STATE OF NEW YORK

8331--A

2025-2026 Regular Sessions

IN SENATE

June 3, 2025

Introduced by Sen. GONZALEZ -- read twice and ordered printed, and when printed to be committed to the Committee on Rules -- committee discharged, bill amended, ordered reprinted as amended and recommitted to said committee

AN ACT to amend the general business law, in relation to requiring transparency from generative artificial intelligence developers for journalism providers

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

Section 1. This act shall be known and may be cited as the "New York artificial intelligence transparency for journalism act".

§ 2. Legislative findings. The legislature hereby finds and declares that:

(a) A free and diverse press was critical in the founding of our democracy and continues to be the lifeblood for a functional society;

(b) New York has a compelling interest in protecting news publishers and broadcasters that report and distribute news from unfair business practices and competition. Every day, journalism plays an essential role in New York and in local communities, and the ability of local news organizations to continue to provide the public with critical information about their communities and enabling news publishers and broadcasters to receive fair market value for their content that is used by others will preserve and ensure the sustainability of local and diverse news outlets;

(c) Communities without newspapers and broadcast news programs lose touch with government, business, education, and neighbors. They operate without journalists working to keep them informed, uncover truth, expose corruption, and share common goals and experiences;

(d) Quality journalism is key to sustaining civic society, strengthening communal ties, and providing information at a deep level;

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

LBD13206-04-5

(e) Seventy-three percent of United States adults surveyed said they have confidence in their local newspaper. Broadcasting remains a dominant and trusted source of news in communities throughout New York;

(f) Studies show that news content comprises a disproportionate amount of generative artificial intelligence training data. News content is especially valuable to artificial intelligence developers because it is high-quality, professional writing created by human beings;

(g) After training, generative artificial intelligence systems continue to access news websites, podcasts, broadcasts and digital platforms in order gain access to fact-checked, accurate and up to date content to produce outputs;

(h) The vast majority of generative artificial intelligence developers do not obtain permission or compensate news publishers or broadcast news operations for accessing their websites, podcasts, broadcasts and digital platforms for the purposes of building and operationalizing their AI tools and services, in violation of copyright law, those sites' and platforms' terms of service and express prohibitions and preferences;

(i) Maximizing the potential of generative AI requires ensuring the sustainability of journalism and the news industry; and

(j) News publishers and broadcast news operations deserve to know when generative artificial intelligence developers have accessed news websites and used their work.

§ 3. Article 21-A of the general business law is renumbered article 21-B and a new article 21-A is added to read as follows:

ARTICLE 21-A

ARTIFICIAL INTELLIGENCE SOURCE DATA TRANSPARENCY

Section 338. Definitions.
338-a. Artificial intelligence source data transparency.
338-b. Enforcement.
338-c. Applicability.
338-d. Severability.

§ 338. Definitions. The following terms, whenever used or referred to in this article, shall have the following meanings:

1. "Artificial intelligence" 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. "Access" means to obtain, retrieve, acquire, reproduce, crawl, index, or request and receive a transmission of content.

3. "Covered publication" means any print, broadcast, broadcast network or digital publication or service which:

a. performs a public-information function comparable to that traditionally served by journalism organizations, such as newspapers, broadcast news operations, broadcast network news operations, magazines and other periodical publications;

b. invests substantial expenditure of labor, skill, and money to create, edit, produce, and distribute content including by engaging natural persons to create, edit, produce, and distribute original text, audio, photo, illustrative, or video content concerning matters or topics of interest or use to members of the public through activities such as observation, video recording events, interviews, research, testing, and analysis; and

c. publishes new content or updates its content on at least a monthly basis and has a process for error correction and clarification.

4. "Crawler" means software that accesses content from a website or other internet source, such as an online crawler, spider, fetcher, client, bot, user agent or equivalent tool.

5. "Developer" means a person that designs, codes, produces, or substantially modifies an artificial intelligence system or service for use by members of the public. The term "developer" shall not include a journalism provider that uses, develops or obtains artificial intelligence systems solely for internal use.

6. "Generative artificial intelligence" means a class of artificial intelligence models that 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.

7. "Journalism provider" means any person that broadcasts or publishes one or more covered publications.

8. "Person" means a natural person, corporation, trust, estate, partnership, incorporated or unincorporated association or any other legal entity.

9. "Artificial intelligence utilization" means to use digital content as data to develop the capabilities of a generative artificial intelligence system, including through setting or changing its learnable weights and other parameters, and includes, in addition to the initial dataset training, further testing, validating, grounding, or fine tuning by the developer of the artificial intelligence system or service.

§ 338-a. Artificial intelligence source data transparency. 1. a. On or before January first, two thousand twenty-seven 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 first, two thousand twenty-two, is made publicly available to New Yorkers for use, regardless of whether the system or service is made available for a fee, the developer of the system or service shall post on the developer's internet website the following information regarding video, audio, text and data from a covered publication used to train the generative artificial intelligence system or service:

(i) the uniform resource locators or uniform resource identifiers accessed by crawlers deployed by the developer or by third parties on their behalf or from whom they have obtained video, audio, text or data;

(ii) a detailed description of the video, audio, text and data from a covered publication used for artificial intelligence utilization, including the type and provenance of the video, audio, text and data and the means by which it was obtained, sufficient to identify individual works;

(iii) whether any source identifiers, terms, or copyright notices were removed from the video, audio, text or data; and

(iv) the timeframe of data collection.

b. The information required to be posted on a developer's internet website pursuant to paragraph a of this subdivision shall not be required where there is an express written agreement authorizing the developer to access the journalism provider's content and the parties agree not to post information relating to the journalism provider's content on the developer's website.

2. a. On or before January first, two thousand twenty-seven, the developer of a generative artificial intelligence system or service who deploys a crawler, either directly or through a third party, in

connection with such system or service shall disclose information regarding the identity of crawlers used by the developer or by third parties on the developer's behalf in a manner clearly accessible by a website operator, including but not limited to:

(i) the name of the crawler including the crawler's IP address, and specific identifier actually used by the crawler when conducting the crawling activity (such as including the identifiers as part of the user agent or other part of the request headers);

(ii) the legal entity responsible for the crawler;

(iii) the specific purposes for which each crawler is used;

(iv) the legal entities to which operators provide data scraped by the crawlers they operate; and

(v) a single point of contact to enable third parties whose websites are accessed by such crawlers to communicate with the developer and to lodge complaints.

b. The information disclosed pursuant to paragraph a of this subdivision shall be available on an easily accessible platform and updated at the same time as any change is made to such information.

c. The exclusion of a crawler by a website operator shall not negatively impact the findability of the website operator's content in a search engine.

§ 338-b. Enforcement. 1. A journalism provider, or a person authorized to act on a journalism provider's behalf, may bring an action in the supreme court for statutory damages or injunctive relief to compel a developer to comply with section three hundred thirty-eight-a of this article. If the court finds that the developer of a generative artificial intelligence system that is made available to New Yorkers for use is not in compliance, the court shall order such compliance and may impose statutory damages to the journalism provider of up to ten thousand dollars for each violation.

2. Consistent with the New York laws of civil procedure, the supreme court may issue an order for disclosure of copies of, or records sufficient to identify with certainty, the text and data used to train the generative artificial intelligence system or service insofar as such text and data pertains to the journalism provider's internet website, broadcasts, podcasts or other digital platforms, including but not limited to:

a. the uniform resource locators accessed by crawlers deployed by developers or by third parties on their behalf or from whom they have obtained text, video, audio or data, and dates and times of collection; and

b. the text and data used for artificial intelligence utilization, including the type and provenance of the text and data and the means by which such text and data was obtained and when.

§ 338-c. Applicability. The provisions of this article shall not be construed to modify, impair, expand, or in any way alter rights pertaining to Title 17 of the United States Code or the Lanham Act (15 U.S.C. 1051 et seq.).

§ 338-d. Severability. If any provision of this article or the application thereof to any person or circumstances is held to be invalid, such invalidity shall not affect other provisions or applications of this article which can be given effect without the invalid provision or application, and to this end the provisions of this article are severable.

§ 4. This act shall take effect immediately.