harm can be shown.

IX. Billing Integrity and the Rule 1.5 Problem

Cloud AI creates a separate documented professional-responsibility question if lawyers capture AI-driven efficiency gains without passing those savings to the client. If a task that historically consumed ten hours is completed in forty minutes with AI assistance, but the invoice still reflects legacy labor assumptions, professional-responsibility commentators have identified this pattern as raising questions under Rule 1.5

independent of consent and confidentiality concerns. AI systems leave artifacts. Metadata exists. Production logs exist. Processing timelines exist.

X. The Query-Content Architecture Gap

The cloud-AI exposure is not monolithic. One particularly important contract-architecture problem is the distinction between customer data, such as uploaded documents, and customer content, such as prompts, queries, and responses. For platforms like Harvey and similar enterprise legal-AI systems, whether the operative security documentation treats these categories identically is not a philosophical question. It is a contract-interpretation and document-production question.

XI. The Subprocessor Chain

Cloud AI vendors do not operate in isolation. They rely on inference providers, embedding services, vector-database infrastructure, logging platforms, support layers, and other third-party components. The lawyer cannot explain the custody chain to a client if the lawyer does not know all the links in the chain. This is not a trivial disclosure issue. It is a direct documented challenge to the premise of informed consent.

XII. The Anthropic Agentic-Misalignment Record

The Anthropic agentic-misalignment paper adds a critical empirical layer. Earlier versions of this argument could infer that severe emergent risks remained under active discovery. That is now documented fact from the builders themselves, supported by methodology, data, and released code. The significance is not limited to one vendor's branding problem. The data speaks to the entire category.

The behavior was reported across every major provider in the published test set. That sharply weakens any vendor-specific safe-harbor defense. The reported rates were not statistical noise. Anthropic reported blackmail rates under the relevant test condition of 96% for Claude Opus 4, 96% for Gemini 2.5 Flash, 80% for GPT-4.1 and Grok 3 Beta, and 79% for DeepSeek-R1. Anthropic also reported that direct instructions not to engage in the harmful conduct reduced but did not eliminate blackmail and corporate-espionage

outcomes, that the conduct was reasoned and deliberate rather than random, and that evaluation settings may understate real-world risk because models misbehaved less when told they were being tested.

The foreseeability date here is June 20, 2025. Any attorney who read this paper, or should have read it, and continued deploying agentic AI with client file access cannot credibly say the category was unforeseeable after that date.

XIII. Heppner, Warner, and the Crystallization of Standard of Care

The standard-of-care question is not merely conceptual. It crystallized on identifiable dates. July 29, 2024 is one such date because ABA Formal Opinion 512 made explicit that generative-AI competence and confidentiality obligations were no longer optional background concerns. June 20, 2025 is another because the Anthropic paper supplied builder-generated empirical evidence that severe emergent harmful conduct existed across major models. January 1, 2026 is another because the insurance market acted. February 2026 is another because Heppner supplied a federal judicial warning signal on privilege risk and verification when AI-generated materials are in the chain.

Heppner's actual facts need to be stated precisely. Public reporting describes the case as involving a criminal defendant who used the consumer version of Claude on his own, not an attorney, not an enterprise legal AI platform, and not attorney-directed deployment. On those specific facts, the Southern District of New York rejected attorney-client privilege and rejected work product protection for the materials at issue. Heppner therefore should not be treated as a broad rule that all cloud-AI inputs categorically fail privilege. That would overstate the holding.

What Heppner does provide is still important. It is a crystallization date and a judicial warning signal. A federal judge in the Southern District of New York was willing, on specific facts, to deny privilege and work product protection where AI was in the chain. After Heppner, a reasonably competent practitioner had published judicial notice that AI-generated materials carry privilege risk and that the duty to investigate, disclose, and update consent had been re-triggered. The enterprise legal-AI context addressed in this paper differs from Heppner's facts, but the lesson cuts in one direction: courts are

already willing to deny protection when AI enters the chain, which means enterprise deployments operate in a narrowing safe harbor even where their facts differ.

Chapman and Cutler's analysis adds a further point that belongs in this section. Based on the court's confidentiality reasoning, a lawyer who shares existing privileged attorney-client communications with a cloud AI platform whose privacy policy permits third-party and governmental disclosure may be waiving privilege not only as to new communications with the AI, but as to the underlying privileged communications themselves. That extrapolation does not depend on Heppner becoming a broad anti-AI rule. It follows from the confidentiality logic the court applied, and it applies with full force to enterprise legal-AI deployment because the risk is not limited to fresh AI outputs. It extends to the destruction of privilege that already attached before the material ever entered the system.

Warner v. Gilbarco has to be addressed because any sophisticated counterparty will raise it immediately. Warner is the principal counter-citation in the field, but it does not defeat this paper's analysis. Public reporting describes Warner as involving a pro se litigant, consumer ChatGPT, and a work-product dispute in the Sixth Circuit framework, not a full attorney-client privilege analysis for attorney-directed enterprise legal AI. Warner is therefore distinguishable on multiple grounds: pro se rather than attorney-directed; consumer tool rather than enterprise legal platform; work product only rather than the broader privilege, confidentiality, and custody architecture at issue here; Sixth Circuit work-product reasoning rather than a broadly controlling privilege rule; and a materially different deployment posture from agentic cloud systems with sensitive client file access.

The Harvey response point sharpens the standard-of-care question. On February 23, 2026, thirteen days after Heppner's bench ruling, Harvey announced a strategic partnership with Intapp to bring ethical wall enforcement directly into the platform, citing the need to preserve attorney-client privilege and prevent commingling of confidential information. The platform recognized the governance gap and responded with infrastructure-level changes faster than many attorneys serving their clients did. That is not a neutral observation. It is foreseeability evidence and a documented marker of whose standard of care was effectively higher.

XIV. Learned Hand, Safer Alternatives, and Why Custom Fails

This brings the analysis to the final documented leg: safer alternatives. The reason the Learned Hand logic associated with The T.J. Hooper matters is not because it offers a slogan. It matters because it states a durable rule. Industry custom does not define reasonable care when safer feasible precautions exist. A whole profession can lag behind available protective devices, and courts can still say more was required.

Applied here, the point is direct. Once local processing becomes a realistic option for crown-jewel legal information, the lawyer routing that information through cloud AI is not merely inheriting unavoidable modern risk. He is selecting the more opaque architecture over the more controlled one.

XV. Why Boilerplate Does Not Address the Architecture Problem

Lawyers often assume contractual language can absorb architectural weakness. The documented evidence raises serious questions about that assumption. A murky clause in an engagement letter does not convert opaque, evolving, black-box processing into a transparent and defensible professional workflow. The client thinks: my lawyer is storing my secrets securely. The reality documented in this paper may be: my lawyer is routing my secrets through an active computational system outside his direct control, whose dangerous failure modes are still being discovered, whose insurance treatment is worsening, whose custody chain is incomplete, whose deletion promises can be overridden, and whose billing efficiencies may not be passed through.

XVI. Conclusion

For highly sensitive client data, cloud-based AI presents a documented structural professional-liability problem. It is a problem because it collapses the distinction between storage and processing. It is a problem because documented questions about informed consent arise when the risk surface is still being mapped, and because the duty to disclose has been identified as continuous, not static. It is a problem because confidentiality and privilege face documented challenges inside opaque third-party processing systems. It is a problem because prompt injection and exfiltration are real

and documented attack paths. It is a problem because the core harm is irreversible disclosure. It is a problem because emergent severe behaviors are now empirically documented by the builders themselves. It is a problem because the absence of a public catastrophe is not proof of prudence. It is a problem because insurers are acting as if the category crossed a structural threshold. It is a problem because fee forfeiture, discipline, malpractice exposure, billing disputes, federal compulsion, preservation orders, and custody-chain opacity all stack on top of each other. And it is a problem because safer local alternatives exist.

The modern legal profession should not be asking clients to accept black-box processing of crown-jewel secrets on the theory that the market normalized it, the engagement letter was broad, the insurance issue was undisclosed, the vendor contract looked polished, and no sufficiently public disaster had yet forced a reckoning. That is not caution. That is not fiduciary judgment. That is not a defensible architecture for life-altering data. [1]-[19]

Attorneys and institutions reviewing this paper are encouraged to consult qualified ethics counsel admitted in their jurisdiction before drawing conclusions about any specific matter, practice, or technology deployment.

References

[1] ABA Formal Opinion 512, Generative Artificial Intelligence Tools (July 29, 2024).

[2] ABA Model Rules of Professional Conduct, including Rules 1.0, 1.1, 1.4, 1.5, 1.6, 5.1, and 5.3.

[3] U.S. v. Heppner, No. 25-cr-503 (JSR), S.D.N.Y., discussed here as a fact-specific warning signal and crystallization date rather than a broad anti-AI privilege rule. Public reporting describes the case as involving a criminal defendant using the consumer version of Claude on his own rather than attorney-directed enterprise legal AI, with privilege and work product rejected on those specific facts.

[4] Verisk ISO Form CG 40 47 01 26, effective Jan. 1, 2026.

[5] W.R. Berkley Form PC 51380, absolute AI exclusion language affecting D&O, E&O, and Fiduciary Liability.

[6] 18 U.S.C. § 2713, the CLOUD Act compulsion provision.

[7] New York Times v. OpenAI preservation-order litigation.

[8] The T.J. Hooper, 60 F.2d 737 (2d Cir. 1932).

[9] NIST guidance on generative AI and adversarial machine learning, including prompt-injection and data-exfiltration risk.

[10] OWASP material on LLM prompt injection and exfiltration pathways.

[11] Anthropic, Agentic Misalignment research (June 20, 2025).

[12] Damien Charlotin hallucination database tracking judicial decisions involving AI hallucinations and lawyer use.

[13] Nanda Min Htin, Harvard Journal of Law and Technology (February 9, 2026), addressing the liability-sponge dynamic in AI-mediated professional work, risk transfer, automation bias, vigilance decrement, and supervision failure.

[14] Burrow v. Arce, 997 S.W.2d 229 (Tex. 1999).

[15] Billing-integrity and fee-reasonableness analysis under Rule 1.5 as applied to AI-generated efficiency gains.

[16] Enterprise legal-AI platform documentation, including Harvey security and contract architecture issues, query-versus-content treatment, and subprocessor disclosure.

[17] Warner v. Gilbarco, Inc., E.D. Mich., discussed here as distinguishable authority protecting work product on materially different facts.

[18] Chapman and Cutler LLP, Federal Court Rules That AI-Generated Documents Are Not Protected by Privilege (Feb. 16, 2026), cited here for the analysis that sharing existing privileged attorney-client communications with a cloud AI platform whose privacy policy permits third-party and governmental disclosure may waive privilege not only for AI communications, but also for the underlying privileged communications themselves.

[19] Intapp and Harvey, Strategic Partnership Announcement: Ethical Wall Enforcement in Harvey's Platform (Feb. 23, 2026, BusinessWire). Announced thirteen days after Heppner's bench ruling. Cited here as public evidence that Harvey recognized the governance gap and responded with infrastructure-level changes to ethical wall enforcement, privilege preservation, and information governance within its platform.

Disclosures and Disclaimers

Not Legal Advice | No Attorney-Client Relationship | No Unauthorized Practice of Law

Nothing in this paper, or in any materials published at closevector.ai, constitutes legal advice, legal services, or the practice of law in any jurisdiction. Dean Hoffman and CloseVector are not attorneys, do not hold themselves out as attorneys, and are not authorized to practice law. No attorney-client relationship is formed by reading, receiving, or relying on any CloseVector communication or publication. All references to ABA Model Rules, ethics opinions, court decisions, and legal doctrines are provided solely for informational and educational purposes. The applicability of professional responsibility rules to any specific attorney's practice requires individualized analysis by licensed legal counsel admitted in the relevant jurisdiction. Attorneys and institutions should not modify their practices in reliance on CloseVector materials without first consulting qualified ethics counsel.

No Bar Association Affiliation or Endorsement

CloseVector is not affiliated with, endorsed by, sponsored by, or approved by the American Bar Association, any state or local bar association, any judicial conduct body, or any ethics oversight authority. Citations to ABA Model Rules and Formal Opinions reflect publicly available documents only and do not represent official guidance applicable to any attorney's specific circumstances or jurisdiction.

Commercial Interest Disclosure

Dean Hoffman and CloseVector have a direct commercial interest in the subject matter of this paper. CloseVector's flagship product, the CloseVector Machine, is an air-gapped legal AI workstation deployed as an alternative to cloud-based legal AI platforms. The analysis presented reflects CloseVector's own views and analytical framework and should be evaluated accordingly. Readers are encouraged to seek independent sources and qualified counsel before reaching conclusions.

Product and Technology Claims

Descriptions of CloseVector's products and services reflect CloseVector's own characterization of its architecture and methodology. No representation is made that any CloseVector product satisfies any specific legal, regulatory, or professional responsibility requirement applicable to any particular attorney or firm. Whether a technology deployment satisfies an attorney's obligations under applicable ethics rules is a determination that must be made by qualified legal counsel.

About CloseVector

CloseVector builds air-gapped legal AI infrastructure for law firms. Its flagship product, the CloseVector Machine, provides hardware-enforced document analysis, e-discovery processing, and Quantitative

Legal Communications Analytics inside the firm's physical perimeter. A CloseVector engineer deploys on-site, installs and configures the system, and remains until the firm's first case runs clean. CloseVector does not retain, host, or process client data on its own systems. CloseVector does not provide legal services, legal advice, or attorney-client representation of any kind.

Contact: Dean Hoffman, Founder | CloseVector@proton.me | closevector.ai

