Proskauer

California Enacts Additional Generative AI Bills Touching on Training Data

New Media and Technology Law Blog on **September 30, 2024**

After several weeks of handwringing about the fate of SB 1047 – the controversial AI safety bill that would have required developers of powerful AI models and entities providing the computing resources to train such models to put appropriate safeguards and policies into place to prevent critical harms – California Governor Gavin Newsom announced that he vetoed the "well-intentioned" bill because it was not necessarily "the best approach to protecting the public from real threats posed by the technology."

However, amidst the hubbub surrounding SB 1047, the Governor in the past month has reportedly signed 17 bills covering generative AI ("GenAI"), including bills on deepfakes, AI watermarking, child safety, performers' AI rights and election misinformation, some of which we discussed in a prior post.

Most notably, over the weekend the Governor signed AB 2013, which would require that, on or before January 1, 2026, developers of GenAI systems or services released on or after January 1, 2022 post a "high-level summary" of the datasets used to train[1] such GenAI systems or services "made publicly available to Californians for use" (as well as post a summary of the training data after a "substantial modification" to a GenAI system or service that is made publicly available to Californians for use). The law lists multiple items that should be part of a "high-level summary," including:

- The sources or owners of the datasets
- The general number of data points included in the datasets and a description of the types of data points within the datasets
- Whether the datasets include any data protected by IP
- Whether the datasets were purchased or licensed by the developer
- Whether the datasets include personal information, or aggregate consumer information as defined under the CCPA

- Whether there was any cleaning, processing, or other modification to the datasets by the developer
- The time period during which the data in the datasets were collected or first used during the GenAI development process
- Whether the GenAI system or service uses synthetic data generation in its development.

Beyond the requirement that the covered GenAI system be "publicly available to Californians," which would perhaps exclude certain enterprise models that are not generally made available to the public, the law also exempts, among other things, GenAI systems whose "sole purpose" is for data security and integrity or used by a federal entity for national defense purposes.

AB 2013 is at the forefront of state AI regulation surrounding GenAI training data sources. GenAI developers generally do not disclose data sources used to train existing or frontier models for a multitude of business or competitive reasons. In addition, with a host of ongoing copyright and privacy-related litigations against major GenAI developers over allegations that such developers used copyrighted material or consumers' personal information for training purposes without authorization, training data sources have been at the center of such claims and not discussed in detail in public.

Other AI-related bills recently signed include:

- AB 1008: The law clarifies that "personal information" under the CCPA can exist in various formats, including, but not limited to, physical formats, digital formats, and abstract digital formats, which would also include "artificial intelligence systems that are capable of outputting personal information." [An Assembly Floor Analysis of AB 1008 law can be found here].
- SB 1120: The law requires, among other things, a health care service plan or disability insurer that uses an AI, algorithm, or other software tool for the purpose of coverage decisions and functions (or contracts for such services through a third party vendor) to ensure that the AI tool bases its determination on specified information and is fairly and equitably applied.
- AB 3030: The law requires a health facility, clinic, physician's office, or group practice that uses GenAI to produce certain written or verbal patient communications pertaining to patient clinical information to include a disclaimer that informs the patient that GenAI was used to generate the communication, along with instructions on how the patient may contact a human health care provider or

employee of the practice.

---

[1] Under AB 2013, “train a generative artificial intelligence system or service” includes “testing, validating, or fine tuning by the developer of the artificial intelligence system or service.”

View original.

Related Professionals

- **Jeffrey D. Neuburger**