STATE OF NEW YORK

6955

2025-2026 Regular Sessions

IN SENATE

March 27, 2025

Introduced by Sen. GOUNARDES -- read twice and ordered printed, and when printed to be committed to the Committee on Internet and Technology

AN ACT to amend the general business law, in relation to establishing the artificial intelligence training data transparency act

The People of the State of New York, represented in Senate and Assembly, do enact as follows:

Section 1. The general business law is amended by adding a new article 44-B to read as follows:

ARTICLE 44-B

ARTIFICIAL INTELLIGENCE TRAINING DATA TRANSPARENCY ACT

§ 1420. Short title. This act shall be known and may be cited as the "artificial intelligence training data transparency act".

§ 1421. Definitions. For the purposes of this article, the following terms shall have the following meanings:

1. "Artificial intelligence" or "artificial intelligence technology" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments, and that uses machine- and human-based inputs to perceive real and virtual environments, abstract such perceptions into models through analysis in an automated manner, and use model inference to formulate options for information or action.

2. "Developer" means a person, partnership, state or local government agency, or corporation that designs, codes, produces, or substantially

EXPLANATION--Matter in *italics* (underscored) is new; matter in brackets [-] is old law to be omitted.

LBD07975-02-5

modifies an artificial intelligence model or service for use by members of the public.

3. "Generative artificial intelligence" means a class of AI models that are self-supervised and emulate the structure and characteristics of input data to generate derived synthetic content, including, but not limited to, images, videos, audio, text, and other digital content.

4. "Substantially modifies" or "substantial modification" means a new version, new release, or other update to a generative artificial intelligence model or service that materially changes its functionality or performance, including the results of retraining or fine tuning.

5. "Synthetic data generation" means a process in which seed data is used to create artificial data that have some of the statistical characteristics of the seed data.

6. "Train a generative artificial intelligence model or service" includes testing, validating, or fine tuning by the developer of the artificial intelligence model or service.

7. "Aggregate consumer information" means information that relates to a group of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a device. Aggregate consumer information does not mean one or more individual consumer records that have been de-identified.

8. "AI model" means an information system or component of an information system that implements artificial intelligence technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.

§ 1422. Data used to train generative artificial intelligence models or services. 1. On or before January first, two thousand twenty-six, and prior to each time thereafter that a generative artificial intelligence model or service, or a substantial modification to a generative artificial intelligence model or service, released on or after January first, two thousand twenty-two, is made publicly available to New Yorkers for use, regardless of whether the terms of such use include compensation, the developer of such model or service shall post on the developer's website documentation regarding the data used by the developer to train the generative artificial intelligence model or service, including a high-level summary of the datasets used in the development of the generative artificial intelligence model or service, including, but not limited to:

(a) the sources or owners of the datasets;

(b) a description of how the datasets further the intended purpose of the artificial intelligence model or service;

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

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

(i) as applied to datasets that include labels, "types of data points" means the types of labels used; and

(ii) as applied to datasets without labeling, "types of data points" refers to the general characteristics;

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

(f) whether the datasets were purchased or licensed by the developer;

(g) whether the datasets include personal information or personal identifying information, as defined in section eight hundred ninety-nine-aaa of this chapter;

(h) whether the datasets include aggregate consumer information;

(i) 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 model or service;

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

(k) the dates the datasets were first used during the development of the artificial intelligence model or service; and

(l) whether the generative artificial intelligence model 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 model or service.

2. A developer shall not be required to post documentation regarding the data used to train a generative artificial intelligence model or service for any of the following:

(a) A generative artificial intelligence model or service whose sole purpose is the operation of aircraft in the national airspace; or

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

§ 1423. Employee data used to train generative artificial intelligence models or services. 1. Any person, partnership, state or local government agency, or corporation that designs, codes, produces, or substantially modifies a generative artificial intelligence model or service using data of which a substantial part is derived from individuals employed or contracted by the entity, regardless if whether the model is made publicly available, shall ensure that the following information is disclosed to each employee whose data is used to train the artificial intelligence model:

(a) the intended purpose of the artificial intelligence model or service;

(b) a description of how the collected datasets further the intended purpose of the artificial intelligence model or service;

(c) a description of the types of data points within the datasets;

(d) whether the datasets include personal information or personal identifying information, as defined in section eight hundred ninety-nine-aaa of this chapter;

(e) the dates the datasets were first used during the development of the artificial intelligence model or service; and

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

2. An entity that uses employee or contractor data to design, code, produce, or substantially modify a generative artificial intelligence model or service shall not be required to disclose the information required by this section if the model or service:

(a) is solely intended to be used in the operation of aircraft in the national airspace; or

(b) is developed for national security, military, or defense purposes and only made available to a federal entity.

§ 2. This act shall take effect immediately.