Stanford University
Human-Centered
Artificial Intelligence

Issue Brief
HAI Policy & Society
April 2026

Data Privacy and Foundation Models: Can We Have Both?

Jennifer King and Tiffany Saade

Introduction

Imagine receiving a security alert from your bank: A fraudster cloned your voice and used it to bypass the bank's digital security measures and empty your bank account. The tool they used? A generative AI model trained on publicly available data, cloning your voice with an old YouTube video you'd forgotten was online. Or consider prompting a chatbot to tell you what it knows about you, and it surfaces deeply personal information gleaned from pseudonymous posts you once made online.

These examples underscore the profound privacy challenges posed by foundation models — large-scale, general-purpose AI models that stand apart in their ability to impact society globally and at scale. These models are the literal foundation upon which large-scale AI is being integrated into countless consumer-sector digital products and services.

In this issue brief, we examine the risks to data privacy, from both individual and systemic perspectives, posed by the training and use of consumer-focused foundation models. Because foundation models are dependent on massive datasets for their development, they pose a broader set of privacy risks than smaller AI systems trained on proprietary or limited datasets. Foundation models may thwart data privacy not only by the

Key Takeaways

Foundation models pose unprecedented and largely unaddressed privacy risks that are broader and harder to address than those posed by traditional AI systems.

These risks emerge across the entire model life cycle — from the mass scraping of personally identifiable information during training, to the memorization and regurgitation of sensitive information in model outputs, to the intimate data that users unwittingly disclose through chatbot interfaces.

Foundation models are also vulnerable to adversarial attacks, including prompt injection, data poisoning, and model inversion, that can circumvent privacy safeguards and expose sensitive personal information.

Existing privacy frameworks, including the EU's GDPR, are fundamentally incompatible with how foundation models are built, yet neither the EU nor the United States has enacted comprehensive rules that could meaningfully change developer behavior.

Without clear regulatory guardrails, the public remains largely dependent on developers to voluntarily protect their privacy. Policymakers must weigh a range of governance mechanisms that require removing personal data from the training data pipeline, increase model transparency, ensure the creation of systems that protect privacy by design, and constrain privacy-infringing model outputs.

use or misuse of the technology itself, but also by the process of building and training them. In addition, they are vulnerable to privacy risks from adversarial attacks. While the risks can be mitigated, without regulatory rules in place, the public is largely reliant on developers to do the right thing to protect the public's privacy, which unfortunately is not always the case.

To reconcile data privacy with foundation models, policymakers should weigh a range of governance mechanisms that ensure the removal of personal data from the training data pipeline, require system architectures that prioritize privacy and data security protections by design, increase the transparency and interpretability of foundation models and their training data, and constrain the outputs of models. Policymakers must confront the many ethical and legal questions over access to and control of personal data as AI model adoption continues to grow.

Privacy Risks During Data Collection and Curation: The Raw Material Underpinning Foundation Models

The raw material upon which foundation models are built is data — vast, unimaginable quantities and types of data. Some models are trained on a single data modality (text, audio, video, still images, etc.), but increasingly foundation models are trained on multiple or all modalities.

Because developers are reluctant to disclose their sources of training data, it is challenging to identify the full scope of sources that are included. Researchers

Because developers are reluctant to disclose their sources of training data, it is challenging to understand the full scope of sources that are included.

generally assume that foundation models include both publicly available and paywalled data. Publicly available data encompasses data included in existing open datasets (e.g., LAION, ImageNet, WordNet, Common Crawl) as well as data scraped from websites across the internet. To build large language models (LLMs), for example, AI developers have already scraped so much English language data (including copyrighted data) from publicly available sources that, by some accounts, they have exhausted the available supply, inspiring developers to engage in creative albeit legally risky strategies to acquire more. Some foundation model developers, large platforms like Google or Meta, also utilize their own proprietary data, including their users' personal data, for model training. Other training sources include data obtained from data brokers or licensed from third parties, such as Google's licensing of Reddit's content. Finally, developers also include users' ongoing interactions with their models, such as users' chats with chatbots, both to improve their performance and train future versions.

The immense volume of data collected to train foundation models means that personal data is inevitably included in these training sets, absent developer efforts to

remove or proactively exclude it. Researchers have identified personally identifiable information (PII) and sensitive information like U.S. Social Security numbers or breached datasets in some training data. Even if we assume that the personal data included in training datasets was acquired exclusively from publicly available sources — for instance, from personal webpages, government websites, or user-generated, public-facing content sites like Reddit — the collection of such vast amounts of data still poses risks to our data privacy. These risks emanate from the scope and scale of this data collection, which makes information that previously existed in obscurity available to AI models.

It's not unusual for people to forget that they've left personal data across the web that can be accessed by anyone. For example, researchers at the University of Washington created the MegaFace database, which made available over 4 million facial images for training facial recognition systems, in part by repurposing decades-old images from the photo-sharing site Flickr.com without the image posters' explicit consent. In 2019, New York Times reporters were able to trace images of young children in MegaFace to their original posters on Flickr, many of whom were parents of the children, and none of whom were aware that the images were in the MegaFace database and being used to train commercial AI facial recognition systems. Some of the children in the photos, now teenagers or older, expressed frustration that their images were being used without their consent or control.

This example and many others like it illustrate how data repurposing for model training not only violates the original context in which the data was disclosed but also individuals' ability to consent — or not consent — to using their personal information for commercial profit. Individuals also have no control over their data once it has been used, with limited to no mechanisms for correction or deletion. What's more, the vast majority of people are completely unaware their personal data is being used for these purposes, making it difficult to impossible for them to object or opt out.

Privacy Risks During Model Training and Inference: Memorization and Model Outputs

While the presence of one's personal information in training data does not guarantee that a foundation model will generate output that reveals it, foundation models have been shown to disclose PII and other sensitive personal information via their outputs.

It is difficult to examine any single data point and interpret its impact on a foundation model's weights and capabilities. Foundation models make inferences and predictions based on a multitude of sources, and the size of these models renders this process difficult, if not impossible, for even model developers to clearly interpret. Predictions about individuals, for example, may range from simple (e.g., an individual's date or place of birth based on public sources) to complex (inferring someone's actual identity from their pseudonymous social media posts). Lifted from the original context in which it was shared, data may enable generative systems to help users reverse-engineer an individual's identity, or to infer details about an individual that go beyond descriptive personal data, such as sexual orientation or political views. It is possible that a model that outputs PII has generated it via inferences it made based on other data, such as predictable facts drawn from demographics, or even derived from relational data from other individuals' data in the training set.

However, models have also been shown to memorize data that was incorporated in training datasets and regurgitate it verbatim. In particular, high-entropy data — data that is uncommon — may be more likely to be memorized. Models also exhibit recency effects, where data seen later in the training process is more likely to be memorized. This is problematic because the process of fine-tuning models, particularly on proprietary datasets (e.g., to optimize the performance of the model for the user's specific context), occurs at the latest point in the development cycle. As a result, individuals' sensitive personal information can be exposed as part of outputs generated by the models unless explicitly disallowed by developers through output "guardrails."

Beyond these immediate privacy harms, the outputs of foundation models can also cause a range of serious societal harms that can be traced back to data collected or inferred about individuals during training.

First, decisional harms — losses of opportunity, liberty, and autonomy — arise when foundation models trained on historically biased data reproduce and embed that bias at scale, including in high-stakes contexts. For example, LLMs used for résumé screening have been found to prefer résumés with white-associated names, while LLMs used for administrative tasks in healthcare settings reproduce gender biases and thereby exacerbate inequities in care provision. Second, economic harms emerge when AI systems produce racially or socioeconomically disparate outcomes. Researchers have found that LLMs used for mortgage applications, for instance, consistently recommend denying more loans and charging higher interest rates to Black applicants compared to otherwise identical white applicants. Third, social and dignitary harms — damage to one's sense of self, social standing, or community belonging — arise when models classify and sort individuals in ways that are overtly or covertly discriminatory, derogatory, and difficult to contest. For example, major U.S. foundation models continue to associate speakers of African American English with archaic negative stereotypes.

As foundation models are increasingly embedded in high-stakes decisional contexts, the aggregation of personal data — and the inferences drawn from it — risks institutionalizing discrimination at an unprecedented scale.

Privacy Risks During Model Use: Interactions and User Inputs

It is impossible to fully anticipate the emergent risks to privacy from the multiplicity of ways that we can interact with foundation models. But one primary layer of risk emerges from how users engage with these models via the interfaces of AI chatbots powered by them. AI chatbots are designed to mimic human conversation and to be excessively flattering and agreeable. This conversational design encourages users to disclose vast amounts of personal and sensitive information: The more a user engages, the more data is collected. While this risk is most acute for users who actively seek health support or have formed parasocial relationships with a chatbot, even those seeking general advice or coding support may inadvertently reveal PII and other sensitive information about themselves or other individuals.

This is concerning from a privacy perspective because chatbot developers currently face little to no oversight when it comes to their handling of increasingly personal and sensitive data provided by their users through chat interfaces. And as foundation model developers build AI agents capable of automating personal tasks, the

Chatbot developers currently face little to no oversight when it comes to their handling of increasingly personal and sensitive data provided by their users through chat interfaces.

incentives to gather and repurpose consumer data across multiple contexts are only growing. Many of them are already exploring different ways to monetize user data: Meta, for example, is repurposing such data to sell targeted ads across its social media platforms, and as of March 2026 OpenAI is testing an ad model that would see it use chat conversation data as context to serve targeted ads within the platform. In short, we are likely to see the existing practices of gathering personal and behavioral online data for targeted advertising replicated in these emergent systems.

Compounding this concern, model developers are not transparent about whether or how they mitigate any associated privacy risks. A recent study co-authored by this brief's authors found that six major U.S.-based LLM developers all default to using customer chat data for model retraining. While some developers offer opt-out mechanisms that allow users to decide that they do not want their chatbot inputs to be repurposed for model retraining, this is not universal across systems: At least two major AI chatbots offer no such opt-out mechanism. Model retraining can extend to any files uploaded by the users, including photos, voice recordings, and documents, and many developers appear to retain such chat data indefinitely. Yet information on whether developers remove identifiable personal information before retraining is sparse. This means that highly sensitive personal information continues to be fed into foundation model training datasets and is raising the stakes of LLM memorization.

Emergent chatbot features also raise privacy concerns. For example, most chatbot platforms have a memory feature that allows users to save their chats to reference later. Increasingly, users are able to personalize their chat interactions by having a chatbot persistently remember preferences and facts about them. As these features evolve and mature, they raise crucial questions about how they function behind the scenes: What is the scope of the time or data they include; where and how is related data stored; could the data be leaked or used for other purposes (e.g., generating detailed behavioral profiles about users, or inferring one's psychological state); are the features subject to data rights requests; and do they respect specific contexts (e.g., work-related versus personal chats)? Law enforcement authorities are especially interested in chat data, which can be subject to criminal legal requests. While this data is similar in scope to the aggregation of our online search queries, the intimate nature of chatbot interactions may result in a qualitative difference where we reveal far more emotional and psychological information to chatbots than can be gleaned from our search histories.

Adversarial Privacy Threats to Foundation Models

In addition to the privacy risks associated with the training and use of foundation models, the models themselves are vulnerable to adversarial attacks that can create additional privacy harms. During

the model training and testing phases, malicious actors can corrupt training data, the model itself, or both to expose personal data. While open and closed foundation models are vulnerable to such attacks, some argue that open models pose a higher risk because malicious actors can gain a deeper understanding of their key features since model weights and architecture are documented publicly. Vulnerabilities at the model level could have widespread and unpredictable downstream privacy consequences since foundation models are embedded throughout the larger AI ecosystem.

During the model training and testing phases, malicious actors can corrupt training data, the model itself, or both to expose personal data.

Prompt Injection: One of the most consequential attack vectors on foundation models has proven to be prompt injection. By inserting malicious instructions or external content in user inputs, attackers can override system prompts, manipulate model behavior, and induce the model to reveal hidden or sensitive information. Research demonstrates that such attacks can lead to "prompt leaking," where models disclose hidden system prompts or confidential instructions. Subsequent work has shown that in real-world LLM-integrated applications — particularly those using retrieval augmented generation (i.e., accessing online search tools) or external APIs — indirect prompt injection can coerce models into exfiltrating data from connected systems or private documents, because the model interprets malicious content as legitimate instructions. The privacy implications of prompt injection attacks are alarming: They can trick foundation models into bypassing traditional security and privacy guardrails to reveal private information.

Data Leakage and Exposure: A key privacy vulnerability of foundation models is the potential for malicious actors to leak private information from training data. Model inversion attacks allow attackers to extract sensitive training data directly from a trained model by reconstructing it from model responses, while membership inference attacks allow them to infer whether a specific data point is part of the model's training dataset. Other methods, such as extractable memorization, allow an adversary to extract elements from the model's training data without prior knowledge of the dataset — a threat that can't be eliminated using current model alignment techniques. All of these techniques can expose sensitive health, biometric, financial, or otherwise personal information, undermining efforts to anonymize data.

Data and Model Poisoning: Foundation models are vulnerable to *data poisoning*, which occurs when malicious actors train a model on corrupted data (e.g., by inserting malware or intentionally biased data directly into training data or webpages that are eventually scraped for training purposes) and thereby alter its outputs in downstream applications. Such attacks can result in decreased model accuracy but also lead to breaches of confidentiality. Research has shown that poisoning even a small fraction of a training dataset can make inference attacks and data extraction more effective and thereby cause significant leakage of sensitive personal information. While high-profile poisoning attacks have so far been limited primarily to the academic and red-teaming sphere, one

could imagine a data poisoning attack occurring at scale against an individual or organization as part of a coordinated campaign, similar to past Google bombing attacks that manipulated search results. Adversaries can also conduct *model poisoning* by compromising the training procedures and model parameters themselves, enabling the attacker to trick the model into revealing PII. Sophisticated model poisoning techniques can also compromise differential privacy mechanisms, thereby weakening model privacy guarantees.

Removal of Safety Guardrails: Adversarial efforts that directly compromise a model's safety guardrails — the last line of defense for privacy — can also introduce broad privacy risks. For example, researchers have demonstrated that it is possible to compromise the safety guardrails of OpenAI's ChatGPT 3.5 Turbo model by fine-tuning it on explicitly harmful data points. Even when there is no malicious intent, the very act of fine-tuning a model — even with "clean" datasets — could distort LLMs' safety alignment. Compromised safety guardrails reveal existing behaviors previously suppressed and could turn models into effective tools for personal data extraction, inference, and re-identification.

Reconciling Data Privacy with Foundation Models

The development and training of foundation models presents serious challenges to existing data privacy paradigms. For example, they challenge the leading global privacy framework — the European Union's General Data Protection Regulation (GDPR) — which enshrines the principles of data minimization and purpose limitation that have been part of the bedrock of data protection for over 50 years. These principles impose limits on how much and what kinds of data can

The development and training of foundation models presents serious challenges to existing data privacy paradigms.

be collected by data processors based on their stated purpose, and they aim to contain a runaway environment where data is collected by companies or governments about individuals without boundaries or consent.

Fundamentally, current methods for building foundation models are incompatible with these principles. EU data protection regulators have conducted multiple investigations into AI developers, raising questions about their compliance with the GDPR, including their use of personal data scraped from publicly accessible posts, their lack of individual notice and consent mechanisms, and their completion of legally required data protection impact assessments. Numerous decisions by EU regulators suggest that foundation model developers will no longer be able to freely scrape all available data for AI training without first considering privacy protection. Even in the United States, which lacks an omnibus consumer privacy or data protection regulation, the Federal Trade Commission (FTC) has launched investigations into multiple developers of purpose-built AI systems over their data collection practices, including their use of personal information obtained via chatbot interfaces.

However, despite these regulatory actions and inquiries, the major data privacy challenges that foundation models pose remain unresolved. Neither the US nor the EU has issued comprehensive legislation or directives that have

significantly changed the data collection behaviors of AI developers. While EU regulators have clarified that legitimate interest is a valid legal basis for training AI models, there remains uncertainty regarding exactly how models will be required to maintain anonymity when personal data is included in training data. The future of a proposal to update the GDPR to weaken the current definition of personal data also remains uncertain. In the United States, the FTC has demanded that developers delete both data and models in instances where companies have misled customers about the use of their data in algorithmic systems. While this approach is unlikely in the case of scraped personal data used for model training, given how much of that data is often public or publicly accessible, data collected directly from users could be subject to such a remedy if the developer uses or reuses the data in ways they did not specify at the time they collected it.

There are many other governance mechanisms regulators should weigh as they seek to reconcile data privacy with foundation models:

1. Reduce and Remove Personal Data from the Training Data Pipeline

Policymakers must enforce privacy protections during the upstream phases of AI development, namely data collection and model training. By focusing on reducing or eliminating personal data from entering the model development pipeline in the first place, policymakers can defend the core privacy principles of data minimization and purpose specification. Doing so will also reduce the opportunity for privacy harms to individuals downstream as well as reduce potential liabilities for developers.

A crucial first step is to enact limitations on the types of data developers can collect or use for model training. In the United States, enacting a federal data privacy law that provides meaningful restrictions on the scope and amount of data companies can collect from consumers across all online contexts, and limits the collection and sales by third parties such as data brokers, is a necessary foundation. Current efforts to reach bipartisan consensus on such a privacy law are at an impasse. In the meantime, policymakers must explore alternative policies to hold developers accountable without forcing the public to shoulder the burden of determining whether their data is being used for model development.

First, policymakers should introduce oversight measures that require developers to remove specific types of personal data included in public scrapes, even if the data isn't explicitly covered by data privacy statutes. Methods such as approximate deletion and machine unlearning have recently attracted attention from developers and policymakers as promising options for data removal from models. However, research has uncovered significant limitations to these approaches, cautioning that the only certain way to remove specific data from a model is to retrain the model without the data. While this is currently an immense and expensive undertaking, research suggests that the process may eventually become more efficient and less high stakes, meaning that these priorities should not be abandoned even if they are impractical today.

Second, policymakers should require companies to limit the collection of data from user interactions with models. Companies should ask individuals to explicitly opt *in* to having their interactions with foundation models repurposed for model training, moving away from the current default that forces users to actively opt *out* if they don't want their data used for model training. Developers should continue to introduce privacy-forward features such as temporary, private chat modes and context-specific protections so that users can conduct sensitive conversations without their transcripts being retained and associated data used for model training. If users opt in to allow their chats to be used for model training, developers should proactively filter and remove PII and other sensitive data prior to using the chats for training.

It is also essential to identify and plug the loopholes that enable the widespread buying, selling, and sharing of consumer data by third parties, particularly data brokers. While brokered data may not currently be a primary source for model training, it is used by many model developers to enhance the data they collect from users, and to build out profiling and ad targeting capabilities. But importantly, there must be automated, scalable methods for the public to exercise these preferences. California, for example, has created an automated opt-out platform for residents to submit data rights requests that all data brokers operating in the state must honor by law. As more Californians remove their data from brokerages, they reduce the vectors by which their personal data can be commodified without their explicit knowledge or consent. Unless these choices are simple and automateable, consumers will continue to bear the burden of one-off requests along with the negative impacts of using their data.

2. Increase Transparency and Interpretability

Demanding transparency and interpretability from frontier model developers for consumers should be a priority for policymakers.

To date, public discussions on AI transparency have been dominated by questions of existential risk and model safety. But increasing transparency throughout the development pipeline with regard to the collection and use of data is also crucial. At an individual level, the current lack of transparency and interpretability puts users on an uncertain footing and prevents them from understanding and anticipating the privacy risks of sharing personal and potentially sensitive data (e.g., health data) through model interfaces. For example, most users today are unaware that the data they provide to foundation model developers may be viewed by human workers, typically outside the U.S., hired to provide feedback on model outputs; in some cases, these workers have reviewed deeply sensitive and revealing personal data, including private videos. At a systemic level, the inclusion of personal data in the construction of foundation models challenges existing data protection principles, especially data minimization and purpose limitation.

Policymakers can only develop privacy protections that guarantee foundation model developers will not exploit or misuse consumers' personal data if they — together with researchers and the general public — can gain detailed information about what data developers collect and how they use and protect it in their models. As we found in our research examining chatbot privacy policies, major foundation model developers' privacy policies omit or are unclear about substantive practices relating to their AI chatbots, despite requirements by the California Consumer Privacy Act that such policies be comprehensive in describing their data practices. The success of data removal requirements also hinges on greater transparency: Even if developers were to successfully remove personal data from their training datasets, questions remain regarding whether their models can still predict the removed information from other data sources.

Regulators must therefore enact data transparency policies that require the disclosure of sources of all training data, both external (scraped) and internally sourced data, including from user interactions with AI products. At a minimum, developers should be required to remove PII and other sensitive personal data from training data and assess the impact of including PII in model training on inferences and outputs. They should also provide a clear process for users to request the deletion of this data, particularly in geographic regions that lack data deletion rights.

Crucially, researchers stress that data transparency policies must demand a high level of specificity in disclosures, must be designed to change the

data collection behaviors of developers, and must include clear enforcement mechanisms. Existing data transparency policies, such as the recently enacted California Assembly Bill 2013 (AB 2013), fail to provide specific enforcement provisions and have led to developers providing short and vague data transparency reports that may be technically compliant but are hardly actionable. Based on the outcome of AB 2013, policies designed to promote transparency between consumers and developers should strive to be specific in their requirements, informative to the intended audience, and enforceable. Though the implementation of these measures requires legitimate considerations — such as technological feasibility or the revealing of trade secrets — their absence may result in transparency policy that is neither meaningful nor successful.

3. Adopt Privacy and Security by Design

Another approach to minimizing privacy risks and harms is to incentivize foundation model developers to create a system architecture that inherently embraces privacy and data security protections in the system's design, from the technical architecture to the user interface.

Already, model providers have missed this mark. In summer 2025, OpenAI was discovered to have posted chats online that users had set to share publicly so they would be indexed by search engines. After considerable privacy-based objections by the public, the company deindexed the chats and discontinued the share feature. Similarly, many users of Meta's AI app mistakenly shared their private chatbot conversations to Meta's public feed due to a new and unclear user interface design. These examples demonstrate the fundamental mismatch between how developers think users should interact and share data with their chatbot products and what users actually expect and want. A privacy-by-design approach would identify and prioritize users' expectations regarding their chat conversations, building and facilitating trust instead of putting the developer's desires for product adoption first.

There is evidence that developers are responding to how their customers are using their models in the wild. After numerous instances of users uploading medical records and bills to get help with diagnoses or contesting medical billing, OpenAI introduced a context window for health-specific conversations with ChatGPT, while Anthropic introduced HIPAA-ready enterprise tools to enable healthcare-related uses of their chatbot, Claude. As foundation model developers begin rolling out advertising in their chat tools, many of the same questions about behavioral targeting that persist across the internet will resurface.

In addition to interface designs that support individual privacy expectations, there are several privacy-enhancing techniques that can help build technical privacy protections into AI models at the system level. Federated learning — a method for training machine learning models in a decentralized manner through remote servers — ensures that user data remains stored on local devices (such as an individual mobile phone), thereby reducing the risk of data exposure and privacy breaches. Another privacy-preserving technique, differential privacy, adds statistical noise to training data that obfuscates individual data points to prevent reidentification from training data. While technical privacy measures are critical, research has shown that even when these techniques are used, models can still be vulnerable to privacy attacks, and they may create a tradeoff between model accuracy and performance. Policymakers should therefore not treat these techniques as easy fixes but rather as additional protective layers that aren't sufficient on their own.

4. Suppress Outputs

Constraining the outputs of foundation models by suppressing the output of specific data types (e.g., phone numbers, social benefit IDs) or generated

Policymakers should not treat [technical privacy measures] as easy fixes but rather as additional protective layers that aren't sufficient on their own.

personal information (whether accurate or not) is another means to reduce privacy risks. However, policymakers should view this method as a supplementary line of defense, rather than a holistic form of privacy protection.

Some researchers view output suppression as a pragmatic alternative to stricter approaches such as model retraining or deletion. As discussed above, removing specific data from trained models does not guarantee that a model will not be able to generate data based on other sources or inferences. Requiring developers to put in place output-suppression guardrails that prevent their models from sharing or generating personal information could therefore be an important policy lever.

However, while such guardrails may suppress certain known types of outputs, they can't feasibly account for the potentially infinite unknown outputs that, in certain contexts, may implicate privacy. For example, developers may find that while suppressing data elements such as Social Security numbers is relatively straightforward, attempting to limit outputs based on contexts such as mental health is challenging and risks being over-suppressive. Furthermore, to maintain output-suppression techniques, developers need to retain the very information they want to prevent their models from including in their output. As a result, while it may be useful for preventing predictable incursions against data privacy, it is not likely to be a panacea for all the potential harms that may arise.

Conclusion

Foundation models present a challenge for data protection and privacy regimes, including existing regulations as well as the global principles that underlie them. The data rights that governed the emergent world of databases and rules-based systems are already being taxed by foundation models' seemingly insatiable appetite for data, creating incentives for companies to both collect more data and encourage its production.

While the recommended actions above can help to manage privacy risks and harms, ethical and legal questions remain that policymakers must confront: Should the principles of data protection continue to stand, or does the potential of general-purpose foundation models require a rethinking of our individual rights and developers' responsibilities? Or more fundamentally, must we reconceptualize how we collect and manage personal data, given the risks it poses to our privacy, autonomy, and civil liberties? As revealed by Anthropic's conflict with the U.S. Department of Defense in February 2026, foundation models have the capabilities to enable both individual and population-level surveillance, and preventing models from being deployed for this purpose is a policy and design decision that rests with the developers. Absent specific legal provisions that would prevent such uses, the public is reliant on developers to prevent their models from being used for these purposes.

In the midst of significant uncertainty, it is certain that tussles over access to and control of data will be an ongoing flashpoint as AI systems grow and expand.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.

The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu.**

Jennifer King is the Privacy and Data Policy Fellow at the Stanford University Institute for Human-Centered Artificial Intelligence (HAI).

Tiffany Saade is a Product Manager for AI Security at Cisco and a 2025 graduate of the Master's in International Policy program at Stanford University.

Acknowledgments

The authors thank Caroline Meinhardt for her patient and thorough editing of this brief, Caroline Yee for her indefatigable research assistance, and Sanmi Koyejo for reviewing an early version of this brief.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**