

STATE OF NEW YORK

6578

2025-2026 Regular Sessions

IN ASSEMBLY

March 6, 2025

Introduced by M. of A. BORES -- read once and referred to the Committee on Science 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:

1 Section 1. The general business law is amended by adding a new article
2 44-B to read as follows:

ARTICLE 44-B

ARTIFICIAL INTELLIGENCE TRAINING DATA TRANSPARENCY ACT

Section 1420. Short title.

1421. Definitions.

1422. Data used to train generative artificial intelligence models or services.

1423. Employee data used to train generative artificial intelligence models or services.

11 § 1420. Short title. This act shall be known and may be cited as the
12 "artificial intelligence training data transparency act".

13 § 1421. Definitions. For the purposes of this article, the following
14 terms shall have the following meanings:

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

22 2. "Developer" means a person, partnership, state or local government
23 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

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

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

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

11 5. "Synthetic data generation" means a process in which seed data is
12 used to create artificial data that have some of the statistical charac-
13 teristics of the seed data.

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

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

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

27 § 1422. Data used to train generative artificial intelligence models
28 or services. 1. On or before January first, two thousand twenty-six, and
29 prior to each time thereafter that a generative artificial intelligence
30 model or service, or a substantial modification to a generative artifi-
31 cial intelligence model or service, released on or after January first,
32 two thousand twenty-two, is made publicly available to New Yorkers for
33 use, regardless of whether the terms of such use include compensation,
34 the developer of such model or service shall post on the developer's
35 website documentation regarding the data used by the developer to train
36 the generative artificial intelligence model or service, including a
37 high-level summary of the datasets used in the development of the gener-
38 ative artificial intelligence model or service, including, but not
39 limited to:

40 (a) the sources or owners of the datasets;

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

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

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

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

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

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

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

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

4 (h) whether the datasets include aggregate consumer information;

5 (i) whether there was any cleaning, processing, or other modification
6 to the datasets by the developer, including the intended purpose of
7 those efforts in relation to the artificial intelligence model or
8 service;

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

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

13 (l) whether the generative artificial intelligence model or service
14 used or continuously uses synthetic data generation in its development.
15 A developer may include a description of the functional need or desired
16 purpose of the synthetic data in relation to the intended purpose of the
17 model or service.

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

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

23 (b) A generative artificial intelligence model or service developed
24 for national security, military, or defense purposes that is made avail-
25 able only to a federal entity.

26 § 1423. Employee data used to train generative artificial intelligence
27 models or services. 1. Any person, partnership, state or local govern-
28 ment agency, or corporation that designs, codes, produces, or substan-
29 tially modifies a generative artificial intelligence model or service
30 using data of which a substantial part is derived from individuals
31 employed or contracted by the entity, regardless if whether the model is
32 made publicly available, shall ensure that the following information is
33 disclosed to each employee whose data is used to train the artificial
34 intelligence model:

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

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

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

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

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

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

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

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

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

55 § 2. This act shall take effect immediately.