Assembly Bill No. 2013

CHAPTER 817

An act to add Title 15.2 (commencing with Section 3110) to Part 4 of Division 3 of the Civil Code, relating to artificial intelligence.

[Approved by Governor September 28, 2024. Filed with
Secretary of State September 28, 2024.]

LEGISLATIVE COUNSEL'S DIGEST

AB 2013, Irwin. Generative artificial intelligence: training data transparency.

Existing law requires the Department of Technology, in coordination with other interagency bodies, to conduct, on or before September 1, 2024, a comprehensive inventory of all high-risk automated decision systems, as defined, that have been proposed for use, development, or procurement by, or are being used, developed, or procured by, state agencies, as defined.

This bill would require, on or before January 1, 2026, and before each time thereafter that a generative artificial intelligence system or service, as defined, or a substantial modification to a generative artificial intelligence system or service, released on or after January 1, 2022, is made available to Californians for use, regardless of whether the terms of that use include compensation, a developer of the system or service to post on the developer's internet website documentation, as specified, regarding the data used to train the generative artificial intelligence system or service. The bill would require that this documentation include, among other requirements, a high-level summary of the datasets used in the development of the system or service, as specified.

The people of the State of California do enact as follows:

SECTION 1. Title 15.2 (commencing with Section 3110) is added to Part 4 of Division 3 of the Civil Code, to read:

TITLE 15.2. ARTIFICIAL INTELLIGENCE TRAINING DATA TRANSPARENCY

3110. For purposes of this title, the following definitions shall apply:

(a) "Artificial intelligence" means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

(b) “Developer” means a person, partnership, state or local government agency, or corporation that designs, codes, produces, or substantially modifies an artificial intelligence system or service for use by members of the public. For purposes of this subdivision, “members of the public” does not include an affiliate as defined in subparagraph (A) of paragraph (1) of subdivision (c) of Section 1799.1a, or a hospital’s medical staff member.

(c) “Generative artificial intelligence” means artificial intelligence 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.

(d) “Substantially modifies” or “substantial modification” means a new version, new release, or other update to a generative artificial intelligence system or service that materially changes its functionality or performance, including the results of retraining or fine tuning.

(e) “Synthetic data generation” means a process in which seed data are used to create artificial data that have some of the statistical characteristics of the seed data.

(f) “Train a generative artificial intelligence system or service” includes testing, validating, or fine tuning by the developer of the artificial intelligence system or service.

3111. On or before January 1, 2026, and before each time thereafter that a generative artificial intelligence system or service, or a substantial modification to a generative artificial intelligence system or service, released on or after January 1, 2022, is made publicly available to Californians for use, regardless of whether the terms of that use include compensation, the developer of the system or service shall post on the developer’s internet website documentation regarding the data used by the developer to train the generative artificial intelligence system or service, including, but not be limited to, all of the following:

(a) A high-level summary of the datasets used in the development of the generative artificial intelligence system or service, including, but not limited to:

(1) The sources or owners of the datasets.

(2) A description of how the datasets further the intended purpose of the artificial intelligence system or service.

(3) The number of data points included in the datasets, which may be in general ranges, and with estimated figures for dynamic datasets.

(4) A description of the types of data points within the datasets. For purposes of this paragraph, the following definitions apply:

(A) As applied to datasets that include labels, “types of data points” means the types of labels used.

(B) As applied to datasets without labeling, “types of data points” refers to the general characteristics.

(5) Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.

(6) Whether the datasets were purchased or licensed by the developer.

(7) Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140.

(8) Whether the datasets include aggregate consumer information, as defined in subdivision (b) of Section 1798.140.

(9) Whether there was any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the artificial intelligence system or service.

(10) The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing.

(11) The dates the datasets were first used during the development of the artificial intelligence system or service.

(12) Whether the generative artificial intelligence system or service used or continuously uses synthetic data generation in its development. A developer may include a description of the functional need or desired purpose of the synthetic data in relation to the intended purpose of the system or service.

(b) A developer shall not be required to post documentation regarding the data used to train a generative artificial intelligence system or service for any of the following:

(1) A generative artificial intelligence system or service whose sole purpose is to help ensure security and integrity. For purposes of this paragraph, “security and integrity” has the same meaning as defined in subdivision (ac) of Section 1798.140, except as applied to any developer or user and not limited to businesses, as defined in subdivision (d) of that section.

(2) A generative artificial intelligence system or service whose sole purpose is the operation of aircraft in the national airspace.

(3) A generative artificial intelligence system or service developed for national security, military, or defense purposes that is made available only to a federal entity.

O