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

AB 1008 (Bauer-Kahan)
Version: June 10, 2024
Hearing Date: July 2, 2024
Fiscal: Yes
Urgency: No
CK

SUBJECT

California Consumer Privacy Act of 2018: personal information

DIGEST

This bill excludes information gathered from internet websites using automated mass data extraction techniques from the definition of "publicly available" in the California Consumer Privacy Act (CCPA). The bill clarifies that personal information under the CCPA can exist in various formats, including specified examples.

EXECUTIVE SUMMARY

The CCPA grants consumers certain rights with regard to their personal information, including enhanced notice, access, and disclosure; the right to deletion; and protection from discrimination for exercising these rights. (Civ. Code § 1798.100 et seq.) In the November 3, 2020 election, voters approved Proposition 24, which established the California Privacy Rights Act of 2020 (CPRA). The CPRA amends the CCPA, limits further amendment, and creates the California Privacy Protection Agency (PPA).

The CCPA defines personal information broadly, but includes an expansive carve out for "publicly available" information. The law also does not make clear how it applies to information in various formats. This bill seeks to address the manner in which new technology is gathering and deploying personal information, especially with respect to the training and deployment of large language models. The bill carves out of the "publicly available" exception information gathered from websites using automated extraction techniques. The bill also clarifies that "personal information" can exist in various formats and provides examples, including physical, digital, and abstract formats.

This bill is author-sponsored. No timely support was received by the Committee. The bill is opposed by a variety of industry associations, including the Consumer Data Industry Association and TechNet.

PROPOSED CHANGES TO THE LAW

Existing law:

1) Establishes the 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) Provides a consumer the right, at any time, to direct a business that sells or shares personal information about the consumer to third parties not to sell or share the consumer's personal information. It requires such a business to provide notice to consumers, as specified, that this information may be sold or shared and that consumers have the right to opt out of that selling and sharing. (Civ. Code § 1798.120.)

3) Prohibits a business, notwithstanding the above, from selling or sharing the personal information of consumers if the business has actual knowledge that the consumer is less than 16 years of age, unless the consumer, in the case of consumers at least 13 years of age and less than 16 years of age, or the consumer's parent or guardian, in the case of consumers who are less than 13 years of age, has affirmatively authorized the sale or sharing of the consumer's personal information. A business that willfully disregards the consumer's age shall be deemed to have had actual knowledge of the consumer's age. (Civ. Code § 1798.120(c).)

4) Provides a business shall not be required to comply with the requirement to place a clear and conspicuous link to opt out if the business allows consumers to opt out of the sale or sharing of their personal information and to limit the use of their sensitive personal information through an opt-out preference signal sent with the consumer's consent by a platform, technology, or mechanism, based on technical specifications set forth in regulations. (Civ. Code § 1798.135.)

5) 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).)

6) Provides that "personal information" does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern. "Publicly available" means information that is:

    a) Lawfully made available from federal, state, or local government records.
    b) Information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media.
    c) Information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience. (Civ. Code § 1798.140(v)(2).)

7) Provides that "publicly available" does not mean biometric information collected by a business about a consumer without the consumer's knowledge. (Civ. Code § 1798.140(v)(2).)

8) Establishes the CPRA, which amends the CCPA and creates the 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).)

9) 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) Provides that information gathered from websites using automated mass data extraction techniques are not considered "publicly available" pursuant to the CCPA.

2) Clarifies that "personal information" can exist in various formats, including:
    a) Physical formats, including paper documents, printed images, vinyl records, or video tapes.
    b) Digital formats, including text, image, audio, or video files.
    c) Abstract digital formats, including compressed or encrypted files, metadata, or the model weights of artificial neural networks.

3) States that the Legislature finds and declares that the provisions of this bill further the purposes and intent of the California Privacy Rights Act of 2020.

COMMENTS

1. California's landmark privacy protection law

As stated, the CCPA grants consumers certain rights with regard to their personal information, as defined. With passage of the CPRA in 2020, the CCPA got an overhaul. Consumers are afforded the right to receive notice from businesses at the point of collection of personal information and the right to access that information at any time. The CCPA also grants a consumer the right to request that a business delete any personal information about the consumer the business has collected from the consumer. However, a business is not required to comply with such a request to delete if it is necessary for the business to maintain the consumer's personal information in order to carry out certain obligations or other conduct. The CPRA added a new category of information, sensitive information, which includes data such as precise geolocation and genetic information. Consumers are additionally empowered to limit businesses' use of such information. A business that sells or shares personal information to third parties is required to notify consumers that this information may be sold or shared and that they have the right to opt out of such sales. (Civ. Code § 1798.120(b).)

2. Stated intent of the bill

According to the author:

> While the development of generative artificial intelligence (GenAI) is creating exciting opportunities to grow California's economy and improve the lives of its residents, this novel technology also carries significant risks for consumer privacy. Over the past decade, massive data breaches have resulted in Californians' sensitive personal information of being posted to the internet without consent. Today, advanced GenAI systems are frequently trained using data obtained through the untargeted and automated scraping of internet websites. Once trained, these systems are capable of accurately reproducing their training data, including Californians' personal information. This bill protects Californians' privacy by clarifying that personal information, as defined in the CCPA, is still considered to be personal information regardless of how it is transmitted or stored.

3. Clarifying and updating what is "personal information" pursuant to the CCPA

"Personal information" is defined under the CCPA 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 law provides a series of examples of information that, if they identify, relate to, describe, are reasonably capable of being associated with, or could be reasonably linked, directly or

indirectly, with a particular consumer or household, are considered personal information. This includes geolocation data, online activity, and inferences drawn from this information to create a consumer profile.

This bill clarifies that this personal information can exist in various formats. This includes physical, digital, and abstract digital formats. This latter category includes compressed or encrypted files, metadata, or the model weights of artificial neural networks.

A large coalition of industry associations, including the Consumer Data Industry Association, write in opposition:

> Including the model weights of artificial neural networks as personal information is not accurate, or even technically actionable, making it challenging or impossible to handle DSARs (data subject access requests), deletion requests, etc. Model weights are mathematical values assigned to connections within an AI model. They do not contain personal data and are not extractable/returnable in a way that is identifiable or can be linked to an individual.

The author argues that new generative AI models are generally capable of reciting large segments of their training data. Therefore, the personal information included within these systems should clearly still be considered personal information, and the bill simply clarifies that.

The author points to ongoing litigation between the New York Times and OpenAI, which stems from the argument that the data being used to train OpenAI's model violates the newspaper's rights with respect to that data. The corollary here is that when these large models are trained on consumers' personal information, it does not suddenly render it no longer personal information.

The author also asserts that text-based generative AI systems are effectively a form of data compression, similar to .zip files. A research paper put out by experts at Google DeepMind and Meta AI entitled "Language Modeling Is Compression" explicitly lay this connection out. The author asserts: "As a result, a business could conceivably use a language model (like ChatGPT or LLAMA) to compress personal information and transfer it to a buyer. This business should be subject to CCPA, just as if they had sent a compressed spreadsheet containing the same information."

In order to make this provision clearer, in response to confusion from opposition, the author has agreed to the following amendment:

Amendment

Amend Section 1798.140(v)(4)(C) to read: "Abstract digital formats, including compressed or encrypted files, metadata, or ~~the model weights of artificial neural networks~~ artificial intelligence systems that are capable of outputting personal information."

The second component of this bill alters the definition of "publicly available." The CCPA provides that "personal information" does not include publicly available information, which is defined to include:

- Lawfully made available from federal, state, or local government records.
- Information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media.
- Information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience.

The law makes clear that "publicly available" does not include biometric information collected by a business about a consumer without the consumer's knowledge. This bill further provides that information gathered from websites using automated mass data extraction techniques is also excluded from what is considered "publicly available." The impetus for such a change is explained by the author:

> In January of 2024, a massive data breach known as the "Mother of all Breaches" exposed more than 26 billion records online. The leak included user data from LinkedIn, Twitter, and other heavily trafficked websites. This was not the first major breach to occur in the past decade: In 2017, Equifax suffered a breach that exposed the personal information – including social security numbers, birth dates, and addresses – of 147 million people. In 2016, Yahoo revealed that the data of all 3 billion of its users had been compromised. Target suffered a data breach in 2013 that affected 40 million credit and debit card accounts and 70 million customer records. Much of this information subsequently appeared online. Most modern GenAI systems are trained using text data sourced from the internet. The ability of GenAI systems to output training data, combined with vast quantities of personal information known to be hosted on the internet, allows GenAI to transmit personal information between individuals without the knowledge or consent of affected consumers.

The coalition in opposition, which includes the American Property Casualty Insurance Association, writes:

> Information in the public domain does not suddenly become nonpublic by virtue of how a person accesses that information—whether it is accessed in person, on paper, using a video tape, electronic copy, or "automated mass data extraction technique," or any number of lawful methods or technologies. AB 1008 would change that, declaring information in the public domain to no longer be public information based on how an entity accessed it, causing any number of unintended or illogical consequences and running afoul of well-established First Amendment rights in receiving and disseminating information. The consequences of doing so range from unintended, to illogical, ill-advised, and harmful, disrupting many industries that legitimately rely on public information to provide services more accurately and efficiently, and potentially even causing great harm by interfering with research or public safety efforts.

The author has agreed to amendments that remove this portion of the bill.

4. Furthering the purpose and intent of the CPRA

Section 25 of the CPRA requires any amendments thereto to be "consistent with and further the purpose and intent of this act as set forth in Section 3." Section 3 declares that "it is the purpose and intent of the people of the State of California to further protect consumers' rights, including the constitutional right of privacy." It then lays out a series of guiding principles. These include various consumer rights such as:

- consumers should know who is collecting their personal information;
- consumers should have control over how their personal information is used; and
- consumers should benefit from businesses' use of their personal information.

Section 3 also includes a series of responsibilities that businesses should have. These include:

- businesses should specifically and clearly inform consumers about how they use personal information; and
- businesses should only collect consumers' personal information for specific, explicit, and legitimate disclosed purposes.

Section 3 also lays out various guiding principles about how the law should be implemented.

The bill helps clarify what is to be considered "personal information" under the CCPA. Therefore, as it explicitly states, this bill "furthers the purposes and intent of the California Privacy Rights Act of 2020."

SUPPORT

None received

OPPOSITION

American Association of Advertising Agencies
American Council of Life Insurers
American Property Casualty Insurance Association
Association of California Life & Health Insurance Companies
Association of National Advertisers
California Association of Realtors
California Chamber of Commerce
California Land Title Association
Coalition for Sensible Public Records Access
Computer & Communications Industry Association
Consumer Data Industry Association
Insights Association
Software & Information Industry Association
State Privacy and Security Coalition, INC.
TechCA
Technet

RELATED LEGISLATION

Pending Legislation:

SB 1223 (Becker, 2024) includes "neural data," as defined, within the definition of "sensitive personal information" for purposes of the CCPA. SB 1223 is currently in the Assembly Privacy and Consumer Protection Committee.

AB 1824 (Valencia, 2024) requires a business that assumes control of all or some part of a transferor business that includes the transfer of a consumer's personal information to comply with a consumer's direction to the transferor pursuant to the CCPA. AB 1824 is currently on the Senate Floor.

AB 1949 (Wick, 2024) prohibits the collection, sharing, selling, using, or disclosing the personal information of minors without affirmative consent from either the minor or their parent or guardian, as provided. The bill provides for regulations to be promulgated by the PPA. AB 1949 is currently in this Committee.

AB 2013 (Irwin, 2024) requires developers of 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. AB 2013 is currently on the Senate Floor.

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 the Senate Appropriations Committee.

Prior Legislation:

AB 947 (Gabriel, Ch. 551, Stats. 2023) included personal information that reveals a consumer's citizenship or immigration status in the definition of "sensitive personal information" for purposes of the CCPA.

AB 1194 (Wendy Carrillo, Ch. 567. Stats. 2023) provided stronger privacy protections pursuant to the CCPA where the consumer information contains information related to accessing, procuring, or searching for services regarding contraception, pregnancy care, and perinatal care, including abortion services.

AB 375 (Chau, Ch. 55, Stats. 2018) established the CCPA.

PRIOR VOTES:

PRIOR VOTES NOT RELEVANT

Assembly Floor (Ayes 61, Noes 15)
Assembly Appropriations Committee (Ayes 11, Noes 3)
Assembly Water, Parks and Wildlife Committee (Ayes 11, Noes 0)
****