June 25, 2024

TO: Members, Senate Judiciary Committee

SUBJECT: AB 1008 (BAUER-KAHAN) CALIFORNIA CONSUMER PRIVACY ACT OF 2018: PERSONAL INFORMATION
OPPOSE – AS AMENDED JUNE 10, 2024
SCHEDULED FOR HEARING – JULY 2, 2024

The undersigned organizations must respectfully **OPPOSE AB 1008 (Bauer-Kahan)** as amended June 10, 2024, which expands the scope of the California Consumer Privacy Act's (CCPA) definition of "personal information", by narrowing the exception for "publicly available" information based on the method by which the data is collected, as opposed to the source or nature of the data itself. At the most basic level, this bill states that information that is placed in the public sphere is publicly available up until the point that a "mass data extraction technique" is used to retrieve it—regardless of whether the public has any legal right to that information, or the creator of the content purposefully made that information available to the public for public consumption.

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. If the underlying objective is to ebb data mining or to reduce the amount of information available for any number of uses, such as training AI models, we encourage you to consider the full ramifications of furthering such public policies; regardless, the state cannot get there by infringing on constitutional rights, including the right to access public records under both the State Constitution and statutory law.

Fifty-five years after the Supreme Court expressly recognized the "right to receive information and ideas", it is disconcerting to see legislation introduced in California suddenly suggesting that citizens do not, or should not, have a constitutionally protected right to freely gather and disseminate information that is in the public domain, and that the bill itself acknowledges *is* publicly available information, simply because of the method and technology used to gather it. Even if it were not the intent of **AB 1008**, it is the message being conveyed by the bill, nonetheless.

The bill unquestionably runs afoul of not only the First Amendment right to receive and disseminate information, but also the California Constitution which protects the right of the people to access public records

The nature of data and privacy interests does not change depending on how it is accessed, let alone a specific type of technology to access it. If information is rendered publicly available because it is in the government records, or because it is available in mass media or lawfully made available to the public by the consumer, or it is made available to a person by the consumer without placing any restrictions on the audience, any privacy interests in that information have changed. Whether someone accessed that information by way of pamphlets that are printed in mass numbers and distributed, or by way of a bot that can search through numerous records to find that information more quickly makes no difference—in fact, stating that the public nature of the same piece of information, received from the same source, suddenly becomes non-public by virtue of the method by which the information is received or disseminated is simultaneously illogical and unlawful.

Both the California Constitution (article 1, Section 3, stating that "[t]he people have the right of access to information concerning the conduct of the people's business, and therefore, the meetings of public bodies, and the writings of public officials and agencies shall be open to public scrutiny") and the California Public Records Act (Gov. Code Sec. 7920.000 et seq; see Sec. 7920.520, stating that "in enacting [the CPRA], the Legislature, mindful of the right of individuals to privacy, finds and declares that access to information concerning the conduct of the people's business is a fundamental and necessary right of every person in this state") recognize the people's right to access public records. Notably, the CPRA prohibits limitations being placed on access to a public record based on the purpose for which the record is being requested if the record is otherwise subject to disclosure. (*See* Gov. Code Sec. 7921.300.) It is therefore hard to imagine that the right could be limited, instead, based on how the information is collected (*e.g*., on an individual basis, or by downloading documents in bulk, or otherwise).

Separately, there is, of course, also the constitutional right to access that information under the First Amendment, which not only protects the speaker's freedom to speak, but the listener's right to receive information and ideas. This, of course, is not a new or untested principle. The U.S. Supreme Court has long recognized that "the right to receive information and ideas" flows from the First Amendment right of speech. (*Stanley v. Georgia* (1969) 394 U.S. 557, wherein the Court first used that precise phrase.) Indeed, it has roots in case law dating back to the early 1940s, when the Court protected the right of a woman to distribute religious materials door-to-door, stating in the words of Justice Hugo Black that "[t]his freedom embraces the right to distribute literature, and necessarily protects the right to receive it." (*Martin v. City of Struthers*, 319 U.S.141, 142-143 (1943).) And as recently as 2011, the Court has affirmed that "the creation and dissemination of information is speech for First Amendment purposes." *Sorrell v. IMS Health, Inc*., 564 U.S. 552, 570 (2011).

As noted above, whether it is intended or not, this bill seems to challenge such principles, suggesting that publicly available information could be rendered nonpublic because of the method used to receive or disseminate it. It is also the clear effect of treating publicly available information as though it is suddenly "personal information" any time that mass data extraction techniques are used to gather the information. By treating it as such, **AB 1008** applies the CCPA restrictions on the collection, use, and disclosure of personal information to publicly available information, extinguishing the right to access and freely share this information. In doing so, the bill overrides the balance of competing interests that was sought by voters when they broadened the CCPA's definition of "publicly available"[1] under Proposition 24, undermining the substantial public and societal benefits of using publicly available information, and running afoul of the First Amendment, as well as article 1, Sec. 3 of the California Constitution. ***The State may not infringe upon these rights to protect a generalized interest in consumer privacy,*** and such broad-brush restrictions on the collection and dissemination of publicly available information are not narrowly tailored to further compelling governmental interests, as would be required to survive strict scrutiny.

AB 1008 overrides the voters' attempt to balance competing interests when expanding the meaning of "publicly available" in Proposition 24 for questionable benefit at best

AB 1008 restates the existing law definition of what is considered "publicly available," as well as the existing exclusion of biometric information collected without the knowledge of the consumer from this definition, verbatim (only changing the applicable legal citation). The only substantive change made to the definition of "publicly available" by this bill, is that it additionally excludes "information gathered from internet websites using automated mass data extraction techniques." The exclusion of "information gathered from internet

[1] AB 375 (Chau and Hertzberg) Ch. 55, Stats. 2018

websites using automated mass data extraction techniques" from what is defined as "publicly available," however, directly contradicts the provisions expressly defining what *is* included under the term.

Specifically, under Proposed Section 1798.140(v)(2)(B)(i)(I)-(III), "'publicly available' includes" any of the following: (I) information that is lawfully made available from government records; (II) 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; and (III) 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. Meaning that existing law recognizes that there are easy ways for consumers and website operators to restrict access to data via audience controls or similar mechanisms, ensuring that the information would not be considered publicly available. In the next paragraph, however, **AB 1008** would amend the definition to state that "'publicly available" does not include information gathered from internet websites using automated mass data extraction techniques"[2], which seems to override the preexisting balance struck between consumer's preferences and the substantial public and societal benefits of using publicly available information.

While the Legislature can certainly do so, assuming that it would be in furtherance of the intent of Proposition 24, it is not clear what benefit is gained here that the consumer could not equally achieve by restricting the information to a specific audience. What *is* clear is that it appears to do so despite infringing upon constitutional rights to receive or disseminate information, or the statutory right to access public records. Take, for example, a scenario where a parent decides to post a message about their missing child on their public social media page or post a public message on their usually *private* social media page, in hopes that the post will be disseminated, far and wide, to help find her. If the information in the post is gathered from a website that helps identify and find missing children using some form of "mass data extraction technique", this otherwise "publicly available" information would no longer be considered "publicly available" despite the obvious intent of the consumer who made it public. Instead, that information would have to be treated the same as "personal information" under the CCPA, which requires compliance with notice requirements prior to the collection, use or distribution of that information (if the website also meets one of the three thresholds to be considered a "covered business").

What happens if a newspaper wishes to collect and use such information in reporting on the uptick of missing children? What if the child is 14 years of age? Now that **AB 1008** would exclude the public post from the publicly available exemption, rendering it "personal information", can the newspaper collect or disclose the story featuring that information without first obtaining the affirmative authorization from the 14-year-old? Furthermore, is the newspaper now a "data broker" under the Data Broker Registry law, as they received the "personal information" via an indirect relationship with the consumer?

All of this is to say: automated data collection and analysis is a critical tool for investigative journalists, social scientists, and US businesses alike, among others, and grouping the definition of the nature of the data with the means of collection of that data is likely to have unintended and potentially illogical consequences for core aspects of the modern internet and free flow of information, and profound impact on core First Amendment activity.

Potential unintended, and otherwise illogical, consequences are inevitable under AB 1008 and innumerable given the lack of clarity in the bill

Even though a person's right of access to public records under CPRA does not hinge on justifiable or unjustifiable uses, the reality is that many industries in California rely on publicly available information for legitimate and beneficial uses. As noted above, by grouping the definition of the nature of the data, with the means of collection of that data, to fundamentally alter what is considered "publicly available" information, **AB 1008** is likely to have unintended and potentially illogical, if not absurd, consequences for core aspects of the modern internet and free flow of information. It will also have costly ramifications for industries reliant on that free flow and significantly limit all automated data collection, which is fundamental to countless

[2] Note, as drafted it is unclear if the party gathering the information is using a mass data extraction technique to gather the information from the website, or if the website from which the information is gathered uses a mass data extraction technique. i.e., "Publicly available does not include information gathered from internet websites using automated mass data extraction techniques" could be read to mean "gathered from internet websites by way of using" or it could mean "gathered, through any means, from the internet websites if the websites use mass data extraction techniques". Presumably, the intended reading is the former ("by way of using").

beneficial uses, such as search engines, web archives, and link previews. Imagine the impact this would have on any number of industries:

For example, **AB 1008** could drastically slow down real estate transactions and significantly disrupt home purchases or refinances depending if realtors could no longer do automated pulls of home sale data from public databases to perform analyses on housing trends, or if title companies' ability to conduct fast and wide searches of publicly available information contained in land records from county recorders were impaired. In the latter instance, while title companies obtain public records in electronic format directly from county recorders in bulk, limitations in recorders' systems dictate that the files are transmitted in digitized – as opposed to digital – form, meaning that data must then be extracted – via OCR[3] and other methods – from the documents to be electronically searchable. Under the broadly drafted provisions of **AB 1008**, however, it is unclear whether these processes – which ensure that title searches can be conducted accurately and promptly – would be prohibited, as they arguably involve "mass data extraction techniques" of publicly available information that is transmitted via the internet. Will a county therefore be found to have committed a data breach for what is a longstanding industry practice intended to effectuate transfer of title to real property? Has the title company unlawfully obtained what they were lawfully provided access to via the state constitution and CPRA?

Adding to the list of potential problems is that fact that the bill does not identify what is meant by "automated mass data extraction techniques". Are databases of court documents and published legal opinions built using "automated mass data extraction techniques" (e.g. courtlistener.com - which is a non-profit free legal research tool)? Is the use of an API to access "tweets" or to verify contractor licenses or doctor and nurse licenses from government websites an "automated mass data extraction technique"? If so, how would it impact something like scraping activity targeted at a social network, which has been upheld by courts?[4] Could it undermine the ability to scrape social media activity during an active shooter situation, where public posts often can provide first responders in time information about where the shooter or victims may be located? How might it set back researchers that rely on web scraping or web research to inform their work? How will it impact the availability of training data? Will it slow, if not halt, the development of fair, safe, and beneficial AI?

Such questions are just the tip of the iceberg. Ultimately, by excluding information gathered from websites by way of "automated mass data extraction techniques" from the CCPA's definition of "publicly available" information **AB 1008** is very likely to disrupt any number of public benefits and bring any number of industries to a grinding halt. Ultimately, it should not matter if the data comes from an API or on microfiche or on paper – if it is publicly available if accessed via a paper copy, the law should not suddenly treat it different if obtained from an API.

The unconstitutionality of the bill aside, it is unclear what public good or interests might result from the passage of AB 1008, that would outweigh the vast harm that will result

At its core, the bill appears rooted in an assumption that any use of publicly available information by the private sector is inherently either nefarious or harmful. Already, this year, we have seen numerous bills on the use of "personal information" [as understood under the CCPA, which includes information that can only indirectly be linked to person or household] to train AI, either seeking to limit the availability of training data for AI or regulate business that do so, as though training AI models with real data is inherently bad. (*See e.g.*, AB 2877 (Bauer-Kahan) and AB 3204 (Bauer-Kahan).)

In reality, when a developer trains a foundation model on web scale data, they are not trying to learn about individuals, and they can implement filters that are designed to prevent disclosure of information about private individuals making it so that models don't respond to these sorts of questions and protect privacy of individuals. Personal information processed to train a foundation model is, in fact, generally limited to information that is incidentally included in datasets. The goal of that training is to help the model learn about statistical relationships in natural language (such as words or characters that often appear in context with other words or characters) or between natural language and corresponding images. Yet, **AB 1008** would

[3] Optical character recognition.

[4] *See hiQ Labs, Inc. v. LinkedIn Corp*., 938 F.3d 985 (9th Cir. 2019), the 9th Circuit case about web scraping, wherein a small data analytics company, hiQ Labs, used automated bots to scrape information from public LinkedIn profiles. The 9th circuit court affirmed the district court's preliminary injunction, preventing LinkedIn from denying hiQ Labs, from accessing LinkedIn's publicly available LinkedIn member profiles.

profoundly limit the availability of data for AI training, as well as for search, or web archive, even if the data used is not readily identifiable or considered "personal information" under most other California laws (the term is given a uniquely broad definition in the CCPA), or would have been considered publicly available information if collected by any other means or using any other technologies.

Moreover, the bill is unlikely to create better privacy outcomes for data subjects by bringing training data under the definition of "personal information" in the CCPA. Indeed, requiring developers to index and search through large volumes of training data for the purpose of responding to data subject requests goes against principles against re-identifying data where the purpose of processing does not require the identification of data subjects. Yet that is precisely what would be required. A model developer is unlikely to be able to verify the identity of a data subject and validate that a particular piece of data relates to that individual without taking additional steps that are contrary to the privacy interests of data subjects, such as indexing training data sets or collecting large amounts of additional personal data from the data subject in violation of data minimization principles.

Carving out a specific category of publicly-available information, even if someone has chosen to share it publicly (e.g., on an ungated blog or social media page), based solely on the method by which it was collected does not make any practical or logical sense. It also expands the CCPA far beyond what CA has chosen to protect in its privacy laws and upends the balance between privacy rights with rights in the free flow of information that was clearly sought by the voters. And to what end? It is neither consistent with constitutional law, nor supported by statutory law, and any resulting benefit to Californians is questionable at best, whereas the list of harms to that would befall them and the entire economy appear innumerable at this time.

AB 1008 is also out of step with California's open data policy

Clearly, **AB 1008** seeks to dissuade collection of publicly available data through automated and machine-readable formats, ignoring many positive applications of such collection. This, of course, is also inconsistent with California's current open data policy. California clearly has sought to maximize access, use, and reuse of government data through portals which support dissemination of data without restriction, as reflected in the State's 2019 Open Data Policy Requirements mandating state agencies to "prioritize the use of open formats that are non-proprietary, publicly available, and that place no restrictions upon their use" and that "systems [be] scalable, flexible, and facilitate extraction of data in multiple formats and for a range of uses…"

Pursuant to the Department of Technology's Technology Letter (TL) 19-01 announcing the Open Data Policy, this policy "promotes more accessible, discoverable, and usable data that impacts economic development and improves government services. In addition, open data encourages informed policy decisions, performance planning, research, and scientific discoveries, and increased public participation in democratic dialogue. Open data is public data collected by the state through its routine business activities and published in a format that is easy to search, easy to download and easy to combine with other data sets from other sources; it does not include private or confidential data about individuals."

Similarly, agencies such as the Secretary of State's office actively encourage the public to access any and all information through public websites, rather than submitting formal public records requests. **AB 1008**'s attempt to limit automated collection of publicly available data, such as public records data, is thus clearly at odds with the previously policy goals of this State.

AB 1008's inclusion of "model weights" as a format that personal data can take is inaccurate

Among other things, **AB 1008** lists different formats in which personal information can exist. These include physical formats such as paper documents or video tapes, digital formats including text, audio or video files, and "abstract digital information" such as compressed or encrypted files, metadata, or "the model weights of artificial neutral networks."

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. A developer therefore could not "delete" personal data from a model weight. It is

also unclear to us, at this time, the impact that this will have on the carveouts from “personal information” for “deidentified information” or “aggregate consumer information.”

Ultimately, we strongly believe that **AB 1008** entirely rests upon a legal fiction that the nature of publicly available information can change depending on how the information is gathered and arbitrarily turns information that is unequivocally within the public domain as “personal information” simply by virtue of it being gathered using “automated mass data extraction techniques”, which is completely undefined, yet seemingly assumed to be dangerous or inherently diametrical to privacy interests. We also believe that the bill will lead to innumerable harmful, unintended or illogical outcomes, disrupting many industries (if not grinding them to a halt) and foreclosing any number of beneficial uses cases, with at least the partial goal of stopping publicly available information from being used to train AI. Because of these concerns, and because the bill unquestionably infringes upon both the First Amendment right to receive and disseminate information, and the constitutional and statutory rights to access public records under California law, we must **OPPOSE AB 1008 (Bauer-Kahan)**.

Sincerely,

Ronak Daylami
Policy Advocate
on behalf of

American Association of Advertising Agencies (4A's), Alison Pepper
American Council of Life Insurers, John Mangan
American Property Casualty Insurance Association, Laura Curtis
Association of California Life and Health Insurance Companies, Matthew Powers
Association of National Advertisers, Travis Frazier
California Association of Realtors, Anna Buck
California Chamber of Commerce, Ronak Daylami
California Land Title Association, Anthony Helton
Coalition for Sensible Public Records Access, Richard J. Varn
Computer and Communications Industry Association (CCIA), Naomi Padron
Consumer Data Industry Association, Zachary Taylor
Insights Association, Howard Fienberg
Software & Information Industry Association, Anton Van Seventer
State Privacy & Security Coalition, Andrew Kingman
TechCA, Courtney Jensen
TechNet, Dylan Hoffman

cc: Legislative Affairs, Office of the Governor
Elise Gyore, Office of Assemblymember Bauer-Kahan
Consultant, Senate Judiciary Committee
Morgan Branch, Consultant, Senate Republic Caucus

RD:ldl