Date of Hearing: April 23, 2024

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
AB 3204 (Bauer-Kahan) – As Amended April 18, 2024

SUBJECT: Data Digesters Registration Act

SYNOPSIS

Compared to an Excel spreadsheet, which stores data in neat, organized columns, artificial intelligence stores data in millions to trillions of connections between "neurons" in a "neural network." When AI is trained on a dataset, each piece of data is digested and spread across these connections. It is not currently possible to "untrain" an AI – to force it to forget specific data – just as it is not possible to force a human to unlearn specific knowledge. AI products are black holes in California's information ecosystem, and the process of "training" these tools acts as an event horizon. Once data has crossed this threshold it cannot subsequently be removed.

The California Consumer Privacy Act (CCPA) of 2018 granted Californians the right to know when businesses collect their personal information, to delete it in certain circumstances, to opt out of its sale, to correct it, if inaccurate, and to limit its use and disclosure. The fundamental irreversibility of the AI training process would seem to interfere with a number of these rights. This bill would create a "data digester registry" in the style of the data broker registry, aimed at understanding where, in California's information ecosystem, AI is being trained using personal information.

This bill is author sponsored and supported by a number of privacy organizations, including Oakland Privacy and Electronic Frontier Foundation. The bill is opposed by a coalition of trade associations including the California Chamber of Commerce, California Bankers Association, and Technet.

SUMMARY: Establishes a registry for entities in California that use personal data to train AI, and tasks the California Privacy Protection Agency (Privacy Agency) with administering this registry. Specifically, **this bill**:

1) Establishes the "Data Digester Registry Fund" within the State Treasury, and provides that it will be administered by the Privacy Agency. Requires that all moneys collected be deposited into the fund to offset the costs associated with this bill.

2) Requires a covered entity to register with the Privacy Agency on or before January 31 each year if, in the previous calendar year, the covered entity met the definition of "data digester."

    a) Defines "covered entity" to mean an organization or enterprise, including, but not limited to, a proprietorship, partnership, firm, business trust, joint venture, syndicate, corporation, association, or nonprofit.

    b) Defines "data digester" to mean a covered entity that designs, codes, or produces an artificial intelligence system or service, or that substantially modifies an existing artificial intelligence system or service, by training the system or service on the personal data of 1,000 or more individuals or households.

c) Defines "training" to mean exposing artificial intelligence to data in order to alter the relationship between inputs and outputs.

3) Upon registering with the Privacy Agency, requires a data digester to:

a) Pay a fee, the amount of which is to be determined by the Privacy Agency.

b) Provide the following information:

i) The name of the data digester and its contact information.

ii) Each category of personal information that the data digester has used to train AI, as personal information is categorized in Civ. Code § 1798.140(v) (1).

iii) Each category of sensitive personal information that the data digester has used to train AI, as sensitive personal information in categorized in Civ. Code § 1798.140(ae).

iv) Each category of information related to a consumer's receipt of sensitive services that the data digester has used to train AI, as sensitive services is categorized in Civ. Code § 56.05(s).

v) Whether the data digester has trained AI using the personal data of minors.

vi) Whether and to what extent the data digester is regulated by the federal Fair Credit Reporting Act, the federal Gramm-Leach-Bliley Act, the federal Driver's Privacy Protection Act of 1994, the Insurance Information and Privacy Protection Act, the Confidentiality of Medical Information Act, the privacy, security, and breach notification rules issued by the United States Department of Health and Human Services, or the privacy of pupil records pursuant to Article 5 of Chapter 6.5 of Part 27 of Division 4 of Title 2 of the Education Code.

vii) Any additional information the data digester chooses to provide concerning its AI training practices.

4) Requires the Privacy Agency to provide notice to any data digester it believes has failed to register within 90 days of the deadline, and to post a copy of the notice online.

5) Provides that a data digester that fails to register is liable for administrative fines and costs in an administrative action brought by the Privacy Agency, according to the following schedule:

a) $200/day the data digester fails to register prior to the date the notice is posted.

b) $5000/day the data digester fails to register beginning the 15th day after the notice is posted (105 days after the deadline)

c) Any fees that were due during the period it failed to register.

d) Any expenses incurred by the Privacy Agency in the course of its investigation and administration of the action, as deemed appropriate by the court.

6) Requires a covered entity that transfers an AI system or service, which is capable of being substantially modified through training on personal data, to inform the recipient of the system or service of their responsibilities under this act.

7) Requires the Privacy Agency to create a website where data digester registration information can be accessed by the public.

8) Permits the Privacy Agency to adopt regulations to implement and administer the provisions of this bill.

9) States that the provisions of this bill shall not be construed to supersede or interfere with the operation of the California Consumer Privacy Act (CCPA).

10) Limits the commencement of an administrative action to 5 years from the date on which the violation occurred.

11) States that the provisions of this bill become operative on February 1, 2025.

EXISTING LAW:

1) Provides, pursuant to the California Constitution, that all people are by nature free and independent and have inalienable rights. Among these are the fundamental right to privacy. (Cal. Const. art. I, § 1.)

2) States that the "right to privacy is a personal and fundamental right protected by Section 1 of Article I of the Constitution of California and by the United States Constitution and that all individuals have a right of privacy in information pertaining to them." Further states these findings of the Legislature:

    a) The right to privacy is being threatened by the indiscriminate collection, maintenance, and dissemination of personal information and the lack of effective laws and legal remedies.

    b) The increasing use of computers and other sophisticated information technology has greatly magnified the potential risk to individual privacy that can occur from the maintenance of personal information.

    c) In order to protect the privacy of individuals, it is necessary that the maintenance and dissemination of personal information be subject to strict limits. (Civ. Code § 1798.1.)

3) Establishes the CCPA. (Civ. Code §§ 1798.100-1798.199.100.)

4) Establishes the Privacy Agency and vests it with full administrative power, authority, and jurisdiction to implement and enforce the CCPA. (Civ. Code § 1798.1899.10.)

5) Defines "personal information" to mean 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. States that personal information includes, but is not limited to, the following if it identifies, relates to, describes, is reasonably capable of being associated with, or could be reasonably linked, directly or indirectly, with a particular consumer or household (Civ. Code § 1798.140(v)):

a) Identifiers such as a real name, alias, postal address, unique personal identifier, online identifier, Internet Protocol address, email address, account name, social security number, driver's license number, passport number, or other similar identifiers.

b) Any personal information described in Section 1798.80(e).

c) Characteristics of protected classifications under California or federal law.

d) Commercial information, including records of personal property, products or services purchased, obtained, or considered, or other purchasing or consuming histories or tendencies.

e) Biometric information.

f) Internet or other electronic network activity information, including, but not limited to, browsing history, search history, and information regarding a consumer's interaction with an internet website application, or advertisement.

g) Geolocation data.

h) Audio, electronic, visual, thermal, olfactory, or similar information.

i) Professional or employment-related information.

j) Education information, defined as information that is not publicly available personally identifiable information as defined in the Family Educational Rights and Privacy Act (20 U.S.C. Sec. 1232g; 34 C.F.R. Part 99).

k) Inferences drawn from any of the information identified in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes.

l) Sensitive personal information.

6) Defines biometric information to mean an individual's physiological, biological, or behavioral characteristics, including information pertaining to an individual's deoxyribonucleic acid (DNA), that is used or is intended to be used singly or in combination with each other or with other identifying data, to establish individual identity. (Civ. Code § 1798.140(c).)

7) Further defines "personal information" to include any information that identifies, relates to, describes, or is capable of being associated with, a particular individual, including, but not limited to, his or her name, signature, social security number, physical characteristics or description, address, telephone number, passport number, driver's license or state identification card number, insurance policy number, education, employment, employment history, bank account number, credit card number, debit card number, or any other financial information, medical information, or health insurance information. (Civ. Code § 1798.80(e).)

a) States that personal information does not include publicly available information that is lawfully made available to the general public from federal, state, or local government records.

8) Defines sensitive personal information to mean any of the following:

a) Personal information that reveals:

i) A consumer's social security, driver's license, state identification card, or passport number.

ii) A consumer's account log-in, financial account, debit card, or credit card number in combination with any required security or access code, password, or credentials allowing access to an account.

iii) A consumer's precise geolocation.

iv) A consumer's racial or ethnic origin, citizenship or immigration status, religious or philosophical beliefs, or union membership.

v) The contents of a consumer's mail, email, and text messages unless the business is the intended recipient of the communication.

vi) A consumer's genetic data.

b) The processing of biometric information for the purpose of uniquely identifying a consumer.

c) Personal information collected and analyzed concerning a consumer's health.

d) Personal information collected and analyzed concerning a consumer's sex life or sexual orientation. (Civ. Code § 1798.140(ae).)

9) Defines "sensitive services" to mean all health care services related to mental or behavioral health, sexual and reproductive health, sexually transmitted infections, substance use disorder, gender affirming care, and intimate partner violence, and includes services described in Sections 6924, 6925, 6926, 6927, 6928, 6929, and 6930 of the Family Code, and Sections 121020 and 124260 of the Health and Safety Code, obtained by a patient at or above the minimum age specified for consenting to the service specified in the section. (Civ. Code § 56.05 (s).)

10) Establishes the Data Brokers' Registry Fund with the State Treasury and empowers the Privacy Agency to administer it. (Civ. Code § 1798.99.81.)

11) Requires a business that meets the definition of "data broker" to register with the CPPA and provide specified information. (Civ. Code § 1798.99.82.)

12) Limits a business' collection, use, retention, and sharing of a consumer's personal information to that which is reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed, or for another disclosed purpose that is compatible with the context in which the personal information was collected,

and not further processed in a manner that is incompatible with those purposes. (Civ. Code § 1798.100(c).)

13) Prohibits a business from selling or sharing the personal information of a child that is 16 years of age or younger, if the business has actual knowledge of the child's age, unless the child, or the child's parent or guardian in the case of children less than 13 years old, has affirmatively authorized the sharing or selling of the personal information. (Civ. Code § 1798.120(c).)

14) Provides that consumers have the right, at any time, to direct a business that collects sensitive personal information about the consumer to restrict the use of that information to only that use which is necessary to perform the services or provide the goods reasonably expected by an average consumer who requests those goods or services. (Civ. Code § 1798.121(a).)

FISCAL EFFECT: As currently in print, this bill is keyed fiscal.

COMMENTS:

1) **Artificial intelligence.** The development of AI is creating exciting opportunities to grow California's economy and improve the lives of its residents. AI can generate compelling text and convincing images in an instant. It can automate painstaking tasks, identify subtle patterns in large datasets, and make accurate predictions in the face of incomplete information. Just as embracing the internet ushered in an era of commercial dominance for California nearly thirty years ago, AI could deliver a second technological golden age to California. But with novel technologies come novel safety concerns. The present bill is predicated on the idea that certain types of information, when used to train AI, are inherently risky. To understand why, it is worth briefly exploring how these systems work.

2) **The importance of training.** AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. This is because AI is not fundamentally different from other computer functions. Its novelty lies in its application: unlike normal computer functions, AI is able to accomplish tasks that are normally performed by humans.

Training is the secret sauce of machine learning; it is the principle innovation that allows modern AI to be both efficient and versatile. During training, a naïve AI is exposed to data and allowed to automatically explore its structure. As the AI explores, it alters itself in an attempt to better represent the data. Each piece of data affects every part of an AI. In a sense, AI "digest" and integrate the data they train on in order to learn, just as humans digest and integrate the foods we eat in order to grow.

AI that are trained on small, specific datasets in order to make recommendations and predictions are sometimes called "predictive AI." This differentiates them from "generative AI," which are trained on massive datasets in order to produce detailed text and images. When Netflix suggests a TV show to a viewer, the recommendation is produced by predictive AI that has been trained on the viewing habits of Netflix users. When ChatGPT generates text in clear, concise paragraphs, it uses generative AI that has been trained on the written contents of the internet.

3) **Haphazard training data.** There is a common saying in computer science: "garbage in, garbage out." The performance of an AI product is directly impacted by the quality, quantity, and

relevance of the data used to train it. Before training, datasets are often categorized to make them easier for AI to work with. Rigorously categorizing the data in a dataset becomes more difficult as the dataset becomes larger, but failing to organize its contents can lead to meaningless, false, or harmful outputs.

The biggest names in AI – OpenAI, Meta, and Google – understand AI's critical need for data better than anyone else. According to a recent New York Times examination, the race to lead in the AI space has become a desperate hunt for digital data. To obtain that data, these tech companies have cut corners, ignored corporate policies and debated bending the law:[1]

> 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.
>
> Last year, Google also broadened its terms of service. One motivation for the change, according to members of the company's privacy team and an internal message viewed by The Times, was to allow Google to be able to tap publicly available Google Docs, restaurant reviews on Google Maps and other online material for more of its A.I. products.

Meta and Google are privy to some of the most sensitive information in the world. In many developing countries, Facebook effectively is the internet.[2] A tremendous number of Californians use Google, or Google Chrome, or Google Drive, or Google Cloud, or Gmail. Amazon has yet to enter the generative AI thunderdome – but when it does, it can sleep peacefully at night knowing a third of the world's cloud computing market is powered by Amazon Web Services.[3]

In their race to obtain vast quantities of training data, major AI developers have not hesitated to move fast and break things. The Stanford Internet Observatory recently discovered that a common image training dataset known as LAION-5B contains many instances of child sexual abuse materials. Their study identified 3226 dataset entries of suspected child pornography, much of which was later confirmed as such by third parties.[4] This dataset was built by automatically scraping the internet, and images containing child pornography were found to have

---

[1] Cade Metz, Cecilia Kang, Sheera Frenkel, Stuart A. Thompson and Nico Grant, "How Tech Giants Cut Corners to Harvarst Data for A.I.," *New York Times*, Apr. 6, 2024, https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

[2] Nesrine Malik, "How Facebook took over the internet in Africa – and changes everything," *Guardian*, Jan. 20, 2022, https://www.theguardian.com/technology/2022/jan/20/facebook-second-life-the-unstoppable-rise-of-the-tech-company-in-africa.

[3] Aditya Rayaprolu, "How Many Websites Run on AWS? Useful AWS Statistics for 2024," *techjury*, Jan. 2, 2024, https://techjury.net/blog/how-many-websites-run-on-aws/.

[4] David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models," *Stanford Internet Observatory*, Dec. 23, 2023.

originated from large, well-known websites such as Reddit, Twitter, Blogspot, and Wordpress, as well as mainstream adult sites such as XHamster and XVideos.

4) **An AI never forgets.** Just as humans cannot intentionally forget information they have learned, it is not currently possible to remove data from a trained AI.[5] Unlike an Excel spreadsheet, which stores data in neat columns, AI stores data in the connections between "neurons" in a "neural network." Every one of these connections is influenced by every piece of training data, and a large model like ChatGPT-4 is reported to have more than 1.7 trillion connections.[6] It is not possible to specifically alter these connections in order to remove data without fundamentally changing the model; as a result, for data to be removed, the model must be retrained from scratch. ChatGPT-4 is estimated to have taken 4-7 months to train in the first place.[7]

What happens when an AI is trained on extremely sensitive information – for example, an individual's DNA sequence, or their social security number, or their intimate photos, or their immigration status? The same thing that happens when an AI is trained on any other type of information: the AI digests it, and then retains it forever. AI are fundamentally different from other forms of data storage. They are black holes in the information ecosystem, with "training" as their event horizons. Once data has crossed this threshold it cannot be removed.

5) **The California Consumer Privacy Act and the California Privacy Rights Act (CPRA).** In 2018, the Legislature enacted the CCPA (AB 375 (Chau, Chap. 55, Stats. 2018)), which gives consumers certain rights regarding their personal information, such as the right to: (1) know what personal information about them is collected and sold; (2) request the categories and specific pieces of personal information the business collects about them; and (3) opt out of the sale of their personal information, or opt in, in the case of minors under 16 years of age.

Subsequently, in 2020, California voters passed Proposition 24, the California Privacy Rights Act (CPRA), which established additional privacy rights for Californians. With the passage of the CCPA and the CPRA, California now has the most comprehensive laws in the country when it comes to protecting consumers' rights to privacy.

In addition, Proposition 24 created the California Privacy Protection Agency in California, vested with full administrative power, authority, and jurisdiction to implement and enforce the CCPA and the CPRA. The Privacy Agency's responsibilities include updating existing regulations, and adopting new regulations.

6) **What this bill would do.** This bill would require businesses and other entities who train AI using personal data to register with the Privacy Agency. In registering, those "data digesters" would be required to pay a fee and disclose which categories of personal data they use to train AI, as well as to disclose whether the data they use to train AI is covered under a number of

---

[5] Stephen Pastis, "A.I.'s un-learning problem: Researchers say it's virtually impossible to make an A.I. model 'forget' the things it learns from private user data," *Yahoo! Finance*, Aug. 30, 2023, finance.yahoo.com/news/un-learning-problem-researchers-virtually-164342971.html.
[6] Reed Albergotti, "Microsoft pushes the boundaries of small AI models with big breakthrough," *SEMAFOR*, Nov. 1, 2023, www.semafor.com/article/11/01/2023/microsoft-pushes-the-boundaries-of-small-ai-models.
[7] Stephen McAleese, "Retrospective on 'GPT-4 Predictions' After the Release of GPT-4," *LESSWRONG*, Mar. 17, 2023, https://www.lesswrong.com/posts/iQx2eeHKLwgBYdWPZ/retrospective-on-gpt-4-predictions-after-the-release-of-gpt.

existing privacy laws. The bill does not require digesters to disclose categories of training data for individual products, nor does it require them to cease using personal data to train AI. It also does not require digesters to disclose the data itself – only which categories the data belong to. Entities that train AI but that do not use personal data are not within the scope of this bill.

7) **Author's statement.** According to the author:

> Artificial intelligence (AI) is an exciting technology with an enormous amount of potential, but there are certain risks associated with its use. One of these is the simple fact that AI systems and services cannot be "untrained." Data that is fed into an AI product during training will forever be a part of that product. This inability to forget is fundamental to the technology, but raises privacy concerns when sensitive personal information is used to train AI. Sequences of DNA, social security numbers, intimate imagery, and immigration status – there are certain types of information that are inherently riskier than others. Californians deserve to know if information of this type is being used in an AI product during training.
>
> AB 3204 creates a registry for businesses and other entities that train AI on personal data. In registering, these "data digesters" would pay a reasonable fee and disclose which types of personal data they have used to train AI. This bill would allow Californians to understand where, across the information ecosystem, their personal data is being utilized to train AI models.

8) **Analysis.** This bill represents a crucial first step towards understanding where, when, and why the personal data of Californians is being used to train AI. The bill is effectively an information-gathering exercise: over time, the data digester registry that this bill creates will increasingly reveal where personal information is falling into AI "black holes" across California's information landscape.

In the coalition of industry associations position letter, the letter begins by comparing the current bill to the Privacy Agency's data broker registry:

> As a general matter, this bill appears modeled off the data broker registry which enables consumers to effectuate their CCPA rights against data brokers. That registry was specifically implemented after the passage of the CCPA, to address a gap in consumer awareness as to the identity of data brokers that might be in possession of their [personal information.] (AB 1202, Chau (Chapter 753, Statutes of 2019)). Creating a central repository was necessary for consumers to identify and initiate requests under the CCPA with third parties with which they do not directly interact. In direct contrast, AB 3204 would now create a central repository of businesses that train AI using [personal information], despite there not being any circumstances comparable to the ones that necessitated the development of the data broker registry that might warrant this registry…

The current bill is, in fact, modeled off the data broker registry. However, it is incorrect to assert that there are no "circumstances comparable to the ones that necessitated the development of the data broker registry…" In the coalition's own words, the data broker registry was created to "address a gap in consumer awareness as to the identity of data brokers that might be in possession of their personal information." The same is true with respect to data digesters: there is currently a gap in consumer awareness with respect to the identity of businesses using personal information to train artificial intelligence.

The coalition letter continues:

> First, the scope of this bill is incredibly broad given the definitions (or lack thereof) referenced, starting with the definition of "data digesters." By requiring any business that trains AI using [personal information] to register as "data digestors," AB 3204 may as well require all CCPA-covered businesses to register, given the breadth of the CCPA's definitions and lack of clarity around what is considered "train[ing] AI." Realistically, this means hundreds of thousands of businesses, as "business" under the CCPA, includes any business that meets one of the CCPA thresholds and that does business in the State of California—not just large California based businesses or large AI developers.

Since this letter was submitted, AB 3204 has been amended to narrow its scope. In the version of the bill in print, a data digester is an entity who "designs, codes, produces, or substantially modifies an artificial intelligence system or service." As a result, the bill mainly targets "developers" of AI tools, rather than the far greater number of California businesses who "deploy" these tools.

The coalition letter continues:

> "…despite relying on the incredibly broad definitions found in the CCPA and being modeled upon the data broker registry, AB 3204 fails to include any of the reasonable, but necessary exemptions that are included in either of those laws."

This is true: as currently written, AB 3204 does not adopt the CCPA's exemptions. Such exemptions include publicly available information, lawfully obtained, truthful information that is a matter of public concern, or consumer information that is deidentified or aggregate consumer information. (Civ. Code § 1798.140(v).) However, unlike the CCPA, which imposes certain affirmative duties on covered businesses, this bill simply requires covered entities to register and describe the categories of information contained with their datasets.

The coalition letter continues:

> Second, companies that are required to register are not simply asked to provide identifying details about high-risk and low-risk AI alike; for each of those models, they are required to identify each category of [personal information], as that term is defined under the CCPA, that the data digester uses to train AI, identified by reference to each applicable subparagraph within that definition of [personal information]—of which there are twelve. They also must do the same when it comes to sensitive [personal information] (SPI), which adds another half dozen categories.

This is true: AB 3204 is premised on the notion that it is important for consumers to know which particular categories of personal data are used to train AI, rather than simply chunking this information as "personal information, sensitive personal information, and personal information related to sensitive services." However, requiring this level of disclosure should not be a huge lift for the data digesters in question. It is hard to imagine the disclosure form these data digesters submit will be longer than a single page. The ~20 categories of information outlined in this bill will each be listed next to a yes/no checkbox. Companies will check boxes for categories of information they use to train AI, and leave the rest blank.

The coalition letter continues:

> Any registered business (not just those subject to CMIA) must then also identify each category of information related to consumers' receipt of "sensitive services", as defined under the Confidentiality of Medical Information Act (CMIA, defining sensitive services as those related to mental or behavioral health, reproductive health, gender affirming care, and a host of other services enumerated under nine other provisions of law), that the data digester uses to train AI, identified by reference to the specific category of sensitive service enumerated in the definition. While likely not the intent, this effectively forces hundreds of thousands of businesses to infringe on the privacy of Californians to provide disclosures with the level of detail demanded by AB 3204.

The "level of detail demanded by AB 3204" is, as described above, a checkbox. It is unclear why this level of disclosure would infringe upon the privacy of Californians. Quite the opposite: requiring data digesters to disclose their use of sensitive information to train AI will allow Californians to make more informed decisions when disseminating their sensitive personal information.

The coalition letter continues:

> Further exacerbating all these issues, is the fact that "[personal information]" under the CCPA, is any "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." While deidentified or aggregated consumer data are exempted from the definition, it still captures information that on its own may not be identifiable, but that when pieced together with other pieces of information, becomes identifiable. This is beyond burdensome. It is incredibly impractical, privacy invasive, and at times completely impossible (certainly, not without violations of many other privacy laws). Imagine businesses having to dedicate employees to determine if they could feasibly trace back each individual piece of information used to train AI to a particular individual, and then also review medical records of those individuals to identify if they were provided sensitive services – even if the business has nothing to do with health care.

Existing law requires data brokers to disclose whether they collect the personal information of minors (Civ. Code 1798.99.82.) This requirement references the same definition of "personal information" as AB 3204. Thus, the issue raised by the industry coalition here should also apply to registered data brokers. As of the time of this writing, the Data Broker Registry lists 456 registered businesses. It can be assumed that these businesses did not find the requirement outlined here to be "beyond burdensome," as they were ultimately able to successfully register.

The coalition letter continues:

> Finally, in addition to imposing significant penalties, fines, fees, and expenses that are problematic particularly for smaller businesses, AB 3204 bill fails to provide any protections or otherwise address copyright and data ownership issues, trade secrets or patents for the information that businesses are required to divulge and that will be made available by the Privacy Agency on a public website. Requiring this level of granular data about each category of PI and SPI used, will invariably force businesses to divulge trade secrets and other highly confidential or patented information, helping their competitors to their own detriment. Again, because this impacts not only businesses developing AI in California, but those doing business in California, should this bill become law, it is highly unlikely that

businesses would risk rolling out certain patented tools and algorithms in California, altogether.

The data broker registry, which serves as a model for the proposed legislation, currently sets its annual registration fee at $400. All other fines, fees, and penalties described in this bill exist downstream of a data digester failing to register. With respect to "trade secrets," this term is defined several times throughout California code:

Health & Saf. Code § 25173: "Trade secrets," as used in this section, may include, but are not limited to, any formula, plan, pattern, process, tool, mechanism, compound, procedure, production data, or compilation of information which is not patented, which is known only to certain individuals within a commercial concern who are using it to fabricate, produce, or compound an article of trade or a service having commercial value, and which gives its user an opportunity to obtain a business advantage over competitors who do not know or use it.

Civ. Code § 3426.1: "Trade secret" means information, including a formula, pattern, compilation, program, device, method, technique, or process, that:

(1) Derives independent economic value, actual or potential, from not being generally known to the public or to other persons who can obtain economic value from its disclosure or use; and

(2) Is the subject of efforts that are reasonable under the circumstances to maintain its secrecy.

Gov. Code § 7924.510: As used in this section, "trade secret" may include, but is not limited to, any formula, plan, pattern, process, tool, mechanism, compound, procedure, production data, or compilation of information that satisfies all of the following requirements:

(1) It is not patented.

(2) It is known only to certain individuals within a commercial concern who are using it to fabricate, produce, or compound an article of trade or a service having commercial value.

(3) It gives its user an opportunity to obtain a business advantage over competitors who do not know or use it.

Pen. Code § 499c: "Trade secret" means information, including a formula, pattern, compilation, program, device, method, technique, or process, that:

(A) Derives independent economic value, actual or potential, from not being generally known to the public or to other persons who can obtain economic value from its disclosure or use; and

(B) Is the subject of efforts that are reasonable under the circumstances to maintain its secrecy.

It is not clear that any of these definitions would consider a yes/no checkbox on a disclosure form for certain categories of personal data to be a "trade secret," especially considering the required disclosure is on a company-wide basis rather than a product-by-product basis. Were

OpenAI to submit a disclosure form indicating they use biometric data to train AI, this bill would not require them to reveal whether that data goes into training DALL-E, or ChatGPT, or any other product.

In summary, this bill would impose modest requirements on data digesters in exchange for providing the Legislature with a fuller understanding of California's AI information ecosystem. The information yielded by the data digester registry will provide a necessary foundation for future legislation, allowing the Legislature to make informed judgments about whether, and how, to further regulate this space.

9) **Related legislation.** AB 2013 (Irwin, 2024) would require a developer of an AI system or service to publicly disclose a description of the training dataset used in the development of the system or service. This bill is currently pending in this Committee.

ARGUMENTS IN SUPPORT:

Oakland Privacy writes:

> A registry, in and of itself, is a limited piece of regulation that cannot do everything to prevent Californians from abuse and mis-use by AI technologies. But in order to fully grasp the full range of companies and nonprofits working in the space, and the scope of the products they are developing and deploying, a registry can be an immensely useful snapshot of the current state of play.

Electronic Frontier Foundation writes:

> A.B. 3204 takes a good first step by increasing the transparency that consumers so often lack when interacting with businesses that use personal information to train artificial intelligence. California has already broken ground as a privacy leader. It is encouraging to see these concepts applied to the unique consumer privacy challenges that artificial intelligence raises. For instance, many consumers are asking, "how can we keep companies from using data about us in their AI models?" Focusing on transparency around data used for training lays a smart cornerstone upon which we hope the state will be able to build.

Transparency Coalition.AI writes:

> The National Institute of Standards and Technology (NIST) defines transparency, in part, as a "process to imply openness and accountability." At its core, AB 3204 will do just this by utilizing existing oversight and enforcement systems to ensure openness and accountability in how personal information is used, bought, and sold in AI systems. This is of particular importance now, as reports increasingly uncover breaches in personal protections as the biggest companies race to take all.
>
> As revealed by an investigative article in the New York Times, *How Tech Giants Cut Corners to Harvest Data for AI,* "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."

> We cannot stress enough the importance of moving in support of this bill now. Our personal information, and that of our children, is fair game without these protections.

ARGUMENTS IN OPPOSITION:

California Chamber of Commerce writes on behalf of a coalition of trade associations:

> Ultimately, it is unclear to us the benefit of creating this registry and requiring such an incredibly cumbersome and privacy-invasive registration process. There are much more reasonable mechanisms that can also ensure that AI developers inform deployers and consumers about the types of AI used to train their consumer-facing AI products. Even if all the legal risks and practical issues of this data digester registry could be addressed, making it possible to comply with bill's disclosures, to what end? The situation addressed by AB 3204 is not comparable to the data broker registry which was necessary to bridge a gap in consumer awareness that effectively precluded them from effectuating their CCPA rights with certain businesses. In contrast, this registry is not a necessary element to transparency.

REGISTERED SUPPORT / OPPOSITION:

Support

Electronic Frontier Foundation
Oakland Privacy
Secure Justice
Transparency Coalition.ai

Opposition

American Council of Life Insurers
Association of California Life and Health Insurance Companies
Association of National Advertisers
California Bankers Association
California Chamber of Commerce
California Credit Union League
California Retailers Association
Computer and Communications Industry Association
Insights Association
Internet Coalition
Los Angeles Area Chamber of Commerce
Software & Information Industry Association
Technet

Analysis Prepared by: Slater Sharp / P. & C.P. / (916) 319-2200