ALM | LAW.COM

New York Law Journal

U.S. and EU Regulators on Diverging Paths Over AI Training Data

By David Owen & Ken Ritz
May 27, 2025

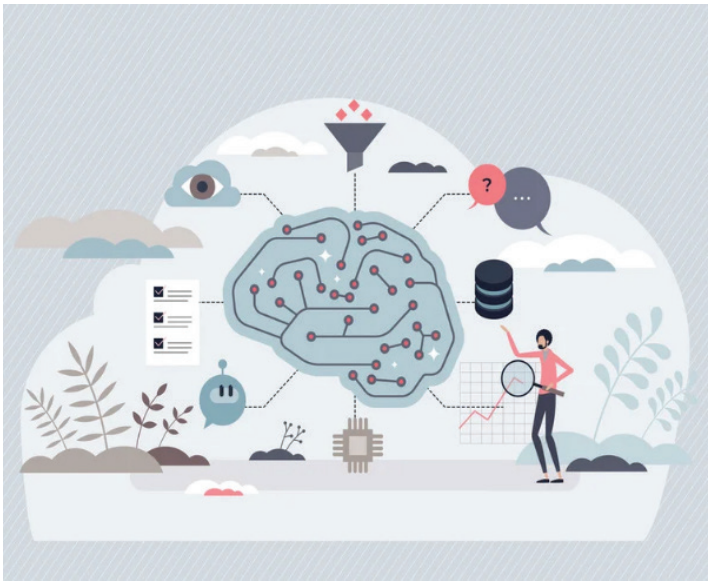
Generative artificial intelligence (AI) has emerged as one of the most potentially transformative technological innovations of our time, and a race is on among governments and tech companies around the world to harness and control this fast developing and disruptive technology.

While most users of ChatGPT likely never consider the amount of training data (the dataset that is used to teach a model how to perform a task) that was assimilated in order to generate useful content in response to their prompts, it is an immense volume of material.

The training data used by GPT-4, OpenAI's latest model, reportedly includes an incredible 1 petabyte of data, the equivalent of 1 million gigabytes, or roughly 22 times the Library of Congress's entire book collection.

AI training data can include anything that can be scraped from accessible digital sources. Data that is doubtful, biased and false is generally a part of the package, as well as social media postings, often including private content that is inadvertently exposed.

The provenance of the data is typically unimportant, sweeping up hacked content and anything inadequately secured. Because AI models cannot effectively train themselves on their own output, known as synthetic data, they require the regular infusion of new training data to evolve and maintain integrity.



An illustration demonstrating artificial intelligence.

As a result, there is now a rapidly expanding demand and market for usable AI training data and for innovative ways to capture more data and refine it to new applications.

While the awesome size and diversity of data available to the public offer enormous potential and opportunity, the indiscriminate gathering and assimilation of data carries a variety of risks and policy concerns.

People whose information or work product has been assimilated improperly or illegally are

essentially without any remedy. In addition, flawed and biased training data can yield inaccurate results, amplify prejudices, produce socially harmful output, and expose users to injury from unreliable results.

Against these concerns is pitted the national and commercial urgency to advance this technology as quickly as possible.

EU Takes A Hands On Approach

Acknowledging the risks of this massive consumption of data into AI models, European Union (EU) regulators have begun to contemplate limitations on the types of data that are permissible for use in AI training sets, including limits on the assimilation of information that is publicly available.

The recently adopted Artificial Intelligence Act (the EU AI Act), among other important measures, sets forth standards for data set quality, validation, and testing in an effort to protect people against risks to health, safety, fundamental rights, and other public interests.

The EU AI Act provides notably coercive authority to regulators, including hefty fines for noncompliance, potentially reaching the higher of 7% of a provider's annual turnover or €35 million, and specifically addresses the issue of data quality in the training of AI models.

Article 10 of the EU AI Act, which is ostensibly effective on August 2, 2026, mandates that training data sets must be "relevant for the intended purpose, representative of the target population, accurate, consistent, unbiased, and complete."

The EU AI Act also calls for transparent and ethical data collection practices especially with respect to certain categories of personal data, which are subject to enhanced safeguards to protect individual rights and freedoms.

A fair amount of uncertainty and disagreement surrounds virtually every element of these new AI proscriptions, along with an expectation that they may be subject to change.

The European Commission has also announced plans to launch a General-Purpose AI Code of Practice, which aims to detail the manner in which providers may comply with their obligations under the EU AI Act and includes a template for summarizing training data used in general-purpose AI models in order to ensure transparency, trust, and compliance with laws in the development and deployment of AI systems.

Most recently, in April 2025, the EU announced its AI Continent Action Plan, aimed at making Europe a global leader in AI. The Action Plan asserts that it focuses on promoting the advancement of Europe's competitiveness in the marketplace and prioritizes the trustworthiness of AI tools, safeguarding and advancing democratic values, upholding fundamental rights, and addressing safety risks specific to AI systems.

Together with the advancement of these new guidelines, the European Commission has actively investigated U.S. tech companies and their data practices relating to AI. In March 2024, the EU launched a probe aimed at companies such as Meta, Microsoft, Snap, TikTok, and X Corp., focusing on how these providers manage the risks of generative AI while offering consumer-facing AI tools.

EU regulators have also opened inquiries regarding big tech's gathering of personal data to develop AI models, citing privacy concerns. These inquiries include inquiries of Google and Meta by the Ireland Data Protection Commission (DPC) concerning whether EU users' personal data is adequately protected before being assimilated into AI models, which led to the delay of Meta's EU launch.

The DPC noted that these inquiries were part of its wider initiative to regulate the processing of personal data in the development of AI models and systems. In April 2025, the DPC announced an investigation of X Corp. over the use of personal data of EU users to train its AI system Grok.

Notably, the inquiry includes a review of the processing of personal data obtained from publicly-accessible social media posts and the extent to which consumers retain control of their personal data even when placing it in the public domain.

If the EU prohibits AI providers from assimilating public postings over potential privacy concerns, it could have a very significant effect on the amount of data that is lawfully available.

The U.S. Signals a Hands Off Approach

In contrast to the EU's developing regulatory posture, the current U.S. administration has signaled that it favors a mostly unfettered use of public data for training by U.S. AI companies. President Donald Trump revoked a previous order by then-President Joe Biden (the Biden Order), which had outlined a framework for managing the proliferation of AI, including provisions for promoting the safety, security, and trustworthiness of AI systems.

Trump issued his own "Order for Removing Barriers to American Leadership in AI" (the Trump Order), calling for federal agencies to revise or rescind all actions under the Biden Order that are "inconsistent with, or present obstacles to" the stated goal of "enhanc[ing] America's global AI dominance."

Other Trump administration officials have explicitly criticized the EU AI Act itself, with Vice President Vance announcing at an AI summit in Paris that "[t]he AI future will not be won by hand-wringing about safety."

Regarding AI training data, the Trump Order asserts only that systems should be free from "ideological bias or engineered social agendas." The current administration also recently declined to sign a pledge by 60+ countries to make AI safe, ethical, and transparent.

U.S. Companies Forge Ahead Under Diverging Regulatory Regimes

Consistent with the evolving U.S. regulatory posture, U.S. tech giants are pushing back against EU regulations, stating that they are stifling innovation and delaying the roll-out of products to consumers.

Driven by the belief that more training data will produce better AI, U.S. companies are reportedly expanding AI data collection unabated, including new plans to harvest and assimilate public data into their AI models. Underlying their determination is a belief that any supposed harms from their AI data collection and processing efforts are mostly abstract and speculative, while the benefits of the AI systems that flow from them are concrete and obvious.

Notably, content creators and others who have objected to AI training data collection practices in the U.S. have had relatively little success with their claims and difficulty identifying specific harm to allege.

Against this backdrop, a recent €30.5 million ($33.7 million) fine issued by the privacy watchdog in the Netherlands to the U.S. company Clearview AI illustrates the growing divide between AI regulators on both sides of the Atlantic.

Clearview is a private company that provides facial recognition technology and an investigative platform primarily to law enforcement and other government agencies. The company has reportedly collected billions of publicly available photos and undertaken to biometrically analyze every face for recognition purposes.

As noted by the Dutch regulators purporting to impose the fine: "Facial recognition is a highly intrusive technology. . . . If there is a photo of you on the Internet – and doesn't that apply to all of us? – then you can end up in the database of Clearview and be tracked." The company's response was equally pointed:

"This decision is unlawful, devoid of due process and is unenforceable," noting that the company "does not have a place of business in the Netherlands or the EU…, does not have any customers in the Netherlands or the EU, and does not undertake any activities that would otherwise mean it is subject to the [General Data Protection Regulation]."

While it seems currently impractical and inefficient for the large U.S. tech companies involved in the AI race to disengage from the EU and its regulators as Clearview has, the rapidly diverging regulatory and enforcement approaches in the U.S. and the EU will likely accelerate consideration of new ways to comply without getting out of the race.

David Owen *is a partner at Cahill Gordon & Reindel.* **Ken Ritz** *is a counsel at the firm.*