FUTURE OF PRIVACY FORUM
1350 Eye Street NW, Suite 350, Washington, DC 20005 | 202-768-8950 | fpf.org

September 11, 2024

Via Electronic Mail

The Honorable Gavin Newsom
1021 O Street, Suite 9000
Sacramento, CA 95814

Dear Governor Newsom:

The Future of Privacy Forum (FPF) writes to you regarding Assembly Bill (AB) 1008, an enrolled bill concerning privacy and artificial intelligence. FPF is a non-profit organization dedicated to advancing privacy leadership, scholarship, and principled data practices in support of emerging technologies in the United States and globally. FPF seeks to support balanced, informed public policy and equip regulators with the resources and tools needed to craft effective regulation.[1]

If enacted, AB 1008 would amend the definition of personal information under the California Consumer Privacy Act (CCPA) to provide that personal information can exist in "abstract digital formats," including in "artificial intelligence systems that are capable of outputting personal information."[2] FPF writes this letter to:

(1) Identify several important ambiguities that would likely require regulatory guidance should AB 1008 be enacted;

(2) Highlight relevant, ongoing research by European data protection authorities as to whether large language models store personal data; and

(3) Recommend that the primary interpreters and enforcers of the CCPA—the California Privacy Protection Agency (CPPA) and the Office of the Attorney General (OAG)—engage in a constructive dialogue with their American and global regulatory peers on this issue.

[1] The opinions expressed herein do not necessarily reflect the views of FPF's supporters or Advisory Board.
[2] A.B. 1008, 2024 Reg. Sess. (Cal. 2024), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240AB1008.

1. **AB 1008's effects on personal privacy and business obligations will turn on the distinction between AI models versus AI systems and the differences between types of AI models.**

AB 1008 would amend Cal. Civ. Code § 1798.140, subd. (v) to clarify that, under the CCPA, personal information *can* exist in various formats such as "abstract digital formats," including in "artificial intelligence systems that are capable of outputting personal information." Neither the bill nor the CCPA defines "artificial intelligence systems." As a result, since there is a spectrum of AI models, architecture, and infrastructure, it is unclear whether the bill pertains to (1) specific types of AI models like LLMs or other formats specifically designed to process or output personal data; and (2) singular models or broader AI systems composed of one or more AI models (which may or may not include LLMs) plus other technologies and processes.

If enacted, the interpretation and enforcement of this provision by the CPPA and OAG will have significant effects on the scope of Californians' privacy rights and the obligations businesses developing and deploying LLM-based AI systems will have under the law. For example, whether personal information exists in AI models will have significant operational implications if businesses therefore have to apply individual data rights under the CCPA (e.g., right to delete personal information) to models, which could disincentivize AI model development in California. Given the significance of this question, FPF writes to highlight emerging research and guidance from some of the European Union's data privacy regulators considering whether LLMs store personal information.

2. **Recent guidance by some European regulators preliminarily propose that large language models do not store personal data.**

In July 2024, the Hamburg Data Protection Authority[3] ("Hamburg DPA") released a discussion paper on personal data and large language models.[4] Although FPF does not endorse or disclaim the paper's findings, which focus specifically on LLMs and personal data as defined in the EU's General Data Protection Regulation ("GDPR") (unlike the broader focus of AB 1008 on AI systems), the DPA's analysis offers a valuable starting point for further expert collaboration and discussion

[3] Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (HmbBfDI). The Hamburg DPA is one of the 16 state-level German Data Protection Authorities, having competence over enforcing the GDPR in the state of Hamburg.

[4] Press Release, HmbBfDI, Hamburg Theses on Person Reference in Large Language Models (July 15, 2024), https://datenschutz-hamburg.de/news/hamburger-thesen-zum-personenbezug-in-large-language-models; *but see* David Rosenthal, *Part 19: Language Models With and Without Personal Data*, VISCHER BLOG (July 17, 2024), https://www.vischer.com/en/knowledge/blog/part-19-language-models-with-and-without-personal-data/ (asserting that whether personal data are contained in an LLM must be "assessed from the perspective of those who formulate the input and those who have access to the output").

because the GDPR's definition of personal data is substantively similar to the CCPA's definition of personal information.[5]

In the paper, the Hamburg DPA provides a technical explanation about how LLMs are developed, covering tokenization (breaking text into units) and embeddings (numerical representations) that capture relationships between these units. The Hamburg DPA views this training process as transforming text into "abstract mathematical representations" that lose "concrete characteristics and references to specific individuals," instead reflecting "general patterns and correlations derived from the training data as a whole."[6] Thus, in the Hamburg DPA's view, these models do not store personal data under GDPR because an LLM does not contain any data that relates to an identified or identifiable person:[7]

> LLMs store highly abstracted and aggregated data points from training data and their relationships to each other, without concrete characteristics or references that 'relate' to individuals. . . . In LLMs, the stored information already lacks the necessary direct, targeted association to individuals that characterizes personal data in [Court of Justice of the European Union] CJEU jurisprudence: the information 'relating' to a natural person.[8]

This ongoing work by the Hamburg DPA is directly relevant to AB 1008 because the bill's original stated purpose was to clarify that, among other things, "the model weights of artificial neural networks" can be a format in which personal information exists.[9] In fact, AB 1008's legislative history specifically notes that the 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."[10] The bill has been amended to no longer directly reference model weights, but the current language may still apply to model weights in AI systems.

AB 1008's legislative history provides further statements of intent that are in tension with the Hamburg DPA's findings. According to the bill's author, one risk that motivated this legislation is that, "[o]nce trained, these [GenAI] systems are capable of accurately reproducing their training data, including Californians'

---

[5] *Compare* Council Regulation 2016/679, General Data Protection Regulation, 2016 O.J. (L 119) Art. 4(1) ("'personal data' means any information relating to an identified or identifiable natural person"), *with* Cal. Civ. Code § 1798.140, subd. (v) ("'Personal information' means 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").

[6] HmbBfDI, *Discussion Paper: Large Language Models and Personal Data* at 4 (July 15, 2024), https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf.

[7] *Id.* at 4–5.

[8] *Id.* at 6.

[9] Bill Analysis, A.B. 1008, Sen. Judiciary Comm., (June 29, 2024), https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202320240AB1008#.

[10] Bill Analysis, A.B. 1008, Sen. Judiciary Comm., (June 29, 2024), https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202320240AB1008#.

personal information."[11] The Hamburg DPA, however, concluded that LLM outputs "do not store the texts used for training in their original form, but process them in such a way that *the training data set can never be fully reconstructed from the model*."[12] Both viewpoints can be correct—that LLMs can reproduce personal information from their training data but that training data cannot be "fully reconstructed" from a model. Nevertheless, this is precisely the kind of difficult, factual, technical discrepancy from which global regulators would benefit from consulting one another.

Importantly, the Hamburg DPA's discussion paper emphasizes that the paper is meant to "stimulate further debate," rather than to be a final say on this issue. There is no consensus so far reached among European DPAs, since the European Data Protection Board—the forum where DPAs all meet and agree on common approaches for the application of the GDPR—has not yet adopted a position on this issue. The findings of this discussion paper are specific to GDPR, but the insights and conclusions will be relevant to comparable definitions of personal data or personal information under other privacy and data protection regimes. Therefore, we recommend that California regulators proactively engage other American and global regulators on this issue and work towards consensus.

3. **California privacy regulators should proactively engage global regulators and open a dialogue on these issues.**

As a leader in the global privacy and data protection regulatory community,[13] the CPPA has previously declared its commitment to collaborating with international data protection authorities, as exemplified by the Agency's recent cooperation agreement with France's Commission nationale de l'Informatique et des Libertés (CNIL),[14] another data protection authority which is considering the application of GDPR to AI systems.[15] The announcement of that agreement stressed that

---

[11] Bill Analysis, A.B. 1008, Sen. Judiciary Comm., (June 29, 2024), https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=202320240AB1008#.

[12] HmbBfDI, *Discussion Paper: Large Language Models and Personal Data* at 4 (July 15, 2024), https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Discussion_Paper_Hamburg_DPA_KI_Models.pdf (emphasis added); *but see* Nicholas Carlini *et al.*, *Extracting Training Data from Large Language Models* (June 15, 2021), https://arxiv.org/abs/2012.07805 (demonstrating how extraction attacks can result in LLMs reproducing training data).

[13] Press Release, Cal. Priv. Prot. Agency, California Privacy Protection Agency Admitted into Global Privacy Assembly (Oct. 27, 2022), https://cppa.ca.gov/announcements/2022/20221027.html; Press Release, Cal. Priv. Prot. Agency, California Privacy Protection Agency Admitted into Asia Pacific Privacy Authorities (APPA) (May 12, 2023), https://cppa.ca.gov/announcements/2023/20230512.html; Press Release, Cal. Priv. Prot. Agency, California Recognized By Dubai International Financial Centre Due to Data Protection Law and Regulations (Aug. 9, 2023), https://cppa.ca.gov/announcements/2023/20230809.html.

[14] Press Release, Cal. Priv. Prot. Agency, CPPA Announces Cooperation with French Data Protection Authority (June 25, 2024), https://cppa.ca.gov/announcements/2024/20240625.html.

[15] In June 2024, CNIL released a series of "AI how-to sheets" regarding application of GDPR to the development and deployment of AI systems. Press Release, AI: CNIL Publishes Its First Recommendations on the Development of Artificial Intelligence Systems, CNIL (June 7, 2024), https://www.cnil.fr/en/ai-cnil-publishes-its-first-recommendations-development-artificial-intelligence-

California law and GDPR both "encourage international collaboration on privacy protections," and that the CNIL agreement is designed to "facilitate joint internal research and education related to new technologies and data protection issues, share best practices, and convene period meetings."[16] This issue—whether and to what degree personal information exists within AI systems—is precisely the type of emerging issue on which the CPPA should seek to be an international thought leader. While the Hamburg DPA's discussion paper is only persuasive and not a final regulatory decision—and recognizing that California may have different policy positions from our European counterparts—engaging with international data protection authorities offers California an opportunity to strengthen relationships with other regulators, share its technical expertise, benefit from the collective wisdom of the global data protection community, and promote international consistency in data protection regulation.

The extent to which AI models, whether standalone or as part of AI systems, contain personal information is a global uncertainty with major implications for privacy, business obligations under privacy laws, and AI development. Cross-border collaboration will enable California regulators to build a shared understanding of personal information in LLMs, non-LLM AI models, and AI systems, even if their policy conclusions differ from those of European regulators. This shared understanding will help assess whether AB 1008 effectively addresses these issues and comports with CCPA requirements, such as data minimization and consumer rights. California has the opportunity to lead by engaging with its international peers to develop a well-informed, nuanced approach based on technical expertise and sound policy. By collaborating with other regulators, the CPPA and OAG can further benefit from shared knowledge and reach a deliberate conclusion.

Thank you,

Jordan Francis
Policy Counsel, Future of Privacy Forum
jfrancis@fpf.org

Beth Do
Christopher Wolf Diversity Law Fellow, Future of Privacy Forum
bdo@fpf.org

---

systems. CNIL has since launched a questionnaire to inform further guidance on the application of GDPR to AI systems, including "to shed light on the conditions under which AI models can be considered anonymous or must be regulated by the GDPR." Press Release, Questionnaire on the application of the GDPR to AI models, CNIL (July 2, 2024), https://www.cnil.fr/fr/webform/questionnaire-sur-lapplication-du-rgpd-aux-modeles-dia-questionnaire-application-gdpr-ai-models.

[16] *Id.*