Bloomberg Law

Professional Perspective

AI, Data Ownership, & Market Share

Sameer Vadera, Kilpatrick Townsend & Stockton

AI, Data Ownership, & Market Share

Contributed by Sameer Vadera, Kilpatrick Townsend & Stockton

Artificial intelligence is a keystone in today's data economy. Over a third of all businesses globally use AI in business applications—an increase of about 270% over the last four years. Businesses typically leverage AI in their applications to drive new revenue streams, gain deeper customer insights, or define product enhancements. But today's AI-powered applications are possible because of Big Data—a massive firehose of data so complex that it cannot be processed by traditional hardware and software tools.

Using AI to support a business application generally involves collecting and curating Big Data into a digestible form, called training data. The training data is then fed into learning algorithms designed to detect patterns and correlations within the training data. The detected patterns and correlations are captured in an AI model, which is trained to make inferences on new, previously unseen data. These inferences are typically incorporated into an intelligent business solution like providing bespoke product recommendations to customers.

Often overlooked, however, training data can yield an enormous business value and competitive advantage in the market. Therefore, assessing the legal frameworks governing the ownership of training data is critical to the success of a data-driven business. This article will briefly summarize three legal frameworks for data ownership in the AI context: the intellectual property (IP) framework, the state competition law framework, and the federal statutory framework.

Market Differentiator

A trained AI model produces inferences about new, previously unseen information. When a trained AI model is built into a business application, the accuracy of the algorithmically generated inferences generally determines the superiority of that business application over competing applications in the market. Business applications that produce highly accurate AI-generated inferences tend to enjoy a larger market share as compared to less accurate competing applications.

For instance, given two competing applications that produce estimated arrival times for traveling on a route, customers tend to use the application that produces more accurate predictions of arrival time. Thus, in the AI space, the accuracy of AI-generated inferences often correlates to market share in terms of usership.

But the accuracy of an AI model's inferences is heavily influenced by the size and quality of the underlying training data from which the model was built. When the training data is sufficiently large and diversified, the patterns and correlations detected by the learning algorithms can be effectively generalized to produce accurate inferences about new and unexpected real-world data. When training data is not sufficiently large and diversified, however, data privacy risks could arise.

Risk of Improperly Trained Models

Several applications now use AI to provide predictive tasks relating to sensitive information. Today, for example, AI applications can predict health issues from medical records or infer the next sentence to write in an email from private emails. When sensitive information is used as training data, however, certain types of improperly trained AI models can unintentionally memorize the sensitive training data and reveal it as an inference in response to new data.

For instance, an AI application that auto-completes phrases is improperly trained using training data that also includes a small set of real credit card numbers. When entering the phrase, "My credit card number is 3," for example, the improperly trained AI application would output a real credit card number that starts with "3," instead of a generalized prediction for auto-completing the phrase. Hackers entering different prefixed texts can "attack" this improperly trained AI model to steal the real credit card numbers.

This phenomenon is also a threat to AI models trained to infer word associations from a business's internal information. A simple inference of uncommon word combinations from that information can reveal trade secrets. Fortunately, there are data science techniques that can prevent this problem; however, this phenomenon is still a significant concern to data privacy, given the fast pace at which AI applications are released to the public.

Impact of Data Loss

Collecting sufficiently large and high-quality training data is tedious and costly. What if the training data falls into the hands of a competitor? Exposing training data to a competitor allows the competitor to bypass the effort and cost involved in collecting and curating the training data. With cheaply acquired training data, the competitor can train or improve its own AI models to inexpensively build a competing application that produces highly accurate inferences. To fend off competing applications, considering data ownership of training data becomes critical.

Losing training data to competitors creates a risk of losing market share, too. For instance, a company in China built a wildly popular mobile application that leveraged AI models to predict arrival times of public transportation buses with high accuracy. To train the AI models, the company invested a large amount of money to affix sensors to city buses. The sensors collected real-time location data at a Big-Data scale and used an IoT network to send the collected data to cloud servers.

The real-time location data collected from buses across the city was curated into training data, which was used to train the AI models to produce accurate predictions of future bus arrival times. The mobile application quickly gained 50 million users, 4 million of which were daily active users. News outlets described the company's mobile application as a must-have for office workers due to its highly accurate predictions of bus arrival times.

A direct competitor reached out to the company to purchase their real-time location data. When rebuffed, the competitor created crawler software that mimicked different users of the company's mobile application by constantly changing its IP address. The crawler software managed to retrieve massive portions of the real-time location data from the mobile application. The competitor then used the real-time location data to train and improve the accuracy of the bus-arrival predictions produced by its own competing mobile application.

Over several months, the competitor siphoned market share from the company's mobile application as users switched to the competitor's mobile application now that the accuracy of the competitor's mobile application improved. While the company went to court in China accusing the competitor of unfair competition, the company still suffered economic loss and damage to its brand as a direct result of losing its training data to its competitor.

IP, State Competition Laws, and Federal Statutes

Given the enormous competitive advantage of training data in the context of AI applications, considering the legal frameworks for data ownership becomes critical to the success of a data-driven business. In general, data ownership is governed by at least three legal frameworks: the IP framework, the competition law framework, and the federal statutory framework.

IP Framework

Copyrights. These can potentially protect curated training data as a whole, but not its individual facts. Big Data itself is not in a form that is ready for analysis. Curating Big Data involves complicated tools that verify, clean, and transform the Big Data into training data. The organizational structure of the curated training data, however, is protectable, as long as the selection or arrangement of the data elements is original or creative. Often, the steps taken to transform Big Data into curated training data are intensive, costly, and specific to the use case of the AI-based application. In this scenario, copyrighting curated training data is not generally a high bar to pass.

Trade Secrets. Training data can be protected under trade secret law, if reasonable measures are taken early to keep the training data secret. Training data derives independent economic value from not being generally known or readily ascertainable through proper means. For instance, training data heavily influences the accuracy of a business's AI-based application, which in turn, drives the business's market share. The steps taken to keep training data under lock and key within a business's network architecture are critical to the prospects of trade secret protection. Trade secret protection can protect individual facts, unlike copyright protection.

State Competition Law Framework

Competition law claims may arise if a competitor steals curated training data. Stealing curated training data is a competitive action. In the context of AI, losing training data to a competitor can result in lost market share. For instance, losing training data enables the competitor to inexpensively compete for market share by improving the accuracy of the competitor's AI-based application with the misappropriated training data. Additionally, the competitor gets to bypass the significant cost

and effort involved in acquiring, storing, and curating the training data. State competition law claims, such as lost profits, unfair competition, and electronic trespass to chattel, have proven fruitful in this scenario.

Federal Statutory Framework

Computer Fraud and Abuse Act. The CFAA prohibits intentionally accessing a computer "without authorization." In the context of AI, losing training data as a result of hacking would likely fall under the CFAA. But what if some of a business's training data is accessible on the business's public-facing website? For example, this is the case when user profiles on a social media platform are publicly searchable.

Competitors can create data scrapers that collect the publicly available user profiles and use them to train AI-models to provide competing services. Many websites define this type of data scraping as unauthorized access in their terms of use because exposing publicly accessible training data to data scrapers can interfere with a business's AI-based services.

Businesses have asserted CFAA claims against competitors deploying data scrapers, but courts have generally interpreted unauthorized access by data scraping as falling outside of the prohibited "without authorization" covered by the CFAA; especially for data collected from public-facing websites. See hiQ Labs, Inc. v. LinkedIn Corporation, No. 17-16783, D.C. No. 3:17-cv-03301-EMC.

Business Impact

The size and quality of training data is a lead factor in determining the accuracy of an AI application. In many lines of business, the accuracy of an AI application is what drives usership. The size and quality of training data heavily influences the market share of a business's AI application. Considering the various legal frameworks that govern data ownership of training data is, therefore, critical for a data-driven business employing AI.

The IP framework (e.g., copyrights and trade secrets), the competition law framework (e.g., state claims of unfair competition), and the federal statutory framework (e.g., asserting the CFAA against data scrapers), among others, are available to pursue to assert data ownership of training data against a competitor.

Given that today there is no comprehensive database protection law in the U.S., taking a kitchen-sink approach when asserting data ownership of training data may be the best strategy to fend off competitors and maintain a dominant market share position.