SENATE JUDICIARY COMMITTEE
Senator Thomas Umberg, Chair
2023-2024 Regular Session

AB 2013 (Irwin)
Version: June 17, 2024
Hearing Date: June 25, 2024
Fiscal: No
Urgency: No
CK

SUBJECT

Artificial intelligence: training data transparency

DIGEST

This bill requires developers of artificial intelligence (AI) systems or services that are made available for Californians to use to post on their website documentation regarding the data used to train the system or service, including high-level summaries of the datasets used.

EXECUTIVE SUMMARY

Owing to recent advances in processing power and the rise of big data, AI's capacity and the scope of its applications have expanded rapidly, impacting how we communicate, interact, entertain ourselves, travel, transact business, and consume media. It has been used to accelerate productivity, achieve efficiencies, liberate us from drudgery, write our college essay, connect with each other, and live longer, fuller lives. It has also been used to constrain personal autonomy, compromise privacy and security, foment social upheaval, exacerbate inequality, spread misinformation, and subvert democracy. For good or ill, its transformative potential seems boundless.

Ultimately, AI systems are only as good as the data used to train them. However, there is very little transparency in what data is used to train these systems and that lack of transparency hamstrings efforts to address and adequately identify many of the issues being raised by AI's rapid development.

This bill seeks to establish basic transparency requirements for developers of AI systems or services that are made available in California. Developers are required to post documentation regarding the data used to train the AI system or service, including high-level summaries of the datasets used in developing the system or service.

The bill is author-sponsored. It is supported by a variety of organizations, including the California Labor Federation and Transparency Coalition.AI. It is opposed by various industry and business associations, including the California Chamber of Commerce.

PROPOSED CHANGES TO THE LAW

Existing law:

1) Establishes the California Consumer Privacy Act (CCPA), which grants consumers certain rights with regard to their personal information, including enhanced notice, access, and disclosure; the right to deletion; the right to restrict the sale of information; and protection from discrimination for exercising these rights. It places attendant obligations on businesses to respect those rights. (Civ. Code § 1798.100 et seq.)

2) Defines "personal information" as information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. The CCPA provides a nonexclusive series of categories of information deemed to be personal information, including identifiers, biometric information, and geolocation data. (Civ. Code § 1798.140(v).) The CCPA defines and provides additional protections for sensitive personal information, as defined, that reveals specified personal information about consumers. (Civ. Code § 1798.140(ae).)

3) Defines "aggregate consumer information" to mean information that relates to a group or category 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 deidentified. (Civ. Code § 1798.140(b).)

4) Defines "security and integrity" as the ability of:
    a) Networks or information systems to detect security incidents that compromise the availability, authenticity, integrity, and confidentiality of stored or transmitted personal information.
    b) Businesses to detect security incidents, resist malicious, deceptive, fraudulent, or illegal actions and to help prosecute those responsible for those actions.
    c) Businesses to ensure the physical safety of natural persons. (Civ. Code § 1798.140(ac).

5) Establishes the California Privacy Rights Act (CPRA), which amends the CCPA and creates the Privacy Protection Agency (PPA), which is charged with

implementing these privacy laws, promulgating regulations, and carrying out enforcement actions. (Civ. Code § 798.100 et seq.; Proposition 24 (2020).)

6) Permits amendment of the CPRA by a majority vote of each house of the Legislature and the signature of the Governor, provided such amendments are consistent with and further the purpose and intent of this act as set forth therein. (Proposition 24 § 25 (2020).)

This bill:

1) Requires the developer of an AI system or service, on or before January 1, 2026, and before each time thereafter that the system or service is made publicly available to Californians for use, regardless of whether the terms of that use include compensation, to post on the developer's website documentation regarding the data used by the developer to train the AI 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 AI system or service, including, but not limited to:
      i. The sources or owners of the datasets.
      ii. A description of how the datasets further the intended purpose of the system or service.
      iii. The number of data points included in the datasets, which may be in general ranges, and with estimated figures for dynamic datasets.
      iv. A clear definition of each category associated to data points within the datasets, including the format of data points and sample values.
      v. Whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.
      vi. Whether the datasets were purchased or licensed by the developer.
      vii. Whether the datasets include personal information, as defined in subdivision (v) of Section 1798.140.
      viii. Whether the datasets include aggregate consumer information, as defined in subdivision (b) Section 1798.140.
      ix. A description of any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the system or service.
      x. The time period during which the data in the datasets were collected, including a notice if the data collection is ongoing.
      xi. The dates the datasets were first and last used during the development of the system or service.
      xii. Whether the 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.

2) Clarifies that a developer shall not be required to post documentation regarding the data used to train an AI system or service for any of the following:
   a) An AI system or service whose sole purpose is to help ensure security and integrity as defined in subdivision (ac) of Section 1798.140.
   b) An AI system or service whose sole purpose is the operation of an aircraft in the national airspace.
   c) An AI system or service developed for national security, military, or defense purposes that is made available only to a federal entity.

3) Provides that for an AI system or service made available before January 1, 2025, the high-level summary must use information reasonable available to the developer, as provided.

4) Defines the following terms:
   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, or produces an artificial intelligence system or service, or substantially modifies an artificial intelligence system or service for use by a third party for free or for a fee.
   c) "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.
   d) "Train an artificial intelligence system or service" includes testing, validating, or fine tuning by the developer of the AI system or service.

COMMENTS

1. Training data and transparency

As stated, training data is the veritable secret sauce for AI systems. With the race to build bigger and better AI systems and models, a battle over data is also being waged:

> The race to lead A.I. has become a desperate hunt for the digital data needed to advance the technology. To obtain that data, tech companies including OpenAI, Google and Meta have cut corners, ignored corporate policies and debated bending the law, according to an examination by The New York Times.

> At Meta, which owns Facebook and Instagram, managers, lawyers and engineers last year discussed buying the publishing house Simon & Schuster to procure long works, according to recordings of internal meetings obtained by The Times. They also conferred on gathering copyrighted data from across the internet, even if that meant facing lawsuits. Negotiating licenses with publishers, artists, musicians and the news industry would take too long, they said.
>
> Like OpenAI, Google transcribed YouTube videos to harvest text for its A.I. models, five people with knowledge of the company's practices said. That potentially violated the copyrights to the videos, which belong to their creators.[1]

Requiring transparency about the training data used for AI systems helps identify and mitigate biases, addressing hallucinations and other problematic outputs, and shines the light on various other issues, such as privacy and copyright concerns. A team of experts from both industry and academia at the Shorenstein Center created a documentation framework for AI which highlights the importance:

> While often categorized as technical, AI systems and their underlying data and models are sociotechnical. In other words, they combine the technical infrastructure and design with the social context in which they are designed, developed, evaluated, and deployed. Accountability for these systems and their impacts requires transparency around their design and creation and how they are intended to be used. In recent years, alongside the exponential increase in data collection and the efforts to develop increasingly powerful machine learning models, there have been notable efforts calling attention to the need for documentation to accompany datasets, models, and AI systems, and to account for the process of creating them.
>
> Documentation is worthwhile for various stakeholders. It improves the understanding of practitioners creating or building datasets, models, or AI systems, which opens up opportunities to reflect on implicit and explicit decisions, ultimately enhancing the reliability of the systems they create. For organizations, it enables knowledge transfer across silos and encourages responsible use. Further, it provides information to users and potentially affected communities that can be used to determine the appropriateness of an AI system or its underlying data or models, thus helping inform consumer choice, advocacy work, regulation development, and regulation enforcement. It also enables

[1] Cade Metz, et al., *How Tech Giants Cut Corners to Harvest Data for A.I.* (April 6, 2024) The New York Times, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html. All internet citations are current as of June 16, 2024.

> recourse in the event of harms caused by or inquiries into the AI system, and accountability regarding who might be held responsible for those harms.[2]

A recent article in the Harvard Data Science Review highlights the importance of transparency and barriers to achieving it without regulation:

> Knowing what is in the data sets used to train models and how they have been compiled is vitally important. Without this information, the work of developers, researchers, and ethicists to address biases or remove harmful content from the data is hampered. Information about training data is also vital to lawmakers' attempts to assess whether foundation models have ingested personal data or copyrighted material. Further downstream, the intended operators of AI systems and those impacted by their use are far more likely to trust them if they understand how they have been developed.
>
> However, in undertaking their analysis, Schaul et al. (2023) concluded that "many companies do not document the contents of their training data—even internally—for fear of finding personal information about identifiable individuals, copyrighted material and other data grabbed without consent."
>
> In public, companies have used different arguments to justify the lack of transparency around their training data. In documentation published at the launch of its GPT-4 model, OpenAI (2023) stated that it would not share detailed information about 'data set construction' and other aspects of the model's development due to "the competitive landscape and the safety implications of large-scale models." The decision not to disclose the data used to train the model was roundly criticized by a number of leading researchers (Xiang, 2023). A recent op-ed in the Guardian argued that companies are using 'speculative fears' to "stop people asking awkward questions about how this particular technological sausage has been made" (Naughton, 2023).[3]

Various sectors are calling on lawmakers to provide some measure of transparency in this space. A group of media organizations and outlets, including the Associated Press and Gannett, recently issued an open letter calling on regulators to require transparency

[2] Kasia Chmielinski, et al., *The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers* (May 21, 2024) Shorenstein Center on Media, Politics, and Public Policy, https://shorensteincenter.org/wp-content/uploads/2024/05/CleAR_KChmielinski_FINAL.pdf.
[3] Jack Hardinges, et al., *We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models* (December 13, 2023) Harvard Data Science Review, https://hdsr.mitpress.mit.edu/pub/xau9dza3/release/2.

as to the makeup of all training sets used to create AI models.[4] At the federal level, Representative Anna Eshoo and Representative Don Beyer introduced the AI Foundation Model Transparency Act, which would direct the Federal Trade Commission — in consultation with the National Institute of Standards and Technology and the White House Office of Science and Technology Policy — to "establish standards for making publicly available information about the training data and algorithms used in artificial intelligence foundation models."[5]

2. Requiring baseline transparency for AI systems in California

This bill begins to address the issue of training data transparency by requiring developers of AI systems and services made available to Californians for use to post documentation of the data used to train the system or service. This includes a high-level summary of the datasets used, including:

- The sources or owners of the datasets.
- A clear definition of each category associated to data points within the datasets, including the format of data points and sample values.
- Whether the datasets include any data protected by copyright, trademark, or patent, requiring the purchase or licensure of the data, or whether the datasets are entirely in the public domain.
- Whether the datasets were purchased or licensed by the developer.
- Whether the datasets include personal information.

Developers must also disclose whether the system or service used or uses synthetic data generation in its development.

The bill makes clear that training includes testing, validating, and fine tuning by the developer of the AI system or service.

According to the author:

> Artificial Intelligence has become nearly unavoidable in Californians' daily lives, with new exciting generative AI tools being introduced daily, and the companies who make up the cornerstones of our digital lives either adopting AI or identifying their existing tools as falling under the AI umbrella. However consumer confidence in AI systems has not grown at the same rapid pace as industry adoption. Many consumers have valid

[4] Bailey Schulz, *Will AI deepen distrust in news? Gannett, other media organizations want more regulations* (August 9, 2023) USA Today, https://www.usatoday.com/story/tech/news/2023/08/09/ai-regulations-media-gannett/70551555007/.

[5] Edward Graham, *Bill sets transparency standards for AI models, including use of copyrighted material* (January 2, 2024) Nextgov/FCW, https://www.nextgov.com/artificial-intelligence/2024/01/bill-sets-transparency-standards-ai-models-including-use-copyrighted-material/393052/.

> questions about how these AI systems and services are created, and if they truly are better than what they seek to replace.
>
> To build consumer confidence we need to start with the foundations, and for AI that is the selection of training data. AB 2013 provides transparency to consumers of AI systems and services by providing important documentation about the data used to train the services and systems they are being offered, including if synthetic data has or is being used to fill gaps in data sources.
>
> Consumers may use this knowledge to better evaluate if they have confidence in the AI system or service, compare competing systems and services, or put into place mitigation measures to address any shortcomings of the particular system or service.

3. Stakeholder positions

A coalition of industry groups in opposition, including TechNet, writes:

> We note that the bill defines "training" to include testing, validating, or fine tuning by the developer of an AI system or service. We are concerned that amendments expanded the scope of the bill even further and effectively captures all data, regardless of risk level. The bill should be narrowed to scope in only high-risk AI systems. Mandating disclosures for low-risk AI unnecessarily burdens businesses for little to no benefit to the public. A system or service is not a "high-risk" AI system or service, for example, if it is only intended to either perform a narrow procedural task or detect decision-making patterns or deviations from prior decision-making patterns but not meant to replace or influence the previously completed human assessment without proper human review.
>
> For disclosures to be meaningful and not overly burdensome, amendments are also needed to narrow various definitions, starting with the bill's definition of "artificial intelligence system or service". AB 2013's current definition of AI system or service is over broad, arguably capturing regression-based models and even the most rudimentary prediction models or machine-based systems that generate content and make decisions using solely linear functions. Such issues can be addressed by recognizing that the system or service must be capable of "operating with varying levels of autonomy," in line with the OECD AI definition.

Chamber of Progress writes in opposition:

> Requiring online platforms to disclose data used to train their artificial intelligence (AI) systems and services on their website stifles competition in the digital marketplace. A healthy, competitive marketplace is essential to promote quality services for consumers and encourages platforms to innovate. The disclosure requirement risks revealing important business information and strategies, even when platforms specifically note that the datasets are protected intellectual property. Additionally, the language "but not be limited to" in such requirements makes the expectations placed on online platforms unclear.

Writing in support, the California Labor Federation argues:

> Artificial intelligence systems have the capability of producing a wide range of outputs ranging from decisions on whether an individual ought to be hired to evaluating an employee's performance. These decisions are monumentally impactful on the lives of everyday Californians, yet the public is not privy to the data used to train the AI systems affecting their livelihoods. From a public transparency view, this is basic information that the public is entitled to in order to understand whether their personal information may have been used as training data, the potential efficacy of the AI system, and what its outputs are predicated upon. Workers and the public cannot be left in the dark when technology of this magnitude is impacting their jobs and lives.
>
> AB 2013 increases public transparency by requiring the developer of an AI system that makes predictions, recommendations, or decisions to publish a description of the datasets used to train the system. The set of required disclosures includes the source of the dataset, who owns it, definitional categories, when the data was collected and when it has been used, whether the data set was purchased, licensed, or found in the public domain, and whether the collected data is being used to synthesize new data sets. AB 2013 provides the public with the information to address AI systems utilizing nonconsensual personal information and training data riddled with implicit and explicit biases.

SUPPORT

California Democratic Party
California Labor Federation, AFL-CIO
Concept Art Association
Los Angeles County Democratic Party
Oakland Privacy

Perk Advocacy
Santa Monica Democratic Club
Secure Justice
Transparency Coalition.AI

OPPOSITION

American Property Casualty Insurance Association
California Bankers Association
California Chamber of Commerce
California Land Title Association
Chamber of Progress
Computer & Communications Industry Association
Insights Association
National Association of Mutual Insurance Companies
Personal Insurance Federation of California
Software & Information Industry Association
TechNet

RELATED LEGISLATION

Pending Legislation: AB 2877 (Bauer-Kahan, 2024) prohibits CCPA covered-businesses that are the developers of AI systems or tools from using the personal information of consumers under the age of 16 to train AI systems or services without first obtaining affirmative authorization, and even with such authorization the data must be de-identified and aggregated before it is used to train. AB 2877 is currently in this Committee.

AB 3204 (Bauer-Kahan, 2024) requires data digesters to register with the agency, pay a registration fee, and provide specified information, prescribe penalties for a failure to register as required by these provisions, require the California Privacy Protection Agency to create a page on its internet website where this registration information is accessible to the public, and create a fund known as the "Data Digester Registry Fund." AB 3204 was held by the Assembly Appropriations Committee.

Prior Legislation: None known.

PRIOR VOTES:

Assembly Floor (Ayes 56, Noes 8)
Assembly Privacy and Consumer Protection Committee (Ayes 8, Noes 1)
****