Date of Hearing: July 2, 2024

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
SB 893 (Padilla) – As Amended June 21, 2024

SENATE VOTE: 37-0

SUBJECT: California Artificial Intelligence Research Hub

SYNOPSIS

The recent explosive growth of the artificial intelligence (AI) industry in California is powered by access to information. Modern generative artificial intelligence (GenAI) systems are trained using nearly all text data that humanity has thus far produced, and the industries at the heart of the AI hype train are hungry for more. California's state government serves as a central repository for information related to the state of California – all 156,000 square miles of it. In the right hands, this information represents an untapped goldmine.

This bill would create a mechanism for government data to be released to external researchers. In doing so, however, this bill creates significant privacy concerns. More than 39 million people currently live in California. Reproductive healthcare, immigration, welfare, and taxes – in order to serve its residents effectively, California frequently collects and retains various types of sensitive information from its residents. In return, Californians rely on government to not take advantage of or reveal these data. The Information Practices Act of 1977 (IPA), modeled after the Federal Privacy Act of 1974, codifies this concept: the IPA is the primary privacy statute governing the collection, maintenance, and disclosure of personal information by California state agencies. In its current form, this bill serves as a blanket exemption to the IPA.

Committee amendments narrow the scope of this bill by limiting the recipients of government data under this bill to researchers at academic institutions. Committee amendments also augment the bill's privacy protections by prohibiting the release of personal information and requiring the California Consumer Privacy Agency to review and approve data prior to its release.

This bill is author sponsored. It is supported by the University of California and Stanford University, as well as a variety of industry associations including the California Chamber of Commerce, TechNet, and the Computer & Communications Industry Association.

Committee amendments, set forth below, narrow the bill's focus and augment its privacy protections.

SUMMARY: Creates the California Artificial Intelligence Research Hub in the Government Operations Agency and tasks it with increasing access to government data and supporting the research into artificial intelligence. Specifically, **this bill**:

1) Requires the Government Operations Agency, the Governor's Office of Business and Economic Development, the California Privacy Protection Agency, and the Department of

Technology to collaborate to establish the California Artificial Intelligence Research Hub in the Government Operations Agency.

a) Permits the Government Operations Agency to collaborate with additional state agencies to establish the hub, as needed.

b) Declares that the hub shall serve as a centralized entity to facilitate collaboration between government agencies, academic institutions, and private sector partners to advance artificial intelligence research and development that seeks to harness the technology's full potential for public benefit while safeguarding privacy, advancing security, and addressing risks and potential harms to society.

c) Requires the Government Operations Agency, the Governor's Office of Business and Economic Development, the California Privacy Protection Agency, and the Department of Technology to consult with academic institutions within the state in establishing the hub.

2) Requires the California Artificial Intelligence Research Hub to do all of the following:

a) Increase lawful access to government data while protecting privacy and safeguarding access to data by developing a streamlined process for researchers at academic institutions to access data collected by state agencies.

i) Exempts trade secrets, as defined in Civil Code section 3426.1.

b) Support access to and development of artificial intelligence computing capacity and technology by building out public computing infrastructure, facilitating access to existing commercial computing infrastructure, or finding ways to reduce costs and other economic barriers research institutions may face in accessing computing infrastructure.

c) Spur innovation in artificial intelligence applications for the benefit of the public.

d) Ensure the development of trustworthy artificial intelligence technologies with a focus on transparency, fairness, and accountability.

e) Provide researchers with increased access to data and computing resources, education, and training opportunities in furtherance of applications of artificial intelligence for benefit to the people of California.

EXISTING LAW:

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

2) Establishes the Information Practices Act (IPA) of 1977, which generally enumerates the requirements applicable to state agencies that collect, maintain, and disclose personal information from California residents, including limitations on permissible disclosure, the rights of residents to know and access the information, and required accounting of disclosures of the information. (Civ. Code § 1798, et seq.)

3) States, in the IPA, 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.)

4) Requires that each state agency maintain in its records only personal information that is relevant and necessary to accomplish the purpose of the agency. (Civ. Code § 1798.14.)

5) Requires that each agency collect personal information to the greatest extent practicable directly from the individual who is the subject of the information rather than from another source. (Civ. Code § 1798.15.)

6) Prohibits an individual's name and address from being distributed for commercial purposes, sold, or rented by an agency unless such action is specifically authorized by law. (Civ. Code § 1798.60.)

7) Defines "personal information," for purposes of the IPA, as any information that is maintained by an agency that identifies or describes an individual, including, but not limited to, the individual's name, social security number, physical description, home address, home telephone number, education, financial matters, and medical or employment history. (Civ. Code § 1798.3(a).)

8) Defines "agency", for the purposes of the IPA, to mean every state office, officer, department, division, bureau, board, commission, or other state agency, except for the California Legislature, agencies within the judicial branch, the State Compensation Insurance Fund, and local agencies, defined to include: counties; cities, whether general law or chartered; cities and counties; school districts; municipal corporations; districts; political subdivisions; or any board, commission, or agency thereof; other local public agencies, or entities that are legislative bodies of a local agency as specified. (Civ. Code § 1798.3(b); Gov. Code § 6252(a).)

9) Requires each agency to keep an accurate accounting of the date, nature, and purpose of each disclosure of a record made pursuant to specified circumstances; and requires each agency to retain that accounting for at least three years after the disclosure, or until the record is destroyed, whichever is shorter. (Civ. Code §§ 1798.25 & 1798.27.)

10) Except as specified, endows each individual with the following rights: to inquire and be notified as to whether the agency maintains a record about them; to inspect all personal

information in any record maintained by reference to an identifying particular of the individual; and to submit a request in writing to amend a record containing personal information pertaining to them maintained by an agency. (Civ. Code § 1798.30, et seq.)

11) Requires each state agency, when it provides by contract for the operation or maintenance of records containing personal information to accomplish an agency function, to cause, consistent with its authority, the requirements of the IPA to be applied to those records; and specifies that for purposes of enforcing penalties for violations of the IPA, any contractor and any employee of the contractor, shall be considered to be an employee of an agency. (Civ. Code § 1798.19.)

12) Establishes the Department of Technology within the Government Operations Agency, under the supervision of the Director of Technology. (Gov. Code § 11545(a).)

13) Requires, upon appropriation by the Legislature, the Secretary of the Government Operations Agency to evaluate the following:

    a) The impact of the proliferation of deepfakes on state government, California-based businesses, and residents of the state.

    b) The risks, including privacy risks, associated with the deployment of digital content forgery technologies and deepfakes on state and local government, California-based businesses, and residents of the state.

    c) Potential privacy impacts of technologies allowing public verification of digital content provenance.

    d) The impact of digital content forgery technologies and deepfakes on civic engagement, including voters.

    e) The legal implications associated with the use of digital content forgery technologies, deepfakes, and technologies allowing public verification of digital content provenance.

    f) The best practices for preventing digital content forgery and deepfake technology to benefit the state, California-based businesses, and California residents, including exploring whether and how the adoption of a digital content provenance standard could assist with reducing the proliferation of digital content forgeries and deepfakes. (Gov. Code § 11547.5(b).)

14) Requires the Secretary of the Government Operations Agency to develop a coordinated plan to accomplish all of the following:

    a) Investigate the feasibility of, and obstacles to, developing standards and technologies for state departments for determining digital content provenance.

    b) Increase the ability of internet companies, journalists, watchdog organizations, other relevant entities, and members of the public to meaningfully scrutinize and identify digital content forgeries and relay trust and information about digital content provenance to content consumers.

c) Develop or identify mechanisms for content creators to cryptographically certify authenticity of original media and nondeceptive manipulations.

d) Develop or identify mechanisms for content creators to enable the public to validate the authenticity of original media and nondeceptive manipulations to establish digital content provenance without materially compromising personal privacy or civil liberties. (Gov. Code § 11547.5(c).)

15) Expresses the intent of the Legislature that policies and procedures developed by the Department of Technology and Department of General Services pertaining to the acquisition of information technology (IT) goods and services provide for all of the following: the expeditious and value-effective acquisition of IT goods and services to satisfy state requirements; the acquisition of IT goods and services within a competitive framework; the delegation of authority by the Department of General Services to each state agency that has demonstrated to the Department of General Services' satisfaction the ability to conduct value-effective IT goods and services acquisitions; and the review and resolution of protests submitted by any bidders with respect to any IT goods and services acquisitions. (Pub. Con. Code § 12101.)

16) Requires the Department of Technology, on or before September 1, 2024, to conduct, in coordination with other interagency bodies as it deems appropriate, a comprehensive inventory of all high-risk automated decisionmaking tools that have been proposed for use, development, or procurement by, or are being used, developed, or procured by, any state agency. (Gov. Code § 11546.45.5(b).)

17) Requires the Department of Technology, on or before January 1, 2025, and annually thereafter, to submit a report, as specified, of the comprehensive inventory to the Assembly Committee on Privacy and Consumer Protection and the Senate Committee on Governmental Organization. This requirement expires on January 1, 2029. (Gov. Code § 11546.45.5(d).)

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

COMMENTS:

1) AI and GenAI. The development of GenAI is creating exciting opportunities to grow California's economy and improve the lives of its residents. GenAI can generate compelling text, images and audio in an instant – but with novel technologies come novel safety concerns.

In brief, AI is the mimicking of human intelligence by artificial systems such as computers. 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. 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.

AI that are trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from GenAI, 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 GenAI that has been trained on the written contents of the internet.

2) The importance of 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:

> 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. A tremendous number of Californians use Google, or Google Chrome, or Google Drive, or Google Cloud, or Gmail. 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. This dataset was built by automatically scraping the internet, and images containing child pornography were found to have originated from large, well-known websites such as Reddit, Twitter, Blogspot, and Wordpress, as well as mainstream adult sites such as XHamster and XVideos.

3) 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. 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. 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.

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.

4) The California Consumer Privacy Act (CCPA) 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 (Privacy 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.

5) The Information Practices Act. The Information Practices Act of 1977 (IPA; Civ. Code § 1798, et seq.), modeled after the Federal Privacy Act of 1974, is the primary privacy statute governing the collection, maintenance, and disclosure of personal information by California state agencies. Along with the substantive provisions of the IPA, the Legislature codified findings and declarations upon its passage justifying the need for the consistent limits on the maintenance and dissemination of personal information by government agencies.

Generally, the IPA places several conditions and restrictions on the collection, maintenance, and disclosure of the personal information of Californians held by state agencies, including a prohibition on the disclosure of an individual's personal information without the individual's consent except in specified circumstances. In addition, the IPA requires that along with any form requesting personal information from an individual, an agency provide notice of information pertaining to the individual's rights with respect to their personal information, the purposes for which the personal information will be used, and any foreseeable disclosures of that personal information.

The IPA also provides individuals with certain rights to be informed of what personal information an agency holds relating to that individual; to access and inspect that personal information; and to request corrections to that personal information, subject to specified exceptions. Finally, when state agencies contract with private entities for services, the contractors are typically governed by the IPA, with few additional privacy protections generally stipulated in the contracts themselves.

6) What this bill would do. This bill would require the Government Operations Agency, in collaboration with the Governor's Office of Business and Economic Development and the California Privacy Protection Agency, to collaborate to create the California Artificial Intelligence Research Hub in the Government Operations Agency. The California Artificial Intelligence Research Hub would be tasked with increasing lawful access to government data, excluding trade secrets, and with generally supporting the development of infrastructure related to artificial intelligence.

7) Author's statement:

> California is a global leader in technological advancement. Much of that leadership has been driven by our world-class higher education systems. Emerging AI technologies are costly and energy intensive, and require broad-based coordination among institutions and other sectors. Shared resources will be vital to the continued development of AI technology in California. The creation of the California Artificial Intelligence Research Hub allows us to pool and leverage the state's financial resources and the intellectual firepower of our academic sector to democratize AI and stop it from becoming monopolized by proprietary interests alone – the tech titans.

8) Committee amendments. Three committee amendments serve to narrow this bill's focus and augment its privacy protections. The first limits access to government data to researchers at academic institutions. The second prohibits the release of personal information, as defined in Section 1798.3 of the Civil Code. The third requires the Privacy Agency to review and approve for release any government data prior to those data being made available. The amended text follows:

> (e) The hub shall do all of the following:
>
> > (1) (A) (i) Increase lawful access to government data while protecting privacy and safeguarding access to data by developing a streamlined process for researchers ***at academic institutions*** to access data collected by state agencies.
> >
> > > (ii) Lawful access to government data increased pursuant to clause (i) shall not include access to trade secrets, as defined in Section 3426.1 of the Civil Code, obtained by the state.
> > >
> > > ***(iii) Lawful access to government data increased pursuant to clause (i) shall not include access to personal information, as defined in Section 1798.3 of the Civil Code, obtained by the state.***
> >
> > (B) In complying with subparagraph (A), the hub shall create a process for eligibility that prioritizes security by limiting who can access the data and for what purpose.
> >
> > ***(C) Any government data made available pursuant to subparagraph (A) must first be reviewed and approved for release by the California Privacy Protection Agency.***

9) Related legislation. AB 302 (Ward, Ch. 800, Stats. 2023) required CDT in coordination with other interagency bodies, to conduct a comprehensive inventory of all high-risk ADS used by state agencies on or before September 1, 2024, and report the findings to the Legislature by January 1, 2025, and annually thereafter, as specified.

SB 896 (Dodd, 2024) would largely codify Governor Newsom's executive order on the use of Generative artificial intelligence (GenAI). The bill requires assessments of the beneficial uses, potential harms, and risks to critical infrastructure of GenAI. The bill calls for the development of guidelines for public sector procurement, uses, and required trainings for the use of GenAI. The bill places obligations on state entities with respect to the use of GenAI and ADS. SB 896 is currently in this Committee.

SB 1047 (Wiener, 2024) establishes the Frontier Model Division in the Government Operations Agency and tasks it with overseeing of the most advances artificial intelligence models. SB 1047 is currently pending in the Assembly Judiciary Committee.

ARGUMENTS IN SUPPORT:

Writing on behalf of an industry coalition, Computer & Communications Industry Association states:

> SB 893 outlines several key responsibilities that would fall under the Hub which would allow it to function as a central facilitator that fosters collaboration among government agencies, academic institutions, and private sector partners to drive forward artificial intelligence research and development. The measure also seeks to responsibly support access to and development of artificial intelligence computing capacity by finding ways to reduce costs and other economic barriers research institutions may face in accessing computing infrastructure. The co-signed organizations believe this is an effective way to promote innovation for public benefit while providing protections for consumer privacy and promoting equity. We believe that SB 893 strikes the right balance of acknowledging the potential immense benefits of artificial intelligence while also safeguarding privacy, advancing security, and addressing risks and potential harms to society.

REGISTERED SUPPORT / OPPOSITION:

Support

California Chamber of Commerce
Computer & Communications Industry Association
Engine
R Street Institute
San Diego Regional Chamber of Commerce
Security Industry Association
Southwest California Legislative Council
Stanford University
TechNet
Technology Industry Association of California (TECHCA)
Tri County Chamber Alliance
University of California

Opposition

None on file.

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