GOODWIN

ALERT · JUNE 10, 2025

California's AB 2013: Generative AI Developers Must Show Their Data

BY Omer Tene Bethany P. Withers Reema Moussa

California Assembly Bill 2013 (AB 2013), known as the Generative Artificial Intelligence: Training Data Transparency Act, was signed into law on September 28, 2024, and is set to take effect on January 1, 2026. The legislation mandates developers of generative artificial intelligence (AI) systems or services to disclose detailed information about the datasets they used to train their models. It appears to use the terms "data" and "datasets" to broadly mean any item (including raw information, structured data, and copyrightable works) used to train the applicable generative AI system or service, as evidenced by the law's suggestion that data could be "protected by copyright." US law does not afford copyright protection to data points but rather to creative expressions.

The law marks a broader legislative effort in California to regulate AI. Other notable laws passed in conjunction with AB 2013 include Senate Bill 942, which requires large AI systems to implement watermarking and detection tools for AI-generated content, and AB 3030, which mandates disclaimers for AI-generated patient communications in healthcare settings. Additionally, the California Privacy Protection Agency has debated and refined a detailed set of regulations, which includes a prescriptive section on automated decision-making technology (ADMT). These new laws reinforce California's role as a leader among states in the AI regulatory, policy, and legislative space, especially as other proposed bills (even if paused for now) are modeled on AB 2013.

At the same time, in Washington, DC, the House of Representatives has recently passed a sweeping 10-year federal moratorium on state regulation of AI systems, AI models, and automated decision systems, as part of President Trump's "One Big Beautiful Bill Act" budget reconciliation package. If enacted, the federal moratorium would preempt California's laws regulating AI and ADMT. For now, however, businesses in this space should keep their eye on legislative developments, while preparing for the implementation of existing state legislation, as the deadline draws near.

Scope of Application

AB 2013 applies to any entity — individuals, corporations, or government agencies — that designs, codes, produces, o substantially modifies a generative AI system for public use by Californians. Generative AI is defined as AI "that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence's training data." Obviously, this would apply to the major generative AI developers such as OpenAI, Anthropic, Google, and Meta. But "substantial modification," which is defined to include updates that materially change the system's functionality or performance, such as retraining or fine-tuning, could expand the application of the law to a vast array of businesses that deploy the leading models in their products. Exemptions exist for systems used solely for security and integrity;

aircraft operation in national airspace; or national security, military, or defense purposes made available only to federal entities.

It's important to note that the law applies to systems or services released or substantially modified on or after January 1, 2022, that is, a full four years before the law comes into effect on January 1, 2026. Consequently, AB 2013 will impact a broad swath of, if not most, AI systems or products.

Core Requirements

Developers who are covered by the law will be required to post documentation on their websites providing detailed information about the data used to train their models, including the:

- **Sources and ownership** of the datasets
- **Purpose and methodology** of data collection, cleaning, processing, or modification
- **Data points** included in the datasets, such as types, labels, and counts
- **Licensing status,** indicating whether datasets are copyright-protected, purchased, licensed, or in the public domain
- **Inclusion of personal or aggregate consumer information,** as defined by the California Consumer Privacy Act
- **Cleaning, processing, or other modification** of datasets by the developer and the respective purpose for each modification
- **Use of synthetic data** in development and/or training of the system
- **Dates** when datasets were first used and the time frame during which the data was collected

Key Challenges

Implementing AB 2013 presents several key challenges for developers of generative AI systems. One of the foremost difficulties lies in assembling comprehensive documentation of training datasets, especially for models that have evolved over time or were built using large-scale, publicly scraped data. Many generative models incorporate data from heterogeneous sources, some of which may lack clear provenance or licensing information, making it difficult to comply with the law's detailed disclosure requirements. Additionally, verifying licensing status or identifying whether personal or consumer data was included may be practically infeasible for developers relying on third-party datasets or pretrained models (particularly from open-source communities). These obligations could also raise trade secret concerns, as disclosing dataset composition and processing methodologies may require revealing competitive or proprietary information, as noted, for example, by the Business Software Alliance in a 2024 press release.

In fact, several major industry groups have voiced strong opposition to AB 2013, citing concerns about what they called its potential to stifle innovation and burden businesses with excessive regulation. One group suggested that the law would force companies to reveal proprietary information, ultimately benefiting large incumbents over smaller startups. Other industry groups have argued that the bill's broad nature and deviation from the focus on "high-risk" AI systems (a model seen already through Colorado's and the European Union's AI acts, which we've covered in How States Are Stepping in to Regulate AI, The EU AI Act Is (Almost) Here, and The World's First AI Regulation Is Here) unnecessarily burden businesses, proposing alignment with the OECD's definition of AI to narrow down the law's scope.

Strategic Considerations

Strategically, organizations will need to reassess their data governance and development workflows. Developers should proactively document dataset usage at every stage — collection, cleaning, annotation, and model training — starting fron January 1, 2022, the law's retroactive cutoff. Companies may need to implement internal compliance frameworks, including audit trails and dataset registries. Legal teams must be closely involved in determining what constitutes a "substantial modification" to a system and whether an exemption applies. There's also a reputational component at play: publishing the information required by AB 2013 could increase transparency and public trust, but it also exposes companies to greater scrutiny

over potential security, bias, intellectual property (IP) infringement, or privacy risks. Given the broader momentum in AI regulation across jurisdictions, aligning compliance strategies with both US and international standards may offer efficiencies and mitigate long-term regulatory risk.

Next Steps for Developers

As organizations prepare for compliance with AB 2013, a critical next step is to initiate comprehensive internal audits of all training data practices, particularly for generative AI systems developed or substantially modified since January 1, 2022. These audits should identify data sources, licensing terms, and the presence of personal or proprietary content. Simultaneously, companies should establish standardized protocols for documenting and tracking dataset composition, usage, and updates, ensuring all relevant information required by the law — such as collection methods, ownership, and synthetic data use — is properly recorded. This could take the form of a data declaration covering these various data characteristics. Given the tension between transparency and security considerations as well as IP protections, businesses should engage legal counsel early to assess what disclosures are required and how to structure them to minimize the risk of revealing trade secrets or sensitive information.

While we'll be monitoring whether the federally proposed moratorium on state AI laws will render this law preempted, taking these steps now will allow organizations to operationalize compliance ahead of the January 1, 2026, enforcement date and position themselves for future regulatory developments and responsible AI governance.

CONTACTS

Omer Tene
Partner

otene@goodwinlaw.com
Boston | +1 617 570 1094

Bethany P. Withers
Partner
Chair, AI & Machine Learning

bwithers@goodwinlaw.com
Boston | +1 617 570 8732

Reema Moussa
Associate

rmoussa@goodwinlaw.com
New York | +1 917 229 7870