For The Defense

dri
The magazine for defense, insurance and corporate counsel

January 2026

Artificial Intelligence

Including . . .

Defending the Algorithm™:
A Bayesian Analysis of AI Litigation and Law

Also in This Issue . . .

Rise of the AI Pro Per:
A New Challenge for the Civil Defense Bar

And More!

THE CENTER
The Voice of the Civil Defense Bar.

Defending the Algorithm™

By Henry M. Sneath, Acacia B. Perko, and Christopher M. Jacobs

This article provides a framework for understanding AI as a probability engine and for defending AI-related litigation.

A Bayesian Analysis of AI Litigation and Law

A Publication of the DRI Center for Law and Public Policy AI Task Force - Part 1

Please attend our live CLE full-day seminar on March 10, 2026 at DRI Headquarters in Chicago entitled:

Artificial Intelligence in Defense Practice: Tools, Prompts, Workflows, Implementation and Governance

While AI has the potential to transform the legal industry and business operations across all sectors, managing its risks through effective legal frameworks and strategic defense methodologies is imperative. This two-part article applies Bayesian (AI) probability theory to predict emerging AI litigation trends and provides practical guidance for defending AI-related claims. It was authored with assistance from Claude® Opus 4.5 Max and research confirmation by Google Gemini 3.0 Pro and Westlaw Precision with AI and Analytics, but edited by humans as AI can make mistakes. Thomas Bayes, mathematician, statistician, philosopher and minister, lived from 1702 – 1761 and generated the mathematical equation that later became known as Bayes Theorem of conditional probability, which is one of the two main schools of thought on probability theory (the other is "frequentist").

The term Defending the Algorithm™ is used and owned by Houston Harbaugh, P.C., and a registered trademark protection is being sought.

Henry M. Sneath is a Director/Shareholder in the firm of Houston Harbaugh, P.C. in Pittsburgh, Pennsylvania. Mr. Sneath is a Business, IP and AI Trial Lawyer and a former President of DRI. He focuses his practice on intellectual property litigation, trade secret misappropriation, AI law, and complex commercial, employment and insurance coverage and bad faith disputes. Mr. Sneath is currently serving on the DRI Center for Law and Public Policy AI Task Force and is working to develop comprehensive AI education programs to be presented by the DRI Center for Law and Public Policy. **Acacia B. Perko** is a Senior Attorney at Houston Harbaugh, P.C. and practices Business, IP, Trade Secret and Employment litigation, and serves as DRI's Trade Secrets Substantive Law Group Chair for the Intellectual Property Litigation Committee. **Christopher M. Jacobs** is a Director/Shareholder at Houston Harbaugh, P.C. and practices Insurance, Business and IP litigation and serves on the DRI AI Task Force. Together, these authors record and publish a blog and podcast series entitled: Defending the Algorithm™: A Bayesian Analysis of AI Litigation and Law and they are frequent speakers on this topic and related IP issues.

Introduction

The legal profession stands at an inflection point. Lawyers and law firms must make AI choices. Artificial intelligence is not merely changing how lawyers work—it is creating entirely new practice areas that intersect with virtually every substantive field of law. From copyright infringement to trade secret misappropriation, contract and computer-access disputes to insurance and bad faith considerations involving AI-driven decision-making, AI-related litigation is proliferating and is coming to a courthouse near you. In this article we use AI probability tools and concepts to augment our human analysis of the intersection of AI and the Law. One thing we can tell you for sure is that this is NOT Y2K.

For defense attorneys and their clients, this emerging landscape presents both significant challenges and opportunities. The challenges are obvious: rapidly evolving technology, uncertain legal standards, massive discovery productions and disputes and high-stakes lawsuits. The opportunities, however, are equally compelling: the chance to shape foundational precedent, develop sophisticated defensive frameworks, and provide clients with strategic and economic advantages as an augmentation of human intelligence in an area where many practitioners are still finding their bearings. And – more and more clients are requesting information on their law firm's AI protocols, training and guardrails. Are you using Narrow Task-Specific AI? Machine Learning? Deep Learning? Generative AI? Is your AI Agentic? In order to best use AI in your defense practice, you need to understand AI and its role as a probability engine designed to mimic human thought, and you and your firm need to understand how to retain human control over the output model.

This article provides a framework for understanding AI as a probability engine and for defending AI-related litigation. We will examine the first wave of cases now shaping the field. Drawing on matters such as *Bartz v. Anthropic, Reddit v. Anthropic and OpenEvidence v. Pathway Medical*, it analyzes how courts are beginning to address novel questions involving AI training data acquisition, Terms-of-Service restrictions, computer-access doctrines, trade secret protection, and the challenges posed by opaque model "black box" architectures. Critically, our analysis uses Bayesian conditional probability theory to calculate or update the likelihood of a hypothesis (Event "A") when given ("|") new evidence of a prior or existing event "B" to calculate a posterior probability of the "A" event occurring. P(A|B) or the probability of "A" given "B"—which is the foundational mathematical structure underlying AI itself—can help us anticipate how these early decisions in AI-related litigation are likely to influence future litigation trends and risk assessments. AI predicts the next most probable word in a textual or verbal response to a prompt or query by the human user of the AI platform. Defense Trial lawyers predict outcomes, litigation cost and other probabilities to clients and opposing counsels all the time; and therefore, whether we realize it or not, we are using Bayesian probability reasoning all day long in the trial lawyer world. AI is trying to mimic our human thought and we need to learn how to prompt it and use it effectively and ethically. We need to understand AI, to allow us better use of AI, to defend AI litigation.

Part I: The Bayesian Framework for Predicting AI Litigation

A. Understanding Bayesian Analysis in Legal Context

Bayes' Theorem is a fundamental building block in the probability-predicting technology of AI and the race to create artificial general intelligence (AGI) and superintelligence. *See* RICHARD E. NEAPOLITAN, *LEARNING BAYESIAN NETWORKS* (2003) (explaining Bayesian inference as a method for updating probability estimates based on new data). It is equally useful in developing an understanding of how we can predict new developments in the law of AI and it is a key method of reasoning that trial lawyers use every day in generating predictions about litigation results. The theorem allows us to update probability estimates as new evidence emerges, making it an ideal framework for analyzing this rapidly evolving legal landscape. Let's use Bayesian AI techniques to evaluate AI litigation and Law.

In practice, Bayesian reasoning helps quantify litigation risk for companies deploying AI systems, particularly as each new decision—favorable or unfavorable—modifies the expected contours of liability. It is especially effective for evaluating the relationship between a client business's "AI-related" operations and the likelihood it will face suit. Bayes Theorem is based on the algebraic equation: P(A|B) = [P(P(B|A) x P(A)] ÷ P(B). To illustrate the framework as applied to AI litigation risk for businesses, consider the core Bayesian equation framed as a predictor of AI-related litigation given a company's deployment and use of AI-related business operations:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

Each component plays a distinct role in assessing litigation exposure:

- **P(AI Litigation | Business Operations) — The Posterior Probability.** This is what we are solving for: the probability that a company will face AI litigation given its own specific business and AI enterprise operations. This is the "updated" risk assessment that we seek after considering new evidence.

- **P(Business Operations | AI Litigation) — The Likelihood.** This asks: Among companies that have faced AI litigation, what percentage had AI business operations similar to yours?

  For example, if (hypothetically) 80% of companies sued for AI copyright infringement were using LLMs for training or content generation as part of their business operations, then this probability equals 0.80 and it would factor into the overall equation.

- **P(AI Litigation) — The Prior Probability.** This is the baseline probability of AI litigation across all businesses, regardless of specific operations. Before the Bartz case (on copyright issues), this might have been estimated at a low probability, but after the Bartz case (see below), it has likely increased to a much higher

probability. The probability in the insurance world is also increasing rapidly but that is what we are testing.

• **P(Business Operations) — The Evidence.**

This is the probability that any randomly selected business engages in your client's specific AI-related operations. While we are not mathematicians nor statisticians, and we cannot right now plug real numbers into this equation, a probability analysis helps us to counsel clients on AI integration risk and provides a framework for strategic litigation planning. While we cannot yet accurately predict the probability of any specific company facing AI litigation without sufficient data, each new case provides additional evidence. As the body of AI case law expands, these updated data points will refine our risk assessments. Again, we are trial lawyers using probability reasoning to understand the probability engine of AI and resultant litigation.

B. Applying Bayesian Analysis to Judicial Precedent

Bayesian analysis becomes even more useful when applied to the growing body of judicial precedent addressing AI training and deployment. Each new ruling represents an incremental data point—new "evidence" in Bayesian terms—that should update our assessment of litigation risk and doctrinal direction.

The Bayesian approach guides us to: (1) treat Judge Alsup's Northern District of California decision in *Bartz v. Anthropic* as our now prior probability or baseline assumption; (2) combine that **prior probability** with new evidence emerging from other cases and forums to create a **likelihood function**; and (3) apply Bayes' Theorem to compute a **posterior probability**; predicting how likely the Alsup framework is to become widely adopted as precedent across multiple federal circuits. *See Bartz v. Anthropic*, No. 24-cv-05417 (N.D. Cal. filed Oct. 2024). Subsequent cases serve as **likelihood evidence**, shifting expectations as the jurisprudence develops. This analytical progression mirrors precisely how courts, defense trial lawyers and litigants adjust positions as new case law develops.

Although not binding outside the Northern District of California, Judge Alsup's opinion is already shaping how litigants frame arguments and how courts may approach the intersection of AI training, copyright doctrine, and fair use. As new opinions arrive, each one functions as a **likelihood input** that modifies expectations regarding the trajectory of AI-related claims.

For example:

• If multiple courts adopt Alsup's transformative-use reasoning, the posterior

probability increases that this framework will become widely accepted.

- If courts narrow or distinguish Bartz, particularly concerning intermediate copying or market harm, the posterior probability adjusts downward.
- If courts place greater emphasis on unlawful acquisition (piracy, scraping), litigation risk shifts toward acquisition-based liability instead of fair use analysis.

Bayesian reasoning therefore provides defense counsel with a structured (albeit somewhat metaphorical) way to evaluate how newly issued opinions should adjust litigation strategy. As case law grows, posterior probabilities become more refined, allowing counsel to give clients more informed assessments of likely exposure and emerging doctrinal trends.

Part II: AI and LLM Copyright Litigation — The Bartz v. Anthropic Framework

A. The $1.5 Billion Settlement and Its Implications

Nowhere is this dynamic more evident than in the earliest and most consequential set of lawsuits confronting AI developers: the copyright actions arising from LLM labs and their model training datasets. Among these, *Bartz v. Anthropic* provides the first major judicial and settlement roadmap for how courts and plaintiffs are approaching AI training practices. The case now functions not only as a doctrinal anchor point, but also as a key "prior probability" within the broader predictive framework for AI litigation. Claude and other Large Language Models (LLMs) train their models on essentially all the electronic data in the world, modified, trained and rewarded by human interaction. They download the internet and then create algorithms to evaluate that massive data tranche in response to human prompts. The algorithm recognizes patterns in the data, and predicts the next most probable word in a textual response to a prompt (or query) from a user. It uses Bayesian probability analysis. The training process for an LLM (like ChatGPT) includes the downloading of copyrighted works done by authors all around the world – books, articles, movies, music, TV shows - everything

In what will represent the largest copyright settlement in U.S. history (when approved by the court), Anthropic agreed to pay $1.5 billion to settle a class-action lawsuit brought by authors and publishers who alleged the company illegally trained its Claude AI LLM system on their copyrighted works. *See* Anthropic Settlement Agreement, *Bartz v. Anthropic*, No. 24-cv-05417 (N.D. Cal. filed Oct. 2024). The settlement, pending approval in the Northern District of California, establishes critical precedent for how AI companies can legally use copyrighted material to train their systems. The "reasonableness" hearing on the class settlement is scheduled for April 2026.

Under the settlement terms, Anthropic will pay approximately $3,000 per book to roughly 500,000 affected authors and has agreed to delete pirated works downloaded from shadow libraries. See id. at 4–6. This substantial payout sends a clear message about the financial consequences of using illegally obtained copyrighted material. Had Anthropic proceeded to trial on the piracy claims, potential damages could have reached multiple billions of dollars—potentially crippling the company given that statutory damages for willful copyright infringement can reach $150,000 per work. *See* 17 U.S.C. § 504(c).

B. Judge Alsup's Groundbreaking Fair Use Analysis

Judge Alsup's partial summary judgment ruling in *Bartz v. Anthropic* (which likely prompted the settlement agreement) provides the first substantive judicial framework addressing whether training AI models on copyrighted works constitutes fair use under 17 U.S.C. § 107. *See* Order on Cross-Motions for Summ. J., *Bartz v. Anthropic*, No. 24-cv-05417 (N.D. Cal. Mar. 2025). While the settlement resolved the piracy claims, Alsup's fair use analysis remains a foundational reference point in the developing law. It establishes a critical distinction between legally obtained training materials—potentially protected by fair use—and pirated training data, which receives no such protection in his opinion. LLM Training is being done not only by the major tech players, but by businesses, law firms and other companies who build or use AI "enterprise" software that can be plugged into an existing business model. Many law firms are building or purchasing enterprise AI for their own databases so that internal searching of data is more efficient and performed by lawyers in a closed system. Firms are marketing themselves as "AI-Powered Law Firms."

In applying the copyright four-factor fair use test, Judge Alsup held that training a large language model on lawfully acquired books is "quintessentially transformative." *See* id. at 23. He emphasized that AI systems do not reproduce or republish the works they ingest; rather, they use them to generate novel and fundamentally different outputs. Alsup likened this process to "any reader aspiring to be a writer," who studies existing works "not to race ahead and replicate or supplant them—but to turn a hard corner and create something different." *See* id.

Nowhere is this dynamic more evident than in the earliest and most consequential set of lawsuits confronting AI developers: the copyright actions arising from LLM labs and their model training datasets.

This transformative-use reasoning is particularly significant because it reframes two factors that traditionally weigh against fair use: (a) the commercial nature of the use; and (b) the wholesale copying of entire works during training. Rather than treating intermediate copying as infringing conduct, Judge Alsup focused on the AI model's ultimate purpose: creating new content, not substituting for the originals. His ruling suggests that courts may view AI training as sufficiently transformative to

override concerns about commercial intent or the scale of copying.

Judge Alsup also addressed Anthropic's digitization of legally purchased books, concluding that the company merely "replaced the print copies it had purchased for its central library with more convenient space-saving and searchable digital copies." *See* id. at 30–31. Because the digitization neither added new copies nor expanded distribution, it qualified as fair use. Judge Alsup's analysis provides a roadmap for defending AI companies when training data is lawfully obtained. But his opinion simultaneously draws a bright line that becomes critical in the next section: fair use does not shield training on pirated or unlawfully acquired materials.

C. The Critical Piracy Distinction

While Judge Alsup's transformative-use analysis provides meaningful protection for AI developers who train models on lawfully obtained works, his opinion draws a bright and consequential boundary: fair use does not extend to materials obtained unlawfully. *See* Order on Cross-Motions for Summ. J. at 22–23, *Bartz v. Anthropic*, No. 24-cv-05417 (N.D. Cal. Mar. 2025). This distinction is central to understanding both his opinion and the resulting settlement.

In addressing Anthropic's use of pirated books obtained from shadow libraries, Judge Alsup was unequivocal. He wrote that Anthropic "downloaded for free millions of copyrighted books in digital form from pirate sites as part of an effort to amass a central library of all the books in the world to retain forever." *See id*. at 2. The court emphasized that such conduct exceeded any fair use protection, regardless of whether the training process itself might otherwise be considered transformative.

Judge Alsup therefore rejected Anthropic's argument that the *source* of the works—licensed, purchased, or scraped—did not matter. Instead, he held that the legality of the acquisition is a threshold requirement. Fair use "cannot sanitize" unlawful copying at the point of acquisition, and AI developers cannot rely on the transformative-use doctrine to shield training practices that begin with pirated datasets. *See id*. at 24.

While Judge Alsup's Northern District of California ruling isn't binding nationwide, this holding has far-reaching implications. Even if courts ultimately conclude that training on copyrighted works is fair use, that analysis depends on the works being lawfully obtained. Training models on materials scraped from piracy-based repositories like Library Genesis or Z-Library therefore cannot be justified by fair use and exposes companies to potentially massive statutory damages.

The piracy distinction is becoming a defining feature of AI copyright litigation. As plaintiffs' attorneys identify whether training datasets include unlawfully sourced material, claims are increasingly being framed around acquisition rather than use. After *Bartz*, the provenance of training data is now a central component of litigation strategy for both plaintiffs and defendants.

PART III: EXPANDING AI LEGAL THEORIES BEYOND COPYRIGHT

A. The Multi-Theory Litigation Landscape

While the *Bartz* decision provides important early guidance on how courts may treat the use of copyrighted works in AI training, it represents only one corner of a rapidly expanding litigation landscape. Plaintiffs' lawyers have already recognized that copyright claims—particularly when confronted with transformative-use arguments—may not always provide the most direct or predictable path to liability. As a result, new lawsuits are increasingly grounded in alternative legal theories that avoid the fair use framework altogether. Applying our Bayesian framework: P(*AI Lawsuits based on claims other than copyright infringement) = High.*

These emerging claims focus less on the expressive content of the underlying works and more on how the data was obtained, what contractual or platform restrictions governed access, and whether companies derived economic benefit from proprietary or user-generated information without authorization. Courts are seeing lawsuits based on:

- violations of website Terms of Service and screen scraping claims;
- unauthorized access under the Computer Fraud and Abuse Act (CFAA);
- unjust enrichment;
- misappropriation of proprietary or compiled databases;
- breach of contract and tortious interference; and
- violations of state consumer-protection statutes.

See, e.g., Compl., *Reddit, Inc. v. Anthropic PBC*, No. 3:25-cv-05643 (N.D. Cal. removed Sept. 2025) (asserting breach of contract, trespass to chattels, unjust enrichment, and interference claims); Compl., *Ryanair DAC v. Booking.com B.V.*, No. 1:20-cv-01191 (D. Del. Aug. 2022) (addressing unauthorized scraping and CFAA-based theories).

This shift reflects a deliberate strategic move by plaintiffs. Even if courts ultimately conclude that AI training on copyrighted works is transformative, companies may still face substantial liability if the data was acquired through impermissible means. These theories bypass the Copyright Act entirely and instead center the litigation on issues of authorization, consent, platform governance, product safety liability and economic value.

For defense counsel, the expansion into these alternative theories presents two challenges. First, companies that believed they were protected by fair use must now contend with claims that do not depend on copyright ownership. Second, AI systems often ingest data from websites, APIs, user platforms, or repositories where the terms of access to those sites vary widely. Without careful documentation of how training data was acquired, defendants may struggle to demonstrate compliance with contractual restrictions or statutory authorization requirements.

This broader legal landscape and the cases discussed in the following sections—*Reddit v. Anthropic, Ryanair v. Booking.com*, and similar suits—illustrate how courts are beginning to grapple with these rapidly evolving theories.

B. Terms-of-Service Violations and Contract Claims

One of the most prominent non-copyright theories emerging in AI litigation involves alleged violations of website Terms of Service (TOS). These claims arise when AI developers use special technology to scrape (fully copy data and meta data)

website content or user-generated material in ways that purportedly exceed or violate site-specific access restrictions. Because TOS provisions govern the relationship between users (or automated agents) and the platform, they create an independent contractual basis for liability that is entirely separate from copyright.

The leading example is *Reddit v. Anthropic*, a lawsuit filed in the Northern District of California. *See* Compl., *Reddit, Inc. v. Anthropic PBC*, No. 3:25-cv-05643 (N.D. Cal. removed Sept. 2025). Reddit alleges that Anthropic systematically harvested vast amounts of Reddit user content despite explicit prohibitions in the platform's TOS. According to Reddit, Anthropic's scraping not only breached contractual terms but also interfered with Reddit's ability to license its data to third parties—thus supporting claims for: (a)breach of contract; (b) trespass to chattels; (c) tortious interference; and (d) unjust enrichment.

These claims are strategically significant because they do **not** require proof of copyright ownership. They hinge instead on:

1. Whether the platform's TOS clearly prohibit automated scraping;
2. Whether the defendant had access or notice of those restrictions; and
3. Whether the scraping interfered with the platform's business interests.

For AI developers and vendors, this reinforces the importance of tracking the provenance of training data and monitoring compliance with the TOS governing the websites and repositories from which data is collected.

C. The Computer Fraud and Abuse Act (CFAA)

Claims under the Computer Fraud and Abuse Act (CFAA), 18 U.S.C. § 1030, have also become a prominent part of the expanding AI litigation landscape. The CFAA provides both criminal penalties and civil remedies against violators. Plaintiffs increasingly contend that the automated scraping of data for AI training constitutes "unauthorized access" or "exceeding authorized access" under the statute. These claims arise even when the scraped content is publicly viewable, reflecting a broader trend toward using the CFAA to challenge large-scale, automated data acquisition.

A leading example is *Ryanair v. Booking.com*, a case from the District of Delaware. *See Ryanair DAC v. Booking.com B.V.*, No. 1:20-cv-01191 (D. Del. Aug. 2022). Although not an AI case per se, the court held that automated scraping by Booking.com of flight data from Ryanair's website violated the CFAA because Booking.com accessed the site in ways that exceeded the scope of their (or any user's) authorized use. The case is increasingly cited in AI-related complaints to support the proposition that automated scraping—whether or not technically "hacking"—may constitute unauthorized access when done in violation of website restrictions.

For AI developers, the CFAA introduces a separate and potentially serious source of liability. Even if the scraped material is not copyrighted, and even if the platform's Terms of Service are ambiguous, plaintiffs may argue that:

1. the developer intentionally accessed a protected computer;
2. the access exceeded authorized use (e.g., prohibiting scraping or automated bots); and
3. the scraping caused loss or harm under the statute.

The CFAA also presents risk for enterprise customers who deploy AI systems that rely on third-party datasets. If a vendor acquires training materials through unauthorized scraping, downstream enterprise users may face claims alleging derivative liability or unjust enrichment.

D. Unjust Enrichment and Quasi-Contract Claims

Unjust enrichment is a growing alternative theory when plaintiffs allege that AI developers gained economic benefit from unauthorized data. Unlike copyright claims—which require proof of ownership and substantial similarity—unjust enrichment focuses on the value conferred and whether the defendant was unjustly enriched at the plaintiff's expense.

In *Reddit v. Anthropic*, unjust enrichment is one of the central claims. *See* Compl., *Reddit, Inc. v. Anthropic PBC*, No. 3:25-cv-05643 (N.D. Cal. removed Sept. 2025). Reddit alleges that Anthropic benefited commercially from harvesting Reddit's platform content, much of which was contributed by users under terms that expressly restricted scraping and reuse. According to the complaint, Anthropic's use of this data: (a) conferred measurable commercial value on Anthropic; (b) impaired Reddit's ability to license its data to other companies; and (c) was acquired without authorization or compensation. These allegations reflect a broader trend in AI litigation: plaintiffs are reframing data not merely as expressive content, but as a commercial asset with independent economic value. By avoiding questions of transformative use, unjust enrichment claims allow plaintiffs to argue that AI developers exploited the market value of proprietary or user-generated data without permission.

From the defense perspective, unjust enrichment claims can be challenging because they do not turn on discrete legal rights like copyright. Instead, they focus on equitable considerations, market dynamics, and the implicit value exchanges involved in data scraping. This creates uncertainty, as courts vary in how they evaluate the "benefit" to the defendant and the corresponding "expense" to the plaintiff.

E. Enterprise AI and Upstream Liability

The expanding landscape of non-copyright AI litigation also raises important questions about upstream liability for companies that deploy enterprise AI tools developed by third-party vendors. Many businesses rely on commercial AI products—such as model-as-a-service platforms or enterprise-level LLM APIs—without full visibility into the provenance of the training data or the vendor's acquisition practices.

This creates meaningful exposure. Even if an enterprise customer never scraped any data itself, plaintiffs may argue that the company benefited from training datasets obtained through unauthorized or other upstream impermissible means. Complaints increasingly allege that downstream users are liable when they: (a) commercially benefit from models trained on improperly acquired data; (b) deploy systems incorporating data scraped in violation of Terms of Service or the CFAA;

or (c) reap the advantages of proprietary databases misappropriated by upstream vendors. *See, e.g.,* Compl., *Reddit, Inc. v. Anthropic PBC*, No. 3:25-cv-05643 (N.D. Cal. removed Sept. 2025) (alleging commercial benefit derived from unauthorized scraping).

Vendor indemnification provisions do not always solve the problem. Many AI providers expressly limit indemnity obligations for claims arising out of training data or data provenance, leaving enterprise customers vulnerable to lawsuits based on theories outside of copyright. This is particularly true when the claims focus on unauthorized scraping, contractual restrictions, or unjust enrichment — areas where traditional copyright defenses, such as transformative use, simply do not apply. From a risk-management perspective, companies adopting AI tools must now consider supply-chain liability in much the same way they would evaluate cybersecurity risk or data-storage risk. Counsel should ensure that procurement teams demand transparency regarding: (a) the training data sources used by vendors; (b) compliance processes for Terms of Service and authorized access; and (c) indemnity provisions tailored specifically to claims of unlawful acquisition or scraping.

PART IV: TRADE SECRET PROTECTION IN THE AI CONTEXT

A. The Defend Trade Secrets Act Framework

Trade secret law offers another significant avenue for claims against AI developers and downstream enterprise users. Unlike copyright—which focuses on the expressive nature of protected works—trade secret doctrine centers on the confidentiality and economic value of information. This distinction makes the Defend Trade Secrets Act (DTSA) and its state law counterparts under the Uniform Trade Secrets Act (UTSA) versatile tools for plaintiffs who allege that AI systems were trained or deployed using confidential or proprietary data. *See* Defend Trade Secrets Act, 18 U.S.C. § 1836 et seq.

Under both the DTSA and state Uniform Trade Secrets Acts, a trade secret includes information that: (a) derives independent economic value from not being generally known; and (b) is subject to reasonable efforts to maintain its secrecy. This definition encompasses a wide range of AI-related material, including model architectures, system prompts, fine-tuning processes, training datasets, unique output structures or formatting logic, and compilations of business intelligence or research data. Plaintiffs increasingly assert that AI developers or vendors have used or misappropriated confidential data during model training, fine-tuning, or product deployment. These allegations do not depend on copyright ownership but instead focus on (a) whether the data had economic value tied to confidentiality, and (b) whether the alleged misappropriation occurred through improper means such as unauthorized access, scraping, or breach of contractual confidentiality restrictions.

For defense counsel, the DTSA introduces several practical concerns. Because modern AI models rely on massive datasets, defendants must be prepared to trace the provenance of training sources and document the processes used to protect third-party or proprietary information. They must also be prepared to show that reasonable steps existed to prevent unauthorized disclosure or use — a requirement that can become complicated when dealing with opaque training pipelines or legacy datasets.

B. The OpenEvidence Litigation: AI System Prompts as Trade Secrets

One of the first major trade secret cases involving AI systems is *OpenEvidence v. Pathway Medical*, a lawsuit alleging that Pathway misappropriated proprietary "system prompts" and confidential architectural elements used to retrieve and evaluate medical research. *See* Compl., *OpenEvidence, Inc. v. Pathway Med., Inc.*, No. 1:25-cv-10471 (D. Mass. filed Feb. 2025). The complaint asserts that OpenEvidence spent years developing specialized prompts, internal logic structures, and workflow protocols designed to guide their platform's evidence-synthesis process. These proprietary prompts embodied OpenEvidence's core intellectual property and, according to the complaint, constituted protectable trade secrets under both the DTSA and state law.

The lawsuit alleges that Pathway used clever queries to trick the OpenEvidence platforms into revealing its own underlying structure, training, prompts and logic. The complaint alleges violations of: (1) the Defend Trade Secrets Act; (2) the Computer Fraud and Abuse Act; (3) breach of contract through violations of OpenEvidence's Terms of Use; and (4) unfair competition under Massachusetts General Laws. Pathway's Motion to Dismiss argues that OpenEvidence failed to allege access to any nonpublic information—contending that "prompt injection" through a public interface is simply lawful reverse engineering. *See id*. (referencing defendant's prompt-injection and reverse-engineering arguments).

This raises a cutting-edge question in AI trade secret law: *Can internal system prompts remain protected as trade secrets if they can be partially inferred or extracted through clever user queries?* The case centers on whether "tricking" a model into revealing aspects of its internal logic constitutes misappropriation—or whether such conduct is merely permissible competitive investigation. The answer will likely influence whether prompt injection can qualify as "improper means" under trade secret statutes and whether Terms of Use alone are sufficient to protect sensitive AI system components, or whether technical safeguards and audit logs are necessary to establish reasonable secrecy measures.

The complaint further alleges that former OpenEvidence employees joined Pathway Medical and improperly transferred confidential materials, including (a) system prompts engineered to retrieve and weight medical research; (b) structured schemas for evaluating evidence quality; and (c) proprietary formatting frameworks for clinical summaries. OpenEvidence maintains that these internal components derived independent economic value from not being publicly known and had been subject to reasonable confidentiality and access controls.

This case underscores that trade secrets in the AI context extend far beyond raw training data. Proprietary prompts, internal weighting systems, model architectures, and workflow engines are increasingly viewed as the functional "core logic" of an AI system—assets with commercial value independent of the underlying datasets. For AI developers, *OpenEvidence* high-

lights the importance of implementing and documenting robust confidentiality safeguards, including role-based access controls, technical restrictions against prompt injection attacks, and clear audit trails for configuration changes.

C. Proprietary Database Protection and AI Training

Another emerging frontier in AI-related trade secret litigation involves claims that AI developers misappropriated proprietary datasets, curated compilations, or high-value research repositories to train or fine-tune models. These cases differ from those involving individual documents or isolated prompts. Instead, they target the aggregated value of structured, curated information that derives its economic significance from the way it has been assembled, organized, or quality filtered. Unlike copyright's fair use doctrine, trade secret law provides no general fair use defense, making these claims particularly potent where confidential compilations are involved. *See, e.g., DirecTV, LLC v. Delvecchio*, 2017 WL 1483374 (D. Conn. Apr. 24, 2017); *Compulife Software Inc. v. Newman*, 959 F.3d 1288 (11th Cir. 2020).

Courts have long recognized that databases—particularly those reflecting expert curation or structured intellectual labor—can constitute trade secrets when they derive independent economic value from not being generally known. In the AI context, plaintiffs allege that companies have unlawfully ingested proprietary compilations such as: (a) subscription-based research databases; (b) curated scientific or medical corpora; (c) specialized financial-market intelligence; or (d) internal business repositories containing customer data, pricing analytics, or operational metrics.

These allegations do not depend on copyright ownership. Instead, they turn on whether: (a) the database was economically valuable because of its confidentiality; (b) the plaintiff took reasonable steps to maintain secrecy; and (c) the defendant acquired, used, or disclosed the information through improper means. Improper means may include scraping behind authentication walls, breaching API-access limits, circumventing rate-limit controls, or exploiting insider relationships to obtain nonpublic information. For AI developers, database-related trade secret claims raise unique risks because even a portion of a proprietary compilation may support misappropriation theories. Courts increasingly focus on whether an AI developer benefited not from the underlying public-domain content, but from the structure, curation, or classification logic of the proprietary corpus—its taxonomies, metadata, weighting logic, or internal tagging. Thus, a defendant may face liability if it used the organizational value of a protected compilation, even if many individual documents were independently accessible.

D. Discovery Challenges: The "Black Box" Problem

A further challenge in trade secret, or really any litigation involving AI systems, arises from the inherent opacity of modern machine learning models. Unlike traditional software—where each instruction is generally human-written and traceable—large language models derive their capabilities from vast parameter sets generated through training (massive downloading and processing of data into an LLM). As a result, neither developers nor users can always explain why the model produced a given output or identify precisely which parts of the training data influenced a particular response. *See* Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Systems*, 3 Big Data & Society 1 (2016).

This opacity complicates trade secret cases in several ways. First, plaintiffs may allege that confidential or proprietary information must have been used in training because the model appears capable of generating outputs that resemble or reference protected material. Defendants, however, face difficulty proving the negative—that the model's behavior is the result of statistical generalization rather than exposure to the plaintiff's dataset. Second, because model weights and internal representations are difficult to interpret, defendants may struggle to demonstrate that (a) protected information was not incorporated into the model's internal logic; (b) the company employed reasonable measures to prevent unauthorized use of confidential materials; or (c) any similarities are incidental rather than the result of misappropriation.

This "black box" problem is not limited to trade secret cases. It has also begun to surface in insurance and bad faith litigation, where courts must evaluate allegedly AI-driven claims or coverage decisions while respecting proprietary algorithms. For example, in *Estate of Lokken v. UnitedHealth Group*, the plaintiffs challenged coverage determinations allegedly made by internal AI systems (nH Predict), raising questions about how much of the algorithmic logic and training data must be disclosed in discovery. *See Estate of Gene B. Lokken et al. v. UnitedHealth Group, Inc. et al.*, No. 0:23-cv-03514 (D. Minn. filed Nov. 14, 2023). Cases like *Lokken* illustrate how opacity can affect judicial review across domains, not just in intellectual property disputes. We will write more about the ongoing *Lokken* case in future FTD articles.

Trade secret law offers another significant avenue for claims against AI developers and downstream enterprise users.

The "black box" nature of AI systems also affects how courts evaluate reasonable secrecy measures. Plaintiffs increasingly argue that companies should implement technical safeguards—such as prompt-injection defenses, audit logs, access controls, and retraining protocols—to prevent inadvertent incorporation or disclosure of trade secrets. This raises the question whether traditional confidentiality measures are sufficient in the AI context, or whether the standard of "reasonable efforts" will evolve to include AI-specific controls.

Conclusion

The first wave of AI litigation has already revealed how quickly courts are adapting established legal frameworks to address

new technological realities. As the *Bartz* decision demonstrates, copyright law will continue to shape the boundaries of lawful model training. At the same time, the rapid expansion of alternative theories grounded in Terms of Service, the CFAA, unjust enrichment, trade secrets and insurance coverage and bad faith law, shows that copyright is only one part of a far broader litigation landscape.

Taken together, these developments illustrate a central theme: plaintiffs are increasingly focused on the inputs to AI systems—the origins of training data including prompts, rewards and bias, the terms of access, the confidentiality of internal logic, and the safeguards employed during training. The Bayesian framework introduced in Part I provides a structured way to anticipate how these cases reshape litigation risk, as each new decision updates the probabilities associated with emerging theories. Defense lawyers need to be prepared to understand AI, understand probability theory and know how to employ their own AI of choice to augment critical legal reasoning and expressions of probability to clients.

For businesses and law firms, proactive diligence is essential. Policies, procedures, guidelines and ethical guardrails must be established for AI use and deployment. Clear documentation of data lineage, compliance with contractual restrictions, and robust confidentiality and ethical controls will be key to mitigating exposure. As courts continue to confront questions involving data provenance, unauthorized access, and model opacity, the underlying principles discussed in this Part 1 provide a foundation for understanding where AI litigation is headed and how practitioners can prepare for what comes next. Look for future Defending the Algorithm™ articles in FTD..