WINSTON & STRAWN LLP

ARTICLE

IP's Real Legal Frontier With AI Is Everything Before the Output

MAY 8, 2026

This article was originally published in Bloomberg Law. Any opinions in this article are not those of Winston & Strawn or its clients. The opinions in this article are the authors' opinions only.

The legal conversation around artificial intelligence and intellectual property has fixated on outputs: Can an AI-generated image be copyrighted? Who owns a patent on an AI-assisted invention?

These are critical questions, but they bypass the crux of the issue. The most significant legal risks begin with model training, long before a system produces anything.

By the time a company is thinking about output liability, it often already has traversed a gauntlet of IP and privacy risks that most legal teams have yet to fully map. For general counsel, the focus should extend to everything that goes into an AI system, and everything that happens to it during the training process.

THE DATA PIPELINE

Large-scale AI models require enormous quantities of training data, and that data rarely arrives in an IP-neutral state. Copyright creates the most obvious exposure. Copyright law protects original works of authorship, but not facts themselves.

When training data consists of literary works, music, or photographs, infringement exposure is heightened. When it consists of facts, the risk shifts: Copyright doesn't protect the facts, but it does protect sufficiently creative compilations of them, and organizations increasingly are structuring datasets to obtain those protections.

The fair use doctrine offers a potential defense where use is sufficiently "transformative." The framework from the US Court of Appeals for the Second Circuit's 2016 ruling in *Authors Guild v. Google Inc.* has been widely cited as a possible safe harbor, but courts have yet to apply it squarely to generative AI training.

The commercial scale of model training creates real tension with the fourth fair use factor's market harm analysis. GCs should treat training data composition as a litigation risk variable, not as a settled legal baseline. Companies that document their data sources and make deliberate choices about licensed versus scraped content will fare meaningfully better in discovery.

Patent exposure at the input stage is subtler but potentially more disruptive. Two theories warrant serious attention.

Under US Code Section 271(a), if a model learns a patented method from training data and executes that method at inference, each inference run may constitute direct infringement—and how the model learned the method is no defense. Under Section 271(b), a company whose model generates instructions for performing a patented process may face induced infringement liability where users foreseeably follow them. The US Supreme Court in *Global-Tech Appliances v. SEB S.A.* held that willful blindness to a known patent is no shield.

Neither theory has been resolved by the Federal Circuit in the AI context. Companies building models on technical corpora—in life sciences, semiconductors, financial engineering, and cybersecurity—should be conducting freedom-to-operate analysis on the methods their models are trained to replicate, not just the outputs they produce.

PERMANENT STAKES

Trade secrets increasingly are the IP protection vehicle of choice for organizations with valuable data. When sharing that data with third parties, sophisticated organizations are crafting narrow licenses, robust confidentiality obligations, favorable risk allocation, and meaningful audit rights to ensure appropriate use.

A particular area of exposure is employee misuse of confidential information within third-party AI tools. When an employee inputs confidential business information into a platform whose terms permit using inputs for training, the harm isn't merely disclosure—it may be the permanent destruction of trade secret status.

Under the Defend Trade Secrets Act and state Uniform Trade Secrets Act analogs, protection depends on secrecy. Once information is incorporated into model weights queried by users worldwide, it is effectively in the public domain. There is no file to delete and no injunction that meaningfully restores the status quo, unlike an unauthorized cloud upload. The trade secret may simply be gone. And if the model later reproduces that information in response to third-party prompts, the deploying company, not just the employee, faces potential misappropriation liability.

The reverse direction is equally serious. Companies training on publicly available data may unknowingly ingest information that was itself misappropriated before entering the public corpus—a former employee's GitHub post containing proprietary source code, or a contractor's blog reproducing internal methodology. Civil liability under the DTSA doesn't require intent. Use of a trade secret obtained through improper means, even indirectly, can constitute misappropriation.

Trade secret risk demands governance frameworks before training begins. These include AI tool vetting protocols, technical protections, thoughtful contracts, compliance monitoring, and training data provenance documentation.

PRIVACY ISSUES

Layer privacy law over any of the above, and the risk calculus multiplies. Training data frequently contains personal information, whether intentional or not. Large publicly available datasets used for large language model training often include personal information subject to privacy laws, particularly in the EU, which takes a strict regulatory approach.

However, the most significant liabilities arise when training data comes from a company's own customers. This use case implicates not just privacy law but foundational IP questions: Who owns the data? Does the company have the rights to use it for AI training? If personal information must be de-identified to satisfy privacy restrictions, does the company have the contractual right to de-identify it, use the result for its intended purpose, and commercialize it? These are enumerated IP rights that may or may not exist in the underlying agreement.

For GCs navigating this terrain, the practical imperative is to treat privacy and IP diligence as a single integrated workstream—not sequential checkboxes, but parallel disciplines applied from the earliest stages of AI development.

The companies that get this right won't just avoid liability. They'll build AI systems on foundations that hold up.

4 Min Read

Related Capabilities

Intellectual Property | Artificial Intelligence (AI)

Related Professionals

Kathi Vidal

Alessandra Swanson

Mary Katherine Kulback