UCLA School of Law

MELODI DINCER
RESIDENT FELLOW & LECTURER
UCLA INSTITUTE FOR TECHNOLOGY, LAW & POLICY

SCHOOL OF LAW
BOX 951476
LOS ANGELES, CALIFORNIA 90095-1476
Email: dincer@law.ucla.edu

California Privacy Protection Agency
2101 Arena Boulevard
Sacramento, CA 95834

Dear Board Members,

We submit this informal comment as a cross-institutional group of technology law researchers who sought answers from leading California businesses about whether they have used our personal information to train their generative AI (GenAI) systems. We believe we are entitled to these disclosures based on our right to know under the California Consumer Protection Act (CCPA). We seek clarity from the Agency after these businesses sent inadequate responses to our data subject access requests (DSARs).

We commend the Agency on its ongoing efforts to establish guidelines around artificial intelligence and automated decision-making technology, including its recent steps to advance rulemaking on related regulations. These actions strengthen the CCPA's mandate to give consumers greater control over their data. But we believe that a missing piece in this conversation is how the CCPA functions to protect personal information that businesses have included in their GenAI training datasets.

The growing presence of GenAI in the current digital landscape is incontrovertible. In a couple of years, California companies' GenAI systems, like OpenAI's ChatGPT and DALL·E and Meta's Llama, have gone from being entirely novel technologies in the eyes of consumers to regular fixtures in educational, business, and creative spaces. Explicit direction from the Agency is most essential and pressing during this period of breakneck growth for this emergent technology. We therefore urge the Agency to engage directly with the question of how Californians can use the CCPA to control whether and how much of their personal information enters GenAI systems.

We ask the Agency to clarify the following questions:

1. Does the CCPA protect personal information used to train generative AI systems?
2. If yes, what details must Californians include in their DSARs, what identity verification may businesses request in response, and how must businesses comply with Californians' DSARs?

In Section 1, we address the first question and discuss why the Agency should find that such personal information is protected under the Act. In Section 2, we focus on the second question based on difficulties we experienced after sending DSARs to several businesses for our data (described further below). Section 3 offers suggestions for how the Agency can address the concerns we raise.

Background

This comment and theDSARs that inspired it are a collaboration between California-based researchers with UCLA's Institute for Technology, Law & Policy, USC's Knowing Machines Project, and NYU's Technology Law & Policy Clinic. To test businesses' CCPA compliance regarding GenAI training data, our researchers sent DSARs to Inflection AI, OpenAI, Google, Meta, Microsoft, Anthropic, and Amazon, seeking disclosure of all data about the individual requester included in the development, training, and/or improvement of any Large Language Model (LLM), Generative Adversarial Network (GAN), Diffusion Model, and/or any similar system.

Beyond meeting the threshold requirement for CCPA compliance, these businesses are among the biggest players in GenAI. We present their responses not as an exhaustive sample, but rather to demonstrate how even the most prominent GenAI businesses are failing to meet their CCPA disclosure obligations.

This table describes the type of responses the researchers received generally:

Business	**Request Method**	**Initial Response**	**Later Response(s)**	**Disclosure(s)**
Amazon	Email DSAR	Redirection to (1) online account information (Your Orders), (2) Privacy Policy, (3) online portal to request Amazon account data	N/A	None
Anthropic	Email DSAR	Confirmation of receipt; notice that business may verify requestor's identity; claim to respond or communicate a decision	N/A	None
Google	Google Form	Confirmation of submission	N/A	None
Inflection AI	Email DSAR	Referral to Privacy Policy	Instructions to type requestor's name and phrase "EXPORT MY DATA [xxxx]" to Pi.AI chatbot	None

Meta	Email DSAR	Notice of inability process DSAR; request to provide examples or screenshots that show evidence of personal information as Meta's genAI models' output(s); redirection to Privacy Policy; redirection to Privacy Center page on GenAI	N/A	None
Microsoft	Microsoft Form (no account), Email DSAR	Further directions to authenticate case with the privacy team via privacy form (requestor needed to also authenticate Microsoft account)	Email confirming request of "access [to] personal data linked to your Microsoft account", with link to portal to download data from privacy dashboard	CSV files: (1) "Product and Service Usage" (2) "Search Requests and Queries"
OpenAI	Email DSAR	Automated response with information on (1) how to export ChatGPT chat history, (2) how to delete an OpenAI account, (3) how to correct inaccurate or incomplete data	N/A	None

Their responses ranged from automated summaries of their privacy policies to counter-demands for the requester to produce the very information they were seeking. All of them failed to comply with our valid DSAR requests, which we crafted based on the CCPA's text. We are a team of experienced researchers in the AI/machine learning field who have the time and ability to carefully craft air-tight DSARs. If the biggest names in GenAI today cannot respond to our requests adequately, how will they respond to requests sent by less sophisticated or informed parties, including typical users of AI systems?

Our research suggests this is a timely opportunity for the Agency to provide guidance both to consumers and businesses alike. In our view, businesses should provide requestors with comprehensive data disclosures, including detailed explanations of all data categories, the potential values they may contain, and the significance of these values in their model training processes.

We therefore ask the Agency to clarify under what authority consumers may exercise their right to know if and what personal information businesses are collecting and/or using to develop and train GenAI systems; this may include providing a template GenAI DSAR for consumers to use in crafting

their requests. Furthermore, we implore the Agency to establish reasonable identity verification protocols and comprehensive disclosure standards for GenAI DSARs. In the alternative, if the Agency finds that the CCPA does not currently provide authority for such a right to know, we ask the Agency to identify the most appropriate regulatory or legislative path to ensuring Californians' personal information remains protected when processed by emergent GenAI systems.

1. The CCPA's Data Protection Rights Should Cover Personal Information Used to Train Generative AI

a. For the CCPA to fulfill its consumer protection mandate, consumers must be able to know if businesses are using their personal information to train generative AI systems.

While the emergence of GenAI is a recent phenomenon, the essential data relationships at play are precisely the kinds that voters and legislators designed the CCPA to cover. The novel quality of a technology should not undermine Californians' hard-fought consumer safeguards within these relationships.

Since its beginnings as a proposed ballot measure ("Prop 24"), the purpose of the CCPA has been to wrest control over data from large corporations and return it to the hands of its rightful owners: consumers. As then-California Attorney General Xavier Becerra told the *New York Times* when the CCPA first took effect in 2019, "[b]usinesses will have to treat that information more like it's information that belongs, is owned by and controlled by the consumer, rather than data that, because it's in possession of the company, belongs to the company."

This promise is embodied in the basic rights that the CCPA gives consumers. The most fundamental among these is the right to know under § 1798.110. Without knowing what personal information a business has collected or processed about you, a consumer cannot make an informed decision about exercising their other rights, such as the right to correct inaccurate information or the right to opt out of sale or sharing of personal information. In this way, the right to know is paramount for fulfilling the CCPA's central mandate of returning control to consumers.

The CCPA governs the data relationships between businesses and consumers even when novel technologies like GenAI reproduce such relationships; there is nothing in the law that limits its impact to only the technologies that existed at its enactment. Under § 1798.110(a), a consumer has the right to request that a business that collects personal information about the consumer disclose to the consumer the categories of personal information it has collected about them, as well as the business or commercial purpose for collecting, selling, or sharing personal information.

As described in the next subsection, GenAI business cannot survive without collecting data for training their models. This step is so important to AI market share that Nick Grudin, a vice president of global partnership and content at Meta, said at one meeting: "[t]he only thing that's holding us back from being as good as ChatGPT is literally just data volume." The use of personal information as part of an AI training dataset should therefore be considered a business or commercial purpose, which is broadly defined under § 1798.140(e) as "the use of personal information for the business' operational

purposes." Without access to personal information, a GenAI business's operations would certainly falter.

A California consumer's right to know under the CCPA should grant them the ability to know what personal data of theirs, if any, has been used to train GenAI. Back in 2019, Becerra foresaw that some consumers would use the CCPA to seek more specific information: "That consumer, so long as they follow the process, should be given access to their information. It could be detailed information, if a consumer makes a very specific request about a particular type of information that might be stored or dispersed," he told the *Times*. Requesting information about the use of one's data in training GenAI systems does not expand the scope of CCPA-required disclosure beyond what already exists—it simply asks businesses to pinpoint the data that they have processed for this particular purpose

Even if a business believes that its use of personal information to train GenAI falls under a statutory exception, the business bears the burden of explaining why the data is exempt. Under § 1798.145(h), "[i]f the business does not take action on the request of the consumer, the business shall inform the consumer…of the reasons for not taking action." The business shall also "bear the burden of demonstrating that any verifiable consumer request is manifestly unfounded or excessive." The Agency should ensure that businesses do not shift this burden onto consumers, who should not be required to justify their data requests when exercising their statutory rights.

Additionally, the Agency's ongoing rulemaking concerning risk assessment and automated decisionmaking technology (ADMT) regulations could be read to reach consumers' personal information in GenAI training datasets. If this is so, the Agency should clarify that any regulation promulgated through this process will apply with equal force to GenAI systems. Our researchers struggled to read the current draft regulations as covering GenAI systems for a few reasons, however. First, the specific examples the Agency gives of ADMT seem to describe non-GenAI systems, like rideshare apps determining driver fares, affect recognition systems used in job hiring, and facial recognition training sets. Second, in the board meeting from March 8, 2024, Neelofer Shaikh explained that "Deep Fakes" is defined in the proposed text for "operating generative models such as large language models." Deepfakes can be produced without resorting to generative models, and so the conflation of deepfakes with generative models creates unnecessary confusion about whether those rules will apply to GenAI systems broadly, especially those that do not produce deepfakes (like large language models).

The Agency can and should clarify both whether the CCPA's right to know extends to GenAI systems and whether the risk assessment and ADMT regulations will apply to GenAI systems that do not profile consumers, produce deepfakes, or facilitate automated or human-in-the-loop decision processes.

b. Under the CCPA, consumers should have the right to decide whether businesses can use their personal information to train GenAI applications.

The CCPA assures Californians that their personal data does not belong to businesses—it belongs to them. At the same time, businesses use large swathes of Californians' personal information to train their GenAI systems. Consistent with their rights under the CCPA, consumers have a right to know

how businesses are using their data as part of AI training processes to exercise the agency and control that the CCPA rightfully promises them.

It is no secret that training data includes information pertaining to specific individuals. OpenAI, for instance, acknowledges that its training information includes "personal information." Studies have also found that, given specific prompting, GenAI models will disclose personal details about individuals that fall within the statutory definition of personal information. Under the CCPA, this category includes, among other things: real names, email addresses, and other identifiers that can be reasonably linked to a particular consumer or household. Names, phone numbers and email addresses were among the personal information that a group of researchers were able to extract from GPT-2. A group of Indiana University researchers were able to elicit similar results from GPT-3.5 Turbo. In both cases, the chatbot responses appeared to be text sequences ripped verbatim from training data—a mere glimpse into the countless pieces of personal information that businesses funnel into GenAI development.

The AI training process is at once time-intensive, expensive, and opaque. An LLM like ChatGPT learns patterns from the internet's massive corpus of text, and then uses these online patterns to predict the words that will appear within certain contexts. It then forms entire text sequences based on the gathered information about which words generally follow other words. The training of one iteration of a chatbot can take place over the course of several months.

Because GenAI development requires continuously wringing fresh data from the internet's various repositories, businesses have resorted to questionable means for compiling their training datasets. Many GenAI tools use models built on data scraped from social media platforms, which means an LLM may be trained on the details of someone's Facebook profile, even if that user never imagined their information would be used in this way. GenAI training datasets have also drawn heavily from Common Crawl, an archive of raw web page data, metadata extracts, and text extracts dating back to 2008. As artist Karla Ortiz recently described at an Agency Board meeting, "[g]enerative AI companies have grossly [o]ver reached and claimed all media and data on the internet as theirs. This includes personal websites, social media forums, heck, even the U.S. government."

As this exhaustive training process rapidly drains the well of available online data, the mandate to protect consumer information becomes all the more urgent. A *New York Times* investigation found that OpenAI, Google, and Meta have waded into legal gray areas in a desperate race to excavate the internet for more data. One route OpenAI took was to develop a tool that transcribed YouTube videos and harvested this text to train their AI models, without regard for whether the data was authorized for such use; Google also transcribed YouTube videos for the same purpose. The investigation found no evidence that Google paid any attention to gathering informed consent from consumers for collection or use of their personal information as part of the automated transcription and training processes.

Because these businesses are hungry for data to bolster their training datasets, personal information is vulnerable to exploitation. The Agency therefore has an opportunity here to play a crucial role in protecting consumers against potential abuses.

Moreover, we know that consumers are concerned about GenAI and seek greater control over their personal data. Among people in the U.S. who have heard of AI, 81% said that as businesses use AI to collect and analyze personal information, this information will be used in ways that people are not comfortable with, according to a Pew Research Center report from October 2023. The survey further found that consumers believe they are better positioned than businesses to control the fate of their online data. Seventy-eight percent of respondents said they trust themselves to make the right decisions about their personal information, compared to only 21% who said they are confident that businesses who have access to their personal information will treat it responsibly.

The Agency has itself acknowledged consumers' desire to opt out of giving their data to GenAI. As Board Member Vinhcent Le said at the March 8, 2024 meeting, "I think there's a lot of dignitary reasons why you wouldn't want your information in [a] generative AI system." Given consumers' well-founded fear that AI systems may manipulate their personal information in troubling ways, it is important that the CCPA enables Californians to safeguard their data in the first instance. Before GenAI businesses have a chance to process personal information, consumers should be able to determine whether they have access to this data at all.

c. In the public debate around GenAI, consumer data rights are crucial for informing policy making and public discourse.

The question of how to regulate AI businesses is one of the most important policy issues today. The Agency is poised to meaningfully contribute to this conversation by increasing consumers' awareness that their personal data is at stake in the debate.

The Agency is no doubt acquainted with the numerous political actions currently underway or being contemplated to address developments in GenAI. Two major bills before Congress, the Generative AI Copyright Disclosure Act and the AI Foundation Model Transparency Act, seek to govern the role of GenAI in the realms of the arts and public life. The White House issued an executive order recognizing that AI "holds extraordinary potential for both promise and peril," and that harnessing AI requires "mitigating its substantial risks." And in California, Governor Newsom has signed an executive order to respond to advancements in GenAI, with an emphasis on deploying AI ethically.

It follows that voters can only have informed opinions on these political questions if they can access clear and accurate information about how the growth of GenAI implicates their personal rights. A fundamental piece of this awareness is understanding how they may have personally—and, likely, inadvertently—contributed to the creation of ChatGPT, Bard, and similar products. Each individual consumer may then decide what they want to do with this information, but in the first instance, the Agency must help them obtain it.

We urge the Agency to address this question sooner rather than later. This February, financial publications reported that OpenAI's revenue had hit $2 billion and was projected to continue growing, thanks to its planned expansions into GenAI applications for the workplace. At a time when businesses are inserting GenAI into every corner of life, the Agency can shape the future of consumer rights by holding these businesses accountable for the personal information they process.

2. Assuming the CCPA Protects Personal Information in GenAI Training Datasets, Businesses Fail to Comply with the CCPA's Right to Know

a. Businesses that Responded to Our DSARs Did Not Comply with the CCPA.

Under the CCPA, businesses are required to provide consumers with transparent access to their personal information upon request, without imposing unreasonable requirements or obstacles. Businesses navigating the intersection of consumer rights and data security must prioritize both data accessibility and protection. When it concerns personal information in GenAI training datasets, our findings suggest businesses fail to make consumer data readily available.

In response to one DSAR, Meta responded, "[t]o help us process your request, please provide examples or screenshots that show evidence of your personal information (for example, your name, address or phone number) in responses from Meta's GenAI models. Once you provide this evidence, we would be happy to investigate further." Instead of Meta proactively ensuring compliance and providing transparent access to the requested data, it placed the onus on the consumer to gather evidence of their personal information already produced in GenAI outputs and provide additional personal information to the company. If it is indeed true that training data falls within the purview of what the CCPA protects, then Meta's response is not in compliance with the Act.

After some back-and-forth emailing, Microsoft eventually responded to one DSAR by providing the researcher with some user data. However, the utility of this data was significantly hampered by its lack of clarity and comprehensibility. Microsoft delivered CSV files that lacked any explanatory notes or legends describing the data in the tables, rendering the categories of data meaningless. For example, columns such as DeviceId, Accuracy, Radius, Latitude, and Longitude were present but contained no values. Furthermore, there was a column titled "aggregation" which consistently listed "daily" as its value in every row, without any explanation of what "aggregation" means, what "daily aggregation" entails, whether other values for aggregation are possible, or what those alternatives might be. To the requesting researcher, this data also appeared to contain browser search histories, which she was not aware Microsoft may be using as training data, and the disclosure did not clarify if that was so. One machine learning researcher who examined Microsoft's files also noted that the terminology they used is not standard in the field, and that it was unclear whether any of this data was used to train any LLMs, GANs, or other machine learning systems. A comprehensive data disclosure that is truly accessible should, at a minimum, include detailed explanations of all data categories, the potential values they may contain, and the significance of these values.

This scenario highlights a critical flaw: if professionals in the field of technology law are unable to decipher their DSAR responses, how can the average citizen, whom the CCPA aims to protect, fare any better? Section 1798.130 of the CCPA states that businesses must offer clear and accessible descriptions of the personal information they collect in their privacy statement. Upon searching Microsoft's Privacy Policy, we were unable to find category descriptions. The requirement to offer clear and accessible category descriptions is crucial to ensure that all consumers, regardless of their technical expertise, can understand the data collected about them, its uses, and their rights over it.

Without such clarity, the rights provided by the CCPA cannot be meaningfully exercised by consumers, thus defeating the law's purpose to empower consumers and safeguard their privacy rights.

Similarly, with Inflection AI, we observed inconsistent responses to identical data export requests, differing only by the names of the researchers. This also contravenes the CCPA if training data is indeed covered by the regulation.The company provided divergent instructions to researchers: one was simply told to send a particular message to receive their data export, while another was redirected to the privacy policy page on the website, without specific guidance on how to actually request their data. This raises questions about the company's adherence to numerous CCPA mandates.

This lack of uniformity in response violates Section 1798.130, which mandates that businesses provide clear and accessible mechanisms for consumers to submit requests for information. When businesses provide varied instructions to consumers, they complicate a process which the CCPA aims to make straightforward and uniform. The CCPA mandates that businesses confirm receipt of verifiable consumer requests within ten days and substantively respond to them within 45 days. When consumers are confused about how they should submit DSARs, however, there is a greater chance that businesses may delay confirming receipt and substantively responding within the mandated timelines.

The CCPA also requires businesses to provide straightforward methods for consumers to exercise their rights. The principle of easy access to privacy controls and rights underpins both Sections 1798.130 and 1798.135. Inflection AI's varied responses fail to uphold this principle in its consumer interactions.

Additionally, the discrepancies in how requests are handled may violate Section 1798.125, which prohibits discriminating against consumers who exercise their CCPA rights. By providing different levels of service or guidance to consumers based on the identical requests, Inflection AI may be discriminating against different requesters. This could especially be the case if the disparity in treatment affects the ability of consumers to effectively exercise their rights under the CCPA.

If industry leaders like Meta, Microsoft, and Inflection AI can respond to our DSARs in such a subpar manner, we are concerned how other businesses might do so. Whether these responses stemmed from ignorance or choice, this is an opportunity for the Agency to provide guidance to businesses by integrating language directly addressing requests for information about AI training data into the Act so that businesses cannot point to ambiguity or lack of specificity to justify their noncompliance. We urge the Agency to establish a clear framework for businesses to follow, ensuring that all DSARs related to GenAI are handled consistently and transparently, thus upholding the principles of consumer rights and data protection embedded in the CCPA.

b. The Agency must establish procedures for requesting and responding to DSARs, ensuring both a practically feasible and straightforward process for individuals to exercise their right to know under the CCPA.

As a consumer protection agency, it is imperative for the CPPA to address gaps to ensure constituents' best interests are at the core of agency action. When a government entity fails to act

appropriately in response to public needs, it undermines the agency's duty to uphold the principles embedded in legislation designed to safeguard consumer rights.

To ensure that GenAI DSARs are handled consistently and in compliance with CCPA regulations, the Agency should develop expected standards for businesses when responding to consumer requests, including clear guidelines around verifying requestors' identities. This involves defining what constitutes a "reasonable degree of certainty" in confirming the identity of the individual making the DSAR, and clarifying what is considered a "burdensome process" to prevent businesses from using this as a loophole to deny legitimate requests. Furthermore, the agency must ensure that responses are not only individualized but also comprehensive, providing all necessary information in a format that is understandable and usable by the consumer. These efforts will not only help maintain compliance with legal standards but also demonstrate a commitment to transparency and the protection of privacy rights. By adopting such standardized procedures, the Agency will better equip itself to address the complexities associated with managing personal data within the realm of GenAI.

3. Recommendations for the Agency

First, we ask the Agency to clarify under which authorities consumers have the right to know what personal information businesses have been using to train their GenAI systems. We believe, for the reasons stated above, that the CCPA already supports this right. If the Agency agrees with this reading, we ask that it state so in an accessible manner and provide clear requirements for businesses responding to DSARs for this information. and initiate steps to educate consumers and businesses alike on how to exercise and fulfill the right.

The Agency should provide guidance to businesses on how to disclose GenAI training data information to requestors and require businesses to do so in a uniform, standardized fashion. By publishing clear compliance guidelines, the Agency can ensure businesses are not excluding GenAI from their CCPA compliance. The responses to our researchers' requests indicate, at the very least, an inconsistency across compliance efforts, and we hope our comments start a conversation around the kind of standards that the Agency aims to establish.

Such standards should, at a minimum, require businesses to provide personal information in a legible way, describing it in simple language written for human requestors, not machine processes. The Agency can also require businesses to disclose how they collected that information, how they store it, and how it was or continues to be used in model training. As opposed to Meta's current approach, the Agency can make it clear that businesses cannot make disclosures contingent on proof that a requestor's personal information appeared in a GenAI output, which artificially limits the scope of right to know where such information is used in training but does not appear in outputs. The Agency can consider how much additional information requestors must provide to businesses to verify DSARs, and how businesses may use existing consumer information to verify requests instead.

The Agency may also consider interpreting the definitions of "personal information" and "publicly available information" to ensure that businesses that scrape the internet to build training datasets still meet their statutory obligations. Section 1798.140(v) of the CCPA defines "personal information" as

not including "publicly available information," which includes "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." This definition should not allow businesses to skirt their CCPA responsibilities simply because a consumer has published personal information on a public social media platform, such as LinkedIn. If a business believes that a piece of information is "publicly available," they should have to explain why that piece of information is exempt from the disclosure requirement.

Additionally, the Agency can inform Californians of their right to request their personal information used in GenAI training. In existing consumer-facing resources, such as the privacy.ca.gov website, the Agency can append guides for making DSARs specific to GenAI, provide example DSAR templates, and/or explicitly include GenAI training data as one of the categories for which consumers can request information.

Second, if the Agency alternatively believes that consumers' right to know if their personal information is being used to train GenAI falls outside the scope of the CCPA, we ask the Agency to establish alternative avenues for consumers to exercise control over this data. One way is to clarify that the draft ADMT regulations apply to GenAI systems broadly and extend consumers' data access rights to GenAI training data. The Agency can also initiate an informal rulemaking or comment period to allow stakeholders to weigh in and further inform the Agency on this specific issue and recommend next steps. We know there is public interest in this question, as community members have spoken out about it during recent board meetings.

In either case, we ask that the Agency specify verification as well as response standards with training data for GenAI in mind to ensure that both consumers and businesses know what to expect. With clear guidelines in place, consumers will feel more confident in their ability to navigate the process of accessing their personal data used in AI systems. Businesses will have the necessary guidance to ensure compliance with CCPA regulations regarding GenAI. The Agency can empower both consumers and businesses to navigate the complexities of GenAI DSARs in a manner that is consistent, transparent, and compliant with the CCPA.

4. Conclusion

Our DSARs revealed significant gaps in business compliance with the disclosure requirements under the CCPA's right to know. Despite the statute's clear mandate to provide transparent access to personal information, responses from leading California businesses fell short, pointing to a need for additional guidelines and enforcement mechanisms.

We strongly urge the Agency to determine that the CCPA grants Californians the right to know whether their personal information has been used to train GenAI systems. We further call upon the Agency to establish standardized procedures for handling GenAI DSARs. By taking decisive and timely action, the Agency can uphold the integrity of the CCPA and fulfill its mission of empowering consumers to exercise their data privacy rights, no matter the technology involved.

***

Thank you for the opportunity to provide feedback during the Agency's rulemaking processes. If the Agency has any further questions, please reach out to Melodi Dincer (dincer@law.ucla.edu) or Jason Schultz (jason.schultz@exchange.law.nyu.edu).

Sincerely,

Melodi Dincer
Resident Fellow & Lecturer
UCLA Institute for Technology, Law and Policy (ITLP)

Michael Karanicolas
Executive Director
UCLA Institute for Technology, Law and Policy (ITLP)

Jason Schultz
Professor of Clinical Law & Director
NYU Technology Law & Policy Clinic

Mike Ananny
Associate Professor
USC Annenberg School for Communication and Journalism

Hamsini Sridharan
Doctoral Candidate
USC Annenberg School for Communication and Journalism

Sarah Ciston
Independent Researcher
(formerly USC Annenberg School for Communication and Journalism)

APPENDIX A

General DSAR Template

To:	support@company.com
Subject:	Data Subject Access Request

To Whom It May Concern:

My name is [researcher name]. I reside in California and am exercising my right under the California Consumer Privacy Act to seek access to the personal information [insert name of company/specific software here] (hereinafter, "you") collects and/or has collected about me and how it is used, shared, and/or sold, whether directly from and/or through me, a third party, an API, and/or a service provider, including but not limited to:

1. The categories and/or specific pieces of personal information you have collected about me;
2. The categories of sources from which you collected that information;
3. The purposes for which you use that information;
4. The categories of third parties with whom you disclose the information; and
5. The categories of information that you sell or disclose to third parties.
5. The categories of information that you sell or disclose to third parties.

My request includes, but is not limited to:

- All data (including data within or associated with text, image, sound, video, or other media) about me that has been included in any development, training and/or improvement of any Large Language Model (LLM), Generative Adversarial Network (GAN), Diffusion Model, and/or any similar system;
- All data (including data within or associated with text, image, sound, video, or other media) about me that has been stored, memorized, retained, or integrated in any other way within any LLM, GAN, Diffusion Model, and/or any similar system;
- All data (including data within or associated with text, image, sound, video, or other media) about me that has been included, either in whole or in part, in any output of any LLM, GAN, Diffusion Model, and/or any similar system;
- All data (including data within or associated with text, image, sound, video, or other media) about me that has been used as part of any Reinforcement Learning from Human Feedback (RLHF) process.

My email address is [insert email here] and phone number is [insert phone number here]. If you need any more information from me, please let me know as soon as possible.

If you cannot comply with my request–either in whole or in part–please state the reason(s) why you cannot comply.

If part of the information is subject to an exception, please state the name and basis of your application of the exception, and to which part of my request you are applying it.

If my request is incomplete, please provide me with specific instructions on how to complete my request.

Sincerely,
[Insert Name Here]

APPENDIX B

Meta's Response Requesting Proof of the Use of Personal Information

---------- Forwarded message ---------
From: **Facebook** <case++aazqhacjslyjac@support.facebook.com>
Date: Mon, Sep 18, 2023 at 8:05 AM
Subject: Generative AI Data Subject Rights #[redacted]
To: [redacted]

Hi W[redacted]

Thank you for contacting us.

Based on the information provided, we were unable to process your request. To help us process your request, please provide examples or screenshots that show evidence of your personal information (for example, your name, address or phone number) in responses from Meta's generative AI models. Once you provide this evidence, we would be happy to investigate further.

If you have any questions about how Meta uses information from our products and services, please see our Privacy Policy: https://www.facebook.com/privacy/policy

To learn more about generative AI, and our privacy work in this new space, you can review the information we have in Privacy Center: https://www.facebook.com/privacy/genai

Thanks,
Privacy Operations

APPENDIX C

Samples from CSV Files Provided by Microsoft in Response to a DSAR

DateTime	DeviceId	AccuracyRadius	Latitude	Longitude	Time	Title
5/5/2023 5:48:19 PM +00:00					5/5/2023 5:53:06 PM +00:00	Intel AX201 Wi-Fi 6 is not working on Ubuntu 21.04
5/5/2023 5:48:19 PM +00:00					5/5/2023 5:52:47 PM +00:00	Missing WiFi adapter Intel AX201 on Ubuntu 21.04
5/5/2023 5:48:19 PM +00:00					5/5/2023 5:48:43 PM +00:00	Linux* Support for Intel® Wireless Adapters
5/4/2023 4:25:53 PM +00:00					5/4/2023 4:26:09 PM +00:00	How To Dual Boot Linux and Windows 11 | Tom's Hardware
5/4/2023 4:11:14 PM +00:00					5/4/2023 4:11:40 PM +00:00	How to Access the Boot Menu in Windows 11 - How-To Geek
5/4/2023 3:44:01 PM +00:00					5/4/2023 3:49:50 PM +00:00	Fix: Why Isn't Linux Detecting My Wi-Fi Adapter? - How-To Geek
5/4/2023 3:44:01 PM +00:00					5/4/2023 3:49:42 PM +00:00	How To Install WiFi Drivers In Linux Mint – Systran Box
5/4/2023 3:44:01 PM +00:00					5/4/2023 3:44:07 PM +00:00	HOW TO INSTALL WIFI DRIVER IN LINUX MINT 19
5/4/2023 3:43:59 PM +00:00						
5/4/2023 2:04:44 PM +00:00						
5/4/2023 2:02:55 PM +00:00						
5/4/2023 2:01:25 PM +00:00					5/4/2023 2:02:41 PM +00:00	Kali Linux vs Linux Mint detailed comparison as of 2023 - Slant
5/4/2023 2:01:25 PM +00:00					5/4/2023 2:01:43 PM +00:00	Linux Mint as a server? - Linux Mint Forums
5/4/2023 2:01:02 PM +00:00						
5/4/2023 1:56:35 PM +00:00						
5/4/2023 1:56:35 PM +00:00						
5/4/2023 1:50:40 PM +00:00					5/4/2023 1:57:36 PM +00:00	How to Find a Windows 10 or 11 Product Key | Tom's Hardware
5/4/2023 1:50:40 PM +00:00					5/4/2023 1:50:46 PM +00:00	3 simple ways to find your Windows 10 product key
5/4/2023 1:41:04 PM +00:00					5/4/2023 1:41:15 PM +00:00	Beelink Forum
5/4/2023 1:39:34 PM +00:00					5/4/2023 1:39:48 PM +00:00	Create installation media for Windows - Microsoft Support

DateTime	EndDateTime	DeviceId	Aggregation	AppName	AppPublisher
9/5/2023 12:00:00 AM +00:00	9/5/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
9/5/2023 12:00:00 AM +00:00	9/5/2023 11:59:59 PM +00:00		Daily	Office Shared Components	Microsoft Corporation
9/5/2023 12:00:00 AM +00:00	9/5/2023 11:59:59 PM +00:00		Daily	account.microsoft.com	Microsoft Corporation
5/5/2023 12:00:00 AM +00:00	5/5/2023 11:59:59 PM +00:00		Daily	Bing	Microsoft Corporation
5/5/2023 12:00:00 AM +00:00	5/5/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
5/5/2023 12:00:00 AM +00:00	5/5/2023 11:59:59 PM +00:00		Daily	Microsoft Edge	Microsoft Corporation
5/5/2023 12:00:00 AM +00:00	5/5/2023 11:59:59 PM +00:00		Daily	MSN Web	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	Office Shared Components	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	Microsoft Cortana and Windows Search	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	Bing	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	Microsoft Edge	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	Microsoft Store	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	MSN Web	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	account.microsoft.com	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	learn.microsoft.com	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	support.microsoft.com	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	support.microsoft.com	Microsoft Corporation
5/4/2023 12:00:00 AM +00:00	5/4/2023 11:59:59 PM +00:00		Daily	techcommunity.microsoft.com	Microsoft Corporation
5/3/2023 12:00:00 AM +00:00	5/3/2023 11:59:59 PM +00:00		Daily	Office Shared Components	Microsoft Corporation
5/3/2023 12:00:00 AM +00:00	5/3/2023 11:59:59 PM +00:00		Daily	Microsoft Cortana and Windows Search	Microsoft Corporation
5/3/2023 12:00:00 AM +00:00	5/3/2023 11:59:59 PM +00:00		Daily	Microsoft Word	Microsoft Corporation
5/3/2023 12:00:00 AM +00:00	5/3/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
4/17/2023 12:00:00 AM +00:00	4/17/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
4/16/2023 12:00:00 AM +00:00	4/16/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
4/15/2023 12:00:00 AM +00:00	4/15/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation
4/14/2023 12:00:00 AM +00:00	4/14/2023 11:59:59 PM +00:00		Daily	OneDrive	Microsoft Corporation

APPENDIX D

Varying Responses to Identical Requests Submitted to Inflection AI

Details of How to Formally Request Data

From: **Aiyappa N U (Pi.ai)** <support@inflectionai.zendesk.com>
Date: Thu, Aug 31, 2023, 3:22 PM
Subject: [Pi.ai] Re: Data Subject Access Request for Data Pertaining to [redacted]
To: [redacted]
Cc: Privacy <privacy@pi.ai>

Hello H[redacted] S[redacted]

Thank you for reaching out to us regarding our privacy terms.

You can find our privacy policy on our website by clicking on the "Privacy" link at the bottom of the page.[https://heypi.com/policy#privacy]

If you have any further questions or concerns about our privacy policy, please don't hesitate to reach out to us. We are always here to help.

Thanks and Regards,

NU Aiyappa

Referral to Privacy Policy

On Sep 7, 2023, at 7:03 PM, [redacted] (Pi.ai) <support@inflectionai.zendesk.com> wrote:

Hello M[redacted] A[redacted]

Thank you for your request to export your data. To proceed with the export, we kindly ask you to provide us with your Name and send a message with the phrase "EXPORT MY DATA [5401]" to Pi as confirmation. Once you confirm, please follow up here and let us know so we can start the export process.

Thank you for choosing our services.