BERGER SINGERMAN
FLORIDA'S BUSINESS LAW FIRM

IP MATTERS: WHAT BUSINESSES SHOULD KNOW BEFORE USING ONLINE CONTENT TO TRAIN AI

November 9, 2026

Artificial intelligence systems need data, often enormous quantities of it. For businesses developing or customizing AI tools, the internet may look like an endless supply of articles, images and other useful content.

But the ability to access content does not necessarily include the right to copy it, store it or use it to train an AI system.

Recent lawsuits filed against Brave Software bring that distinction into focus. The New York Times and several other newspapers allege that Brave copied their articles, incorporated the content into its search products, supplied it to AI companies and offered technology that bypassed publisher paywalls. The complaints assert separate claims involving direct and secondary copyright infringement, circumvention of access controls and unjust enrichment. The allegations have not been proven, but the cases provide an important warning for any business using online content to develop or operate AI systems.

Publicly Available Does Not Mean Free to Use

Content does not lose copyright protection simply because it appears on a public website. A person may be able to read an article or view an image without charge, but that does not automatically authorize a business to make thousands of copies, create a permanent repository or use the material to build a commercial product.

Scraping is not automatically unlawful, and using copyrighted material to train an AI system may present a fair-use argument depending on the circumstances. But fair use is not a blanket exemption for anything found online.

The analysis may depend on:

- What content was collected;
- How the business obtained it;
- Whether the content was copied or permanently stored;
- Whether the use complied with applicable website terms;
- Whether access controls were avoided;
- What purpose does the AI system serve; and
- Whether the system reproduces or substitutes for the original material.

The details matter. Scraping a public webpage, copying content into a training repository and bypassing a paywall are different acts that may create different legal claims.

The Method of Access Can Create Its Own Risk

A business might have a credible fair-use argument regarding AI training and still face exposure based on how the material was acquired.

Website terms may restrict automated collection, commercial reuse or AI training. A subscription may authorize employees to read content without permitting the company to download it in bulk. Circumventing a paywall, login requirement or other technological control may create additional risk under the Digital Millennium Copyright Act, even when the underlying copyright question remains unresolved.

Technical restrictions are not merely inconveniences to engineer around. They may provide evidence that the content owner did not authorize the proposed use.

A License Is Only as Good as the Rights Behind It

Purchasing data from a vendor does not automatically solve the problem. Calling a collection of material "licensed," "curated" or "publicly available" does not establish that the provider had the authority to supply it for AI training.

A vendor cannot transfer rights it never possessed.

Before acquiring content or engaging a provider, businesses should determine:

- Where the material originated;
- Whether the provider collected it directly or obtained it from another source;
- Whether commercial AI training is expressly permitted;
- Whether the data may be stored, modified and combined with other information;
- Whether the resulting model may continue to be used after the license ends;
- Whether the provider will defend and indemnify the customer against infringement claims; and
- Whether the provider can produce records supporting its claimed rights.

A promise that the data is "safe to use" is not a chain of title.

Do Not Outsource the Diligence

Many businesses will not scrape content or train a foundational model themselves. They may purchase a data set, customize a third-party platform or hire a vendor to build an AI tool. That changes who performs the work, but it does not eliminate the customer's potential exposure.

Contracts should address data sources, permitted uses, compliance with access restrictions, ownership of inputs and outputs, recordkeeping, infringement claims and responsibility for removing disputed material. Businesses should also understand whether their information will be used to train a vendor's broader models.

If a provider cannot explain where important training material came from, that uncertainty may become the customer's problem when a claim arises.

Build a Defensible Process Before Training Begins

Removing information after it has been incorporated into an AI model may be technically difficult, expensive or impossible. The time to evaluate the data is before it enters the system.

Businesses developing or deploying AI should consider:

- Maintaining records identifying the source of each data set;
- Separating licensed, public-domain and internally created content;
- Reviewing website terms, subscriptions and access restrictions;
- Limiting collection to material necessary for the intended purpose;
- Testing outputs for reproduction of protected content;
- Establishing procedures for disputed or improperly sourced material; and
- Coordinating among legal, technology and information-security teams.

The law surrounding AI training is still developing. Businesses do not need to stop innovating while courts resolve these issues, but they should be able to answer three basic questions: What data are we using, where did it come from and what gives us the right to use it?

If those answers are unclear, the company may be training more than its AI system. It may also be training the plaintiffs' lawyers where to look.

Berger Singerman's Intellectual Property Team advises businesses regarding artificial intelligence, copyright, data licensing and technology agreements. Please contact Geoffrey Lottenberg or another member of the Intellectual Property Team with questions about developing or using AI systems and training data.