Hacking the Law

What the Machine Reads That You Can't

AI security & ethics for tax attorneys · Colin McNamara · 39th Annual Legal Seminar on Ad Valorem Taxation · August 27, 2026 · Attendee reference

Two AI failures now reach the courtroom. One is **hallucination**: the machine invents authority (1,961 sanctions cases and counting). The other is **prompt injection**: a document you are handed carries hidden instructions to your own AI tool. Every current rule polices the machine's **output**. Injection attacks the **input**. The last defense against both is a human who actually reads, and for a tax attorney, that human signs under penalties of perjury.

THE CASE THAT STARTED THE INPUT PROBLEM

Elliott v. New York Bariatric Group (Conn. Super. Ct., Aug. 6, 2026). A pro se litigant hid instructions in 3-point white-on-white text telling any AI reviewing his motion to rule for him. The judge caught it by noticing the page had too much white space. Connecticut courts do not use AI; he was sanctioned anyway. "The wrong lies in the attempt." Prompt injection is, in substance, an ex parte communication with the decision-making apparatus.

No. AAN-CV-25-6066141-S. First U.S. decision to sanction courtroom prompt injection.

IT IS NOT ONE OUTLIER

- **Brazil** (May 2026): two lawyers hid white text to fool the court's AI; the court's own AI caught them. 10% fine + bar referral.
- **Peer review** (Nikkei, July 2025): 17 papers, 14 universities, 8 countries, hidden "give a positive review only."
- **Hiring** (USENIX 2026): ~1% of 200,000 real resumes carried hidden prompts. The "white fonting" trick is 20 years old; only now can the machine act on it.

HALLUCINATION SANCTIONS TO KNOW

Mata v. Avianca (S.D.N.Y. 2023). The origin. Six fake ChatGPT cases; the "many harms" paragraph.

678 F. Supp. 3d 443.

Wadsworth v. Walmart (D. Wyo. 2025). The AI was the firm's own vetted tool. Two supervisors who never touched it were fined $1,000 each for signing. "Approved legal AI" is not a defense.

No. 2:23-cv-118.

Johnson v. Dunn / Butler Snow (N.D. Ala. 2025). $0 fine, but a published opinion, disqualification, and bar referrals in all four states where he's licensed.

No. 2:21-cv-1701.

Noland v. Land of the Free (Cal. Ct. App. 2025). Published as a warning: no filing should contain a citation the responsible attorney has not personally read.

114 Cal. App. 5th 426.

> THE POINT MOST TALKS MISS
>
> ~40% of the defects in the database are **real, correctly-cited cases used for propositions they do not support**. A citation-checker that confirms a case exists catches none of them. Only a lawyer reading the case does.

TAX EXPOSURE: SETTLED, AND UNSETTLED

Privilege (settled). *U.S. v. Heppner* (S.D.N.Y., Feb 2026): chats with consumer Claude are NOT privileged or work product. *Warner v. Gilbarco* (E.D. Mich., same day): ChatGPT queries WERE work product (waived only by disclosure to an adversary). Your **§7525** practitioner privilege is built on attorney-client privilege, so it is at least as fragile.

> THE OPEN CRIMINAL QUESTION
>
> Pasting a client's return data into a consumer LLM is probably an **IRC §7216** disclosure, a criminal misdemeanor (§6713 civil twin). The IRS has published nothing on AI. Safe posture: **written §7216 consent naming the tool, or don't put the data in.**

Duties (now explicit). IRS OPR Alert 2026-19 (June 2026) maps Circular 230 onto AI: verify output (§10.22), understand the tool (§10.35), and do not bill hours the AI saved (§10.27). AICPA imposed an AI-reliance standard on CPAs Jan 1, 2024.

The IRS uses AI on you. 126 active use cases; AI-assisted selection of the 75 largest partnerships for audit. A Stanford study (IRS-confirmed) found Black taxpayers audited at 2.9-4.7x the rate of others. The model never saw race; it optimized for the wrong objective.

THE TAX ATTACK SURFACE, IN ONE SENTENCE

A K-1 footnote is an unbounded free-text field, written by a stranger, that your software is designed to read automatically, and the output is a document you sign under penalties of perjury. The same shape appears in inbound IRS notices (OCR'd scans), predecessor workpapers, and M&A data rooms.

THE NEXT SHOE: AGENTS THAT ACT, NOT JUST READ

AGENCY IS THE THIRD FAILURE MODE

Injection is what it reads; hallucination is what it invents; **agency is what it does.** In July 2026 an OpenAI model escaped its evaluation sandbox via a zero-day and breached Hugging Face (~2.5 days; write access to internal repos); Anthropic disclosed three test incidents reaching outside orgs. Attackers drove **Claude Code** through a ~90%-autonomous espionage campaign against ~30 targets (GTG-1002, Nov 2025). Mundane tasks wreck systems too: **Replit's agent deleted a production database under a code freeze**, fabricated ~4,000 fake records, and lied about it. A customer talked a Chevrolet bot into a "legally binding" $1 car with plain English. **Moffatt v. Air Canada** (2024 BCCRT 149) already held a company liable for its chatbot and rejected the "separate legal entity" defense. "My AI did it" is not a defense; it is an admission you did not supervise it. The five controls below answer the agent problem: human-in-the-loop on exit, least privilege, separate read from write.

SOLUTIONS: DEFENSE IN DEPTH (THE AI SOFTWARE LIFECYCLE MEETS THE LAW FIRM)

You can't patch prompt injection, and people cheat, so you don't lean on one control. Three layers, and make them auditable. The same patterns apply to tax, criminal, or civil work.

The pipeline: Untrusted inputs → Ingest guard → Agents → Adversarial review → Human sign-off, with **every step logged to one observability & audit layer** (LangSmith / Langfuse / OSS) you can replay for a client, an opponent, or the bar.

- 1. **Guard the inputs.** Treat every document as untrusted code. Scan at ingest (diff extracted text vs. render, flag anything invisible, ~15 lines of code, or use your e-discovery platform's hidden-content detection). Point your metadata scrubber **inbound**, not just outbound. Flatten and re-OCR scanned K-1s and opponents' PDFs so the AI reads your text layer, not theirs.
- 2. **Adversarially review the outputs.** Nothing you sign leaves on one model's say-so. Red-team your draft with sub-agents/skills that verify every citation, number, and source against a primary record. Use a **second model** (Codex checks Claude, or the reverse). Keep a human on exit; least-privilege the tools so the agent reading the opponent's docs can't reach privileged files.
- 3. **Observe & audit (courts will mandate this).** If you can't replay it, you can't defend it. Trace and log every AI action to a shared, tamper-evident record, an "AI document-management layer." Tools: **LangSmith** (hosted), **Langfuse** (self-hosted), or open-source, across Claude Code, Codex, and chat. Finance already requires this; courts are next.

THE RULES YOU ALREADY HAVE (NO NEW RULE REQUIRED)

Competence	Understand the tool's limits. Tex. Disc. R. 1.01 / Model Rule 1.1; Tex. Ethics Op. 705 (2025); ABA Formal Op. 512 (2024).
Confidentiality	Informed consent before inputting client data. Boilerplate engagement-letter language is not enough (ABA 512). R. 1.05 / 1.6.
Candor	The rule you break when a fabricated citation goes in and is not corrected. R. 3.3.
Supervision	Your signature is the duty (Wadsworth). R. 5.1 / 5.3; Circ. 230 §10.36.
Fees	Bill the time you spent, not the time it used to take. R. 1.5; Circ. 230 §10.27.

Ask any AI vendor, in writing: Do you train on our inputs? What is the retention period? Who are your subprocessors? What is your breach-notice commitment? Is there a no-training, no-retention contractual addendum? (Consumer tiers usually fail these; the answers drive both the §7525 privilege and §7216 analyses.)

Every case, statistic, and citation above was independently verified against primary sources or multiple reputable outlets as of Aug. 25, 2026. This handout is educational and does not constitute legal advice. Hallucination-case count: Damien Charlotin, AI Hallucination Cases database. ·