AI Prompt-Injection Hacking Creates Emerging Legal Risks

By **Joel Bush and Charles Gray** (September 2, 2026)

Joel Bush

Charles Gray

On Aug. 6, in Elliott v. New York Bariatric Group in the Connecticut Superior Court, Judge Walter M. Spader Jr. issued what appears to be the first U.S. decision sanctioning a prompt-injection attack aimed at a U.S. court.

Pro se plaintiff Matthew Elliott filed motions containing white-on-white text instructing any AI reviewing his filings to agree with his arguments and grant relief the clerk had denied.

Judge Spader noticed unusual white space while reviewing the docket on paper and found the hidden instructions.

The court found the conduct irreconcilable with the good faith certification required by Connecticut Practice Book Sections 4-2(b) and 4-9, invoking its inherent authority and duty of candor. It likened the concealed instruction to an ex parte communication, delivered through a channel the opposing party can neither see nor answer.

It rescinded Elliott's electronic filing privileges; future filings must be made in person on paper. Unlike Connecticut's 2026 AI rules, which focus on verifying AI output such as hallucinated citations, Elliott's conduct manipulated the input.

With no square U.S. precedent, Judge Spader rested on Connecticut law and inherent authority, citing only as illustration Elisandro Martins de Barros v. Renato Ribeiro de Lima, a Brazilian case decided on May 12, where hidden instructions targeted a court's AI system and the tribunal fined two lawyers 10% of the claim and referred them for discipline.

He distinguished Tov Realty LLC v. Suarez, decided by the Connecticut Supreme Court on June 9, which sanctioned an attorney over AI-fabricated citations, on intent, finding Elliott's conduct deliberate rather than negligent. The practical warning is immediate: For in-house counsel, the exposure is on the reading side. A hidden command in an incoming filing, production or exhibit can silently skew an AI summary.

Spader also names a failure mode the profession has not reckoned with: These tools tend toward agreeableness, so a litigant who prompts only for support gets it, fluently. When the

argument fails, the author concludes the loss was illegitimate rather than wrong. He urges the bar to ask a tool to test a position as readily as to advance it.

Why Prompt Injection Matters

Prompt injection is adversarial text designed to make an AI system depart from its intended task. In a direct injection, the attacker types the instruction into the tool. In an indirect injection, the instruction is hidden in content — such as a filing, email, resume, web page or PDF — that another person later asks the tool to analyze.

The attack works because an LLM receives operator instructions and external text in one context stream; there is no reliable architectural wall that makes the latter data only. White-on-white or tiny text may be invisible to people, but remains visible to text extraction or optical character recognition, so the model can treat it as an instruction.

The result can be misleading summaries, suppressed issues, fabricated conclusions, disclosure of system prompts or confidential data, or unauthorized actions by an AI agent. The key distinction is not whether the output looks polished; a compromised summary may look accurate.

Trade Secret Risks

Under the Defend Trade Secrets Act, AI system prompts, model configurations and proprietary instructions may be trade secrets if they derive value from secrecy and are protected by reasonable measures.

No court has definitively resolved whether prompt injection is a form of reverse engineering, but it likely would be considered an improper means: It uses deception to make a system disclose information it is designed to withhold.

E.I. duPont deNemours & Co. v. Christopher, decided by the U.S. Court of Appeals for the Fifth Circuit in 1970, frames improper means by commercial morality; Compulife Software Inc. v. Newman, decided by the U.S. Court of Appeals for the Eleventh Circuit in 2020, recognizes that even a public interface can be exploited; and Alcatel USA Inc. v. DGI Technologies Inc., decided by the Fifth Circuit in 1999, treats deception used to gain access to protected software as improper.

In OpenEvidence Inc. v. Doximity Inc., filed in the U.S. District Court for the District of

Massachussetts on June 20, 2025, the plaintiff alleged prompt-injection-based extraction of AI trade secrets, offering an early example of how these claims may be litigated. Lawful reverse engineering starts with a lawfully obtained product and independent effort; prompt injection instead induces a direct disclosure through deception.

Employment and Cybersecurity Risks

Applicants can hide instructions or fabricated skills in resumes and applications to influence AI screening, advancing unqualified candidates or distorting evaluations. The same indirect-injection risk becomes more serious when an AI agent has access to sensitive data, receives untrusted content, and has an outbound channel — what security researchers call the lethal trifecta.

A successful injection can then exfiltrate data or take unauthorized action without a human noticing. The EchoLeak incident involving Microsoft 365 Copilot last year illustrates the concern: A malicious email reportedly caused silent exfiltration from connected enterprise services. Human review and least-privilege controls remain necessary even where model defenses improve.

Litigation and Professional Responsibility Risks

The risk runs in two directions. A lawyer or party who hides instructions in a filing, discovery response or correspondence to manipulate a court's or counterparty's AI review can face sanctions under the court's inherent authority — Federal Rule of Civil Procedure 11, and professional duties of candor and fairness (Model Rules 3.3 and 3.4), as well as Model Rule 8.4(d)'s prohibition on conduct prejudicial to the administration of justice.

Elliott shows that a court may impose a targeted filing restriction without dismissing the case; Barros shows a tribunal reaching for a monetary penalty and a disciplinary referral.

The opposite risk is relying on AI that has been manipulated by adverse filings, third-party documents or incoming correspondence. Judge Spader made the same point to the bar, warning that the risk runs not only to what counsel files but to what counsel feeds its own tools from the other side, naming an opponent's production, a witness statement and an expert report.

Failing to screen inputs or verify consequential outputs may implicate Model Rule 1.1's competence requirement and Rules 5.1 and 5.3's supervision duties. Nonlawyer actors —

competitors, applicants, customers or pro se litigants — are not bound by lawyer-specific rules, but their conduct may still support trade secret, fraud, unfair competition or computer access claims.

Patent Prosecution: The Same Attack, Higher Stakes

The exposure is not confined to litigation, and prosecution is where the incentives are sharpest. Since January 17, 2024, the specification, claims and abstract of a U.S. utility application must be submitted in DOCX format, or the applicant must pay a non-DOCX surcharge under Title 37 of the Code of Federal Regulations, Section 1.16(u), currently $430 for a large entity.

Examiners then search that text with AI: The U.S. Patent and Trademark Office's similarity search tool, in use by utility examiners since 2022, is trained on disclosure text and classification data.

Judge Spader's remedy does not translate to the USPTO. A court can revoke e-filing and require paper; the USPTO cannot practically do the same, because paper filing carries its own $400 fee under Section 1.16(t), avoidable only through the Patent Center.

The deterrent must come from consequences instead, and in prosecution they are harsher. Every paper presented to the USPTO carries a certification under Title 37 of the Code of Federal Regulations, Section 11.18(b), that the party's statements are true to its own knowledge and that it made an inquiry reasonable under the circumstances, and Section 11.18(c) authorizes striking the paper, terminating the proceeding and disciplinary referral.

The duty of candor and good faith under Section 1.56 runs alongside it, and the office has confirmed that these existing duties are not displaced when AI is used in practice before it.[1]

Concealed text intended to steer an examiner's tools is difficult to characterize as anything but an attempt to deceive. Proven by clear and convincing evidence of but-for materiality and specific intent, the consequence is not a sanction the client absorbs but an unenforceable patent.[2]

The inbound vector deserves more attention. An applicant does not control everything that enters its own file. A third party may place a document there directly, through a preissuance submission of prior art under Title 35 of the U.S. Code, Section 122(e), and

Title 37 of the Code of Federal Regulations, Section 1.290, and information disclosure statements routinely carry third-party references and translations.

A competitor with an interest in narrow claims has both a motive and, after Elliott, a demonstrated method. Prosecution counsel should read the text layer of every inbound reference before it reaches an analysis tool, and do the same for its own applications before filing, where concealed text is not merely a litigation risk but a defect in the patent.

Best Practices: Practical Guidance for In-House Counsel and Business Clients

Because no model-layer technique reliably stops every injection, defense should assume some will succeed and limit what the AI can access, disclose or do. The following controls address the so-called lethal trifecta and the reading-side risk.

- ***Screen documents for hidden text.*** Compare rendered pages with extracted text; use select-all, plain-text paste, or PDF flattening to surface invisible content before sending documents to an AI tool. Treat flags as a review signal because OCR and accessibility layers can be legitimate.
- ***Keep a human in the loop for consequential decisions.*** Require human review of source materials before relying on AI for hiring, legal analysis, contract review, compliance or other outcomes affecting rights or obligations.
- ***Vet vendors' AI guardrails.*** Ask how the tool separates instructions from supplied content, sanitizes inputs, filters outputs, tests for indirect injection, reports vulnerabilities and uses customer data.
- ***Apply least privilege to AI agents.*** Limit data, tools, credentials and outbound channels to what the task requires; require confirmation for sensitive actions; log activity and rotate tokens. Removing one leg of the lethal trifecta can prevent catastrophic exfiltration.
- ***Govern shadow AI with written policies.*** Maintain an inventory of sanctioned and unsanctioned tools, prohibit sensitive inputs absent approval, set retention and training rules, and train personnel on prompt-injection and confidentiality risks.
- ***Plan for incidents.*** Maintain and test an AI-specific response plan covering suspected injection, compromised credentials, preservation of prompts and logs, isolation of the tool, escalation, and required legal, contractual, regulatory and insurance notices.

Key Takeaways

Connecticut has now shown that prompt injection aimed at a U.S. court can produce sanctions. Such conduct manipulates input, not merely AI output.

Indirect injection creates a reading-side risk. Hidden content can silently bias summaries and decisions made by counsel, employers and businesses.

Trade secret exposure is real. Deceptive extraction of protected AI instructions likely qualifies as improper means, not lawful reverse engineering.

AI agents magnify cybersecurity and employment risk. Resume manipulation and the lethal trifecta make human review and least privilege essential.

Patent practice raises the stakes. USPTO filings are machine-readable by rule and searched with AI tools, and concealed instructions there risk Section 11.18 sanctions and an unenforceable patent.

Prompt-injection law and practice are developing quickly. Organizations should treat AI inputs as untrusted, preserve human judgment and document proportionate safeguards as courts and regulators continue to define the standard of care.

Joel D. Bush and Charles W. Gray are partners at Kilpatrick Townsend & Stockton LLP.

[1] 89 Fed. Reg. 25609 (Apr. 11, 2024).

[2] Therasense, Inc. v. Becton, Dickinson & Co., 649 F.3d 1276 (Fed. Cir. 2011) (en banc).