Google

June 2024

Generative AI and Privacy

Policy Recommendations Working Paper

This is a working paper, reflecting views as of June 2024. We will update the principles and recommendations outlined in this document as the technology continues to develop.

Executive Summary

It's an exciting time for the development of AI, including the wide availability of generative AI (GAI) tools and the release of powerful new models such as Google's Gemini.[1] GAI is a type of machine learning model that can take what it has learned from the examples it has been provided to create new content like text, images, music, and code. Such capabilities can simplify day-to-day tasks, spur scientific discovery, and help solve global challenges.

We believe our approach to AI must be both bold and responsible. That means developing AI in a way that maximizes the positive benefits to society while addressing the challenges. We also know that building AI responsibly is only possible when we do it together - as part of a collective effort involving researchers, social scientists, industry experts, governments, creators, publishers, and people using AI in their daily lives.

Policymakers watching the development of GAI have asked questions about GAI's use of personal data and its intersection with privacy. In this paper, we offer insights into how and why GAI interacts with personal data and some initial views on how organizations and policymakers can apply long-standing privacy principles to protect personal data. We address some of the more pressing questions about privacy and GAI, such as the need for publicly available data and personal data to develop GAI, the difficulty of deleting personal data from pre-training data sets and the influence of such data from the models, and how data minimization can be achieved even with large datasets. While GAI requires use of personal information, this can be done in a way that safeguards privacy, through a framework that addresses privacy concerns while supporting the responsible use of publicly available data.

Section I gives an introduction to why we believe it is important to focus on the future of privacy in the age of AI.

Section II describes the lifecycle of personal data in a GAI system, including initial collection, pre-training of a GAI model, model fine-tuning, and its outputs. GAI systems process data from the open web and beyond to produce outputs in response to user prompts. Developers of GAI tools must ensure that each stage of the data lifecycle includes privacy safeguards that address relevant risks, from data collection to potentially greater output issues like personal data leakage.

Section III describes how developers and deployers can apply foundational privacy principles, such as accountability, transparency, user controls, and data minimization to GAI. The section also addresses safeguards for model outputs and the need for greater protections for teens and children.

Policymakers have an important role to play in helping foster the responsible deployment of these new tools. There is time to get this right, and we support creative solutions like regulatory sandboxes to offer space for a risk-based and proportional consideration of privacy issues. We look forward to continuing the discussion among stakeholders to promote privacy safeguards that enable further development of important AI tools.

To achieve these goals, we offer four concrete recommendations in Section IV:

1. Balance Benefits and Risks, ensuring that privacy safeguards built on longstanding principles apply to GAI in ways that are proportional to its benefits and risks.

2. Focus on the Outputs, so that privacy standards cover the results of AI products used by businesses and consumers.

3. Protect Access to Publicly Available Data, avoiding restrictions on the processing of public data needed to train AI models.

4. Invest in Opportunity Research, seizing AI's new privacy and security opportunities.

AI is a foundational and transformational technology, with the capacity to help and inspire people in almost every field of human endeavor. We hope that what follows progresses conversations about how GAI can best develop in a privacy-protecting way.

01 Section I: Introduction -AI Opportunity and Responsibility

Google's mission has always been to organize the world's information and make it universally accessible and useful. We're excited about the transformational power of AI, including GAI, and the helpful new ways people are using it. The opportunities are vast — from research that revolutionizes scientific discovery, to products that make everyday tasks easier and more productive for billions of people.

When Google was founded 25 years ago, most searches happened on computers in homes, computer labs, or libraries. And, from its earliest days, machine learning helped improve Search and other products, whether by understanding Search queries even when the spelling wasn't quite right, or by offering translation between commonly used languages, or by blocking billions of spam emails. Now, AI is making it possible to search with new inputs, like searching with your camera, humming a tune, or powering a translation tool in real time across over 100 languages — with the goal of soon reaching 1,000 languages.[2] AI will continue to unlock breakthroughs in many fields of human endeavor, as it already has in medical diagnosis,[3] and we expect that these gains will be felt by people, businesses, and organizations everywhere. Our recently released AI opportunity white paper provides concrete policy recommendations on how governments, civil society, and companies can work together to help AI benefit as many people as possible.[4]

At the same time, we must be clear-eyed that AI not only provides opportunities but also poses risks. Our CEO, Sundar Pichai, has said that "AI is too important not to regulate, and too important not to regulate well."[5] So, what will it take to get AI public policy right? We need proportional, flexible, and risk-based regulatory frameworks; constructive, open-minded dialogue between regulators and companies, such as through regulatory sandboxes; globally interoperable standards; and policies that promote progress while reducing risks of abuse. Achieving these goals in a time of rapid technological change requires the patience to get it right—support for even deeper regulatory capacity and collaborations across governments, the private sector, academia, and civil society.[6]

Leaders and practitioners in the privacy community are actively working to define how privacy principles apply in an AI context. Privacy principles that guide AI protections—like transparency, control, data minimization, accountability, and more—are not new, and they offer a strong foundation for AI governance.[7] However, as the technology and its uses are rapidly evolving, consensus on how exactly to apply those principles to GAI is still developing. More work is needed to add clarity in this space—and to set the conditions for the societal trust that is necessary to unlock the full promise of AI.

As important as it is to understand and address risks, it's equally important to identify the opportunities. Far too little attention has been given to the potential for GAI to power improvements in privacy, whether by supercharging cyber defenses or making privacy compliance easier for businesses large and small. These applications are just beginning to unfold but have the potential to identify, address, and reduce privacy and security risks.[8]

It remains early days, and this paper does not attempt to address all of the legal and policy issues that may arise. As GAI technology evolves and our own experience with it grows, we expect to learn from the policy community and continue contributing to the global dialogue.

02

Section II: Generative AI and Personal Data

GAI models are a type of machine learning that can take what the model has learned from provided examples to create new content, such as text, images, music, and code. Compared to other forms of AI, such as structured learning algorithms, GAI and other deep learning algorithms need significantly more or significantly higher quality input data. The technical process of "learning" for the large language models (LLMs) used in GAI, for instance, begins with training the model to identify relationships and patterns among words in a large dataset. Through this process, a GAI model will learn "parameters," which represent the mathematical relationships in data. Once the model has learned these parameters, it can generate outputs based on the parameters, including in response to natural language interactions with users.

A GAI model is developed in stages, including data collection, pre-training, and fine-tuning. As discussed further below, personal data often makes up a small, but critical, portion of these datasets to help ensure the accuracy of the models. Importantly, the processing that occurs is agnostic about whether or not the training data is personal data; the data is used to find patterns and is not used to seek to identify a human being or process data specifically about or in relation to a particular individual.

Data Collection

The amount of data needed for training generally includes millions or billions of data points that typically come from a wide range of sources such as web documents, books, and code, and includes image, audio, and video data.[9] Model developers also may use proprietary data and data licensed from third parties.[10] While personal data typically makes up a very small fraction of training data, it may include, for example, names and biographical information published in Wikipedia; professional data, including job titles and educational histories posted on a professional website; and commentary on living persons, for instance, from newspaper articles or other publications.[11] The inclusion of some personal data helps reduce bias and can improve model accuracy and performance.

Large-scale data collection has been an important part of improving the performance of general LLMs that cover a broad range of topics and domains. State-of-the-art algorithms based on large neural networks enable effective analysis and encoding of a vast number of statistics about the training corpus, and have achieved unprecedented performance on a wide range of applications.

Once data is collected, it is cleaned and pre-processed to make it suitable for training the model. These processes can involve removing irrelevant and low-quality content, deduplication, and filtering, as described further in the “Data Minimization” overview later in this paper.

Pre-training

The first stage of training a GAI model is often called pre-training. Pre-training helps LLMs, for instance, learn patterns and relationships in data and use them to predict the next probable words or images in a sequence. For example, as an LLM learns, it can predict that the next word in “rock and ___” is more likely to be ”roll” than “shoelace.” Although LLMs can generate articulate responses that may give the impression that they are retrieving information, LLMs do not inherently understand the words they are generating, the concepts they represent, or their accuracy, which is why they can sometimes produce answers that, while sounding plausible, contain factual errors.

In a text-to-image model, to associate images with the concept of “cat,” for instance, GAI would be trained on large amounts of cat photos. Over time, it associates images of animals with whiskers, fur, pointy ears, and other cat-like features with the idea of a “cat.” This allows the model to take an input, such as “cat wearing an ice cream hat,” connect what it has learned about cats, ice cream, and hats, and generate a new corresponding image, even if it has never seen an image of a cat wearing an ice cream hat in its training data.

Fine-tuning

Pre-trained models can be used directly. However, many pre-trained models are then further trained on specialized data for use in a specific application. At this stage, additional customer- or user-specific personal data such as documents or images could be included to tailor outputs at a user level. For Gemini, we may fine-tune the model using user-generated data such as conversations and feedback to better answer user questions, focusing on safety, factuality, and creativity.

Complexity

Information about people is useful in understanding language without losing meaning and context. LLMs use such data to learn how language incorporates concepts about relationships between people and the world. For instance, a model that generates text about a historical or current event will need to be able to correctly identify and use the proper names of people, places, and organizations involved.[12] Thus excluding, masking, or filtering out personal data from training data could hinder an LLM's ability to understand language and harm the quality of the model's outputs. Similarly, as GAI applications like personalized medicine, individualized tutors, and personal agents evolve and grow, models will need to continue to learn from personal data to be effective and meet consumer expectations.

While it may be technically feasible to identify certain types of more common and recognizably structured categories of personal data within training data, even this presents its own challenges at scale. It can be difficult or impossible for a model to distinguish between fact and fiction, whether a person is living or dead, or whether a word is a name and what data may be reasonably linked to it. Trying to filter such personal data out from training data would result in quality issues and additional unpredictable results. All filters have false positive and false negative rates that depend on the types of content and the configuration of the filter. High false positive rates are especially detrimental to training data sets because they cause the removal of large amounts of valuable training data that are not actual personal data. In some cases, false positives are unavoidable and can create significant problems for model training. For example, Google found that the format of a particular identifier being considered for filtering also matched the format of the four X and Y coordinates that describe a two-dimensional box. The exclusion of this critical numerical data could have appreciably degraded the model's understanding of important numerical and geometric concepts.

Another challenge arises when trying to remove the influence of data from an already-trained model. It is difficult to identify exactly which training data points are responsible for a specific result. Even if it were possible to perfectly identify responsible training data points, the potential to remove the influence of certain training examples from a trained model is an unsolved computer science research problem. While retraining the model without certain personal data is theoretically possible as an alternative solution, it is extremely resource intensive and may have quality implications. Google initiated the first machine unlearning research challenge, with the goal of advancing the state of the art and encouraging the development of efficient, effective, and ethical unlearning algorithms.

03 Section III: Generative AI and Data Protection Considerations

So how best to apply foundational privacy principles, such as accountability, transparency, user controls, and data minimization? What are the best ways to promote responsible collection and use of data for model training, safeguards for model outputs, and enhanced protections for children? In this section, we provide an initial set of considerations. We look forward to working with regulators and the privacy community on additional issues at the intersection of GAI and privacy, such as the legal basis for data processing, sensitive data protections, and purpose limitations.

Accountability

Organizations that develop or deploy GAI models should be transparent and accountable for explaining the privacy principles they follow, maintaining an internal privacy program that contemplates documenting their privacy practices.[13] This principle is core to ensuring that organizations responsibly protect data, without prescribing exactly how to do that work in a fast-evolving sector.

AI governance at Google is built around our AI Principles, which we released in 2018, and our AI Privacy Practices.[14] Our AI Principles have two parts, including both seven clear commitments of how we will build and use AI responsibly, along with specific applications of AI technology that we will not pursue. As our experience in this space deepens, this list may evolve. These principles help ensure that the way we develop AI is aligned with core human values.

Our fifth AI Principle, "incorporate privacy by design," helps guide our privacy practices in the development of AI technologies. To carry out this principle in GAI, we use structured product launch processes to ensure that, wherever appropriate, GAI products provide an opportunity for notice and consent, encourage architectures with privacy safeguards, include appropriate transparency and control over the use of data, and employ data anonymization techniques and other privacy protections. Our launch reviews rely on teams of engineers, product specialists, and/or legal counsel, charged with taking reasonable steps to assess relevant privacy risks.[15]

Google has a robust process for AI development, which includes risk assessment frameworks, ethics reviews, and executive accountability processes in place to implement the AI Principles and practices at both the development and deployment stages.

Transparency

GAI can be difficult for even experts to understand, so it is important to inform users about data practices, empowering them to make appropriate choices. Developers and operators of GAI can provide transparency through multiple mechanisms—including privacy policies, terms of service, in-product notifications and disclosures, and centralized, easy-to-access resource hubs. Privacy policies should include transparency to users about personal data collection and processing, which should include information about model training, including examples where possible. Transparency can also be achieved through additional educational content, including blogs and help centers that share examples to help make data practices more easily understandable. It will also be important to build a wider public understanding of the operation of these technological tools.

Google's fourth AI principle is "be accountable to people," and this helps guide our transparency efforts. Google has published numerous blog posts and more formal academic articles about our LLMs to facilitate understanding and increase transparency about how these models work. Each product will require its own thoughtful disclosures that are tailored to, and appropriate for, the user experience.

For our Gemini models, we created an additional notice that sits alongside the Google Privacy Policy, which is presented to all users as an interstitial before they use Gemini. We also have made available a FAQ page linked from the Gemini Privacy Notice and from Gemini's user interface that gives more information on how Gemini works, how it collects data, users' rights, how users can exercise these rights, and more. Finally, we published a new Gemini Privacy Help Hub,[16] which gives users centralized access to relevant privacy-related disclosures.

Our Cloud solutions take a similar approach to transparency in the enterprise context. Google Cloud builds privacy protections into its architecture and provides meaningful transparency over the use of data, including clear disclosures and commitments regarding access to a customer's data. This includes visibility into any records regarding the actions that Google personnel take when accessing customer content.[17] In addition, Google Cloud does not use data provided to it by its customers to train its own models without the customer's permission. In our standard GAI implementation for enterprise customers, the data at rest of the organization remains in the customer's cloud environment.

User Controls

An important part of responsible, human-centered AI is empowering users to make clear choices and to control their data as appropriate. User control plays a key role in ensuring fairness and guaranteeing individuals' rights to privacy and data protection. Empowering users to manage their personal data helps establish a foundation of trust and provides tools for people to exercise their rights.

As a signed-in experience, Gemini provides users the ability to safely manage their Gemini experience. Through Gemini Activity settings, users have the ability to stop saving their Gemini activity, as well as to access, review, and delete their Gemini conversations.

Google Cloud's enterprise customers likewise have the ability to control their data. Across our Vertex AI Platform, including Vertex AI Search and conversation, the data or content that the Vertex AI (GAI) customer inputs and the output generated by Vertex AI (GAI) is considered "Customer Data," and Google processes Customer Data[18] only according to the customer's instructions.[19]

Data Minimization

The responsible deployment of GAI includes reducing the amount of personal data needed across the lifecycle of a GAI system without reducing its quality. Setting out these minimization goals will help ensure that the data used is necessary and proportionate to the purposes for which the data is processed. It's important to recognize that there are different data minimization considerations in the training of AI models versus the products and services built upon those models. The latter may be where the greatest risks arise but also where the most practicable data minimization strategies, such as data deletion controls, are most feasible.

All LLM systems face the need to optimize both accuracy and representativeness of training data as well as data minimization. In the context of AI systems, data minimization should not mean that only small volumes of data should be used in the model training. It is possible to process personal data and still comply with data minimization[20] by selecting training data that is adequate, relevant, and limited to what is necessary for the purposes of the AI system.

There are many techniques to implement data minimization at the product or user-interface level. For example, when a user interacts with Gemini, a number of data minimization standards take effect. Users are provided a clear disclosure that they should not enter confidential information in Gemini conversations or any data the user would not want a reviewer to see or Google to use to improve its products, services, and machine-learning technologies. And, by default, user Gemini activity older than 18 months is auto-deleted, and users can turn off auto-delete or change the auto-delete period to 3 or 36 months.[21]

Automated and manual techniques may be available to help filter out personal data, train with less personal data, and delete user prompts, covering the lifecycle of personal data in GAI systems. One automated method may be to filter out problematic content, as appropriate and feasible before training a GAI model. This can include data from fraudulent websites or machine-generated spam pages, made up predominately of lists of keywords or boilerplate text.[22] Another automated method is data deduplication, a method to prevent the use of multiple copies of the same data in training, which reduces training data memorization and thus repetition in model outputs.[23] Researchers have found that deduplication can reduce the likelihood of the model memorizing given text, which could result in fewer outputs that include a regurgitation of information, including personal data.[24] Similarly, in some cases it will make sense for an LLM to use filters to remove certain personal data from the data (e.g., identity documentation and payment card information). Further, research continues on how to train machine learning models with less personal data; researchers, for instance, are exploring the use of machine learning to generate synthetic data, which can reduce the need for personal data for model training.[25]

Of course proportionality has always been an important part of data minimization, inherent in the concept of whether and how data is necessary for the processing activity.[26] The rapid progress of GAI makes it important to avoid assumptions about what data is needed to train a model, how long data needs to be maintained, and the impact of deleting, pseudonymizing, or anonymizing data. Overly restrictive application of data minimization could make models less accurate and useful, as described above. A model is more likely to generate low-quality or inaccurate information if its training data includes an insufficient amount of reliable information or examples.

Data Output Safeguards

AI-generated content can generate voices, images, and even behaviors and can share information about people. This technology can unleash great benefits to society, such as the ability to democratize access to learning across languages, but it also can be misused and pose unique challenges to the authenticity and control individuals have over their own likeness and expression. Output safeguards are one way that GAI can, over time, increasingly prevent the spread of private personal data or inappropriate, offensive, or harmful content.

The kind of GAI we see today is a new technology that is designed to be capable of generating new content, including creative outputs. However, newly generated texts or images will often contain inaccurate information. These types of outputs by LLM-based services are also referred to as hallucinations, which is a response from an LLM that may appear coherent and presented confidently but is not based on fact. Hallucinations are more likely to occur if the response is not limited to a deterministic output well grounded in training data or real-world information.

However, there are growing safeguards and technologies that can help address this issue. For instance, Gemini's "double-check" feature can help users check its answers by evaluating whether there is content across the web that substantiates the response.[27] When a statement can be evaluated, the user can click the highlighted phrases and learn more about supporting or contradicting information found by Google Search. Another safeguard is the use of an output filter.[28] Before generating an output for a user that includes a name, for instance, LLM developers can consider using a public name detector or another classifier to determine whether a person's name is likely in the query, and if so, whether that person is a public or private figure. If the classifier determines the question likely contains the name of a non-public figure, it can take action to not respond. Similarly, before generating the output for the user, an LLM can check whether there is a removal request for a specific result. Where a removal request has been approved, the model deployer can seek to suppress corresponding outputs. Of course, output safeguards are not perfect and evolve based on research advances and experience over time.

Deceptive lifelike content, which GAI can create in the form of a "deepfake," raises a number of concerns. While this paper is focused on privacy concerns centered on the collection and use of data reasonably linkable to a person (i.e., "personal data"), the protection of a person's likeness, including their name, image, and other distinct characteristics, also necessitates thoughtful legal and ethical frameworks that balance GAI innovation with the preservation of human rights and dignity. Google has a prohibited use policy that applies to GAI, which, among other restrictions, prohibits users from creating content that impersonates an individual (living or dead) without explicit disclosure in order to deceive.[29]

On the technical side, we are implementing or exploring multiple complementary solutions, including: (1) watermarking (adding invisible information to generated content that can later be detected);[30] (2) metadata (affixing information to content files that denotes whether they were AI-generated and by which model); and (3) digital signatures (also known as "hashing & logging" —generating perceptual hashes of content at the time of generation for purposes of later detection, similar to reverse image search).

Privacy Protections for Teens & Children

Children and teens are increasingly using technology in all aspects of their lives, and they are doing so at younger ages. Technology can be a force for good—allowing them to learn more effectively, explore educational interests, build life skills, and connect with friends and family. There are extraordinary opportunities for GAI to be responsibly used in ways that protect, respect, and empower children and teens. For example, GAI can help younger people to ask questions that a search engine couldn't typically answer and to pose follow-up questions to help them dig deeper.[31]

Of course, there are risks that come with the use of this technology, and realizing these benefits while minimizing the risks requires close collaboration between experts, parents, educators, and minors themselves. It also means ensuring that minors have the knowledge and resources to use these tools more safely and effectively. GAI products likely to be used by minors should take into account their needs in the development of onboarding flows, in-product notices, and educational resources.

Companies building GAI products that are available to minors should invest in AI education and literacy programs for this group.[32] This includes explaining in age-appropriate language both the opportunities and limitations of the technology, how to interact with GAI tools, and how to use GAI to empower, assist, and inspire.

Flows and notices should educate users that AI can and will make mistakes, encourage users to check information, and reinforce that the technology is rapidly evolving. Resources can, in understandable and age-appropriate ways, explain GAI functionalities and limitations and how data is collected, used, and protected.

Last but not least, it will be important to develop safeguards (such as content filters) and product policies that help prevent the generation of inappropriate, offensive, or harmful content. Safeguards can help prevent and filter the sorts of harmful content that can pose unique risks to minors based on their developmental stages. Products should responsibly handle queries related to bullying, self-harm, and illegal or age-gated substances or activities, which could include not providing responses, offering disclaimers, or directing users to resources to seek help. Dedicated teams should proactively monitor for new trends and to address them or fine-tune enforcement systems to quickly address emergent potential harms.

04

Section IV: Recommendations

Our approach to AI must be both bold and responsible. That means developing and deploying AI in a way that maximizes the positive benefits to society while addressing the challenges. It is possible—and in fact critical—to embrace that tension productively. The best way to be truly bold in the long-term is to be responsible from the start.

Building AI responsibly is only possible when done together, as a collective effort involving researchers, social scientists, industry experts, governments, creators, publishers, and people using AI in their daily lives. Our March 2023 Policy Agenda for Responsible Progress in Artificial Intelligence outlines a collective agenda to unlock opportunity by maximizing AI's promise, promoting responsibility while reducing risks of abuse, and enhancing global security while preventing malicious actors from exploiting the technology.

Below, we elaborate upon this agenda in the context of privacy and AI, focusing on four key actions for the privacy community—including policymakers, regulators, civil society, and the private sector—to protect privacy and personal data while also advancing societal benefits.

Focus on AI Products; Seize New Privacy and Security Opportunities

#1 Balance Benefits and Risks, ensuring that privacy safeguards built on longstanding principles apply to GAI in ways that are proportional to its benefits and risks.

Core, unifying principles of privacy law are an important foundation for responsibly advancing AI. For many years, these privacy principles have formed the bedrock of privacy protections, and they remain applicable and effective when it comes to GAI.[33] We recommend that stakeholders consider three key areas:

- Apply Privacy Protections Proportionally, Taking Benefits into Account. Privacy regulations around the world are written with the goal of being adaptive, risk-based, and proportional, intended to be flexible and technologically neutral. Principles-based laws protect the data privacy rights of the individual, while balancing these rights against other fundamental rights and societal goals. It is important to weigh both data protection rights and the societal benefits of economic, medical, scientific, and other advances. Policymakers and data protection authorities should remain future-oriented as technologies evolve, engaging with technologists and industry as novel privacy issues arise.

- Foster Cross-Industry Dialogue, Sandboxes, and other Cooperation Mechanisms. The entire ecosystem benefits when privacy guidance and other regulatory initiatives are coherent, interoperable, and consistently applied. Aligned approaches minimize fragmentation in global markets and allow companies to focus on privacy outcomes rather than paperwork. The intersection of privacy and AI demands dialogue among all stakeholders to address complex questions such as how data minimization should apply to model training, how individual rights can best be exercised, and who is best placed to fulfill privacy obligations across a complex ecosystem. We will need cross-industry dialogue, regulatory sandboxes, and diverse expertise to test innovative products and develop nuanced and nimble regulatory frameworks.

- Harmonize Efforts Internationally and Across Regulatory Domains. We will need regulatory cooperation to harmonize privacy, governance, competition, content regulation, safety, unlawful discrimination, and international trade commitments. Policymakers and regulators should aspire to regular coordination, sharing expertise, and exploring how to manage tensions and tradeoffs that arise between their areas.[34] International fora can help advance these conversations, share best practices, and promote interoperability. Policymakers and regulators can lean into work, for example, at the Global Privacy Assembly, the OECD, the Global CBPR Forum, and the European Data Protection Board, Ibero-American Data Protection Network, the Network of African Data Protection Authorities, the Asia-Pacific Privacy Authorities, and the Asia Pacific Economic Cooperation.

#2 Focus on the Outputs, so that privacy standards cover the results of AI products used by businesses and consumers.

GAI has two distinct phases: (1) the development and training of the AI model; and (2) the building and operation of applications — the products and services that a consumer or business uses. These two phases can have different data uses, privacy implications, risk levels, and opportunities to implement safeguards.

In general, the focus of privacy controls should be at the application level, where there may be both greater potential for harm (such as greater risk of personal data disclosure), but also greater opportunity for safeguards. Leakages of personal data, or hallucinations misrepresenting facts about a non-public living person, often happen through interaction with the product, not through the development and training of the AI model. At the same time, the product stage offers opportunities for privacy safeguards and protection against inappropriate, offensive, or harmful content, such as through enforcement of usage policies, limitations on how the application interacts with personal data, watermarks, fine tuning of models, filters, classifiers, and other output safeguards in addition to enhanced transparency and user controls.

We therefore recommend that the privacy community focus the application of privacy protections to the AI products used by businesses and consumers, which generally align with areas of greatest potential for risk mitigation.

#3 Protect Access to Publicly Available Data, avoiding restrictions on the processing of public data needed to train AI models.

Publicly available information is at the core of how AI models are trained; it is foundational to model quality and functionality. Part of this publicly available information may incidentally include personal data. Although there are techniques to reduce certain highly specific personal data collected and processed in the training phase, currently, personal data is key to training models to understand language and cannot be removed easily or without potential quality implications. Doing so is particularly important as we consider adversarial dynamics online. Bad actors, including criminal organizations and hostile intelligence services, are free to train AI models using publicly-available data, and will not adhere to regulations. Restrictions on the use of this data will create a dynamic where bad actors can develop more powerful models than rule-bound companies, creating more than just privacy risks.

Data about people can help LLMs learn how language incorporates concepts about relationships between people and the world. For instance, a model that generates text about a historical or current event will need to be able to correctly identify and use the proper names of people, places, and organizations involved in the event. Excluding, masking, or filtering out personal data pulled from the open web from training data could hinder an LLM's ability to understand language and can impact the quality of the model. Additionally, to fully remove such data from a large model could require retraining the model, which can be incredibly resource intensive and may be unfeasible on a short timeline.

AI technologies can deliver high-quality results for people around the world. To do so requires preserving cross-border data flows and continuing to advance the agenda for trusted data flows with new urgency. Not only are data flows vital for these systems, but they also enable a more global, equitable training set in which AI tools can learn to engage effectively in many languages. Already today, data localization comes with negative consequences for resilience, cybersecurity, the economy, and more. The pitfalls of data localization will only grow as data flows contribute to the enhanced productivity and innovation of AI systems. And as authoritarian surveillance states maintain an advantage in the quantity of modeling data over rule-of-law nations, it will be important for democracies and partners with shared values to work together, not at cross-purposes.

For these reasons, we recommend that the privacy community avoid the use of privacy law to restrict the processing of public data and preserve trusted data flows, which are at the heart of responsible AI innovation.

#4 Invest in Opportunity Research, seizing AI's new privacy and security opportunities.

Today, there is far too little conversation happening on how AI can support and enhance privacy goals. But there is huge potential to unlocking the power of data in AI systems while also protecting privacy. We recommend stakeholders consider three key areas:

- Use AI to Power Consumer Privacy. AI offers a substantial opportunity to improve consumer privacy and support companies' efforts around privacy compliance. AI, for example, can help organizations understand privacy-related feedback for large numbers of users [35] or help identify privacy compliance issues.[36] With more GAI-powered privacy tools on the horizon, we need dedicated efforts on how to advance privacy through AI, as well as ensure such tools aren't unintentionally restricted by well-intended privacy regulations.

- Promote Privacy-Enhancing Technologies and Incentivize their Research and Application. Technologies like differential privacy, multiparty computation, and federated learning all have the potential to contribute to privacy in AI systems. Other more emergent techniques, like generating synthetic data and machine unlearning, also have promise. However, more work is needed to stimulate their development and application to AI use cases and overcome current limitations.[37] Lawmakers, regulators, and international organizations can help accelerate and broaden the adoption of privacy-enhancing technologies by companies of all sectors and sizes. This could include promoting privacy-enhancing technologies in laws and regulations. Increasing attention and investment on this work could come through investments, research prizes, and other incentives for privacy-enhancing AI technologies. Policymakers could also consider incorporating the use of AI privacy technologies into "best practice" recommendations as a mitigating factor in enforcement actions and in assessing sanctions and/or levels of fines.

- Supercharge AI-powered Cyber Defenses. The privacy community should support work to supercharge AI-powered cyber defenses, and study whether and how such tools could advance protections for personal data. The cybersecurity community is rapidly advancing the state of the art of how AI tools can be used to secure networks. AI-driven tools, for example, can more rapidly detect threats or identify when employees may have excessive permissions to network resources. Any of these capabilities can support privacy goals by reducing the potential for unauthorized access or leaks. But there may also be opportunities to build or implement AI tools in other ways that directly support the protection of personal data, such as tools that could rapidly identify, flag, or stop instances of unauthorized data exfiltration. The privacy community can also help support expectations that GAI products incorporate best practices for security by design[38] and work with the cybersecurity community to address threats, such as "training extraction" and "prompt injection" that have a particular nexus to personal data.

We hope this working paper will contribute to a deep and collaborative dialogue on the intersection of privacy and GAI. To fully unlock the benefits of GAI, we must focus on the harms we want to avoid and the risks we want to mitigate, as well as on the potential we want to achieve. We look forward to working with policymakers, regulators, and civil society on how best to protect individual privacy in the age of GAI.

[1] See Sundar Pichai and Demis Hassabis, Introducing Gemini: our largest and most capable AI model, The Keyword (Dec. 6, 2023), https://blog.google/technology/ai/google-gemini-ai.

[2] Google Lens, Google, https://lens.google (last visited Feb. 7, 2024); Krishna Kumar, Song stuck in your head? Just hum to search, The Keyword (Oct. 15, 2020), https://blog.google/products/search/hum-to-search; Jeff Dean, 3 ways AI is scaling helpful technologies worldwide, The Keyword (Nov. 2, 2022), https://blog.google/technology/ai/ways-ai-is-scaling-helpful.

[3] Andrew Carroll, A breakthrough to better represent human genetic diversity, The Keyword (May 10, 2023), https://blog.google/technology/health/first-pangenome-reference-nature-paper-ai.

[4] Google, The AI Opportunity Agenda, https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/AI_Opportunity_Agenda.pdf (last visited Feb. 7, 2024).

[5] Sundar Pichai, Google CEO: Building AI responsibly is the only race that really matters, Financial Times (May 3, 2023), https://www.ft.com/content/8be1a975-e5e0-417d-af51-78af17ef4b79.

[6] Kent Walker, A shared agenda for responsible AI progress, The Keyword (Mar. 28, 2023), https://blog.google/technology/ai/a-shared-agenda-for-responsible-ai-progress.

[7] As with all our products, our GAI work is guided by our Privacy Principles, Responsible AI Practices, and our AI Principles. See Privacy Principles, Google, https://safety.google/principles (last visited Feb. 7, 2024); Responsible AI Practices, Google, https://ai.google/responsibility/responsible-ai-practices (last visited Feb. 7, 2024); Our Principles, Google, https://ai.google/responsibility/principles (last visited Feb. 7, 2024) ("AI Principles"). Protecting privacy requires us to be rigorous in building all our products to be private by design. We uphold responsible data practices as outlined in our Privacy Principles and leverage our deep research and technical advances in AI, hardware, and cloud computing to continually improve our approach. As we slowly and deliberately deploy GAI, we are building in privacy safeguards, user controls, and meaningful information about our products' data practices to protect our users and their privacy.

[8] While not all of these AI tools may not specifically constitute "GAI," they demonstrate how AI tools can be leveraged to protect security and privacy.

[9] Google, Gemini: A Family of Highly Capable Multimodal Models, at 5, https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (last visited Feb. 8, 2024) ("Gemini Technical Report").

[10] Google DeepMind, AI Safety Summit: An update on our approach to safety and responsibility (Oct. 27, 2023), https://deepmind.google/public-policy/ai-summit-policies/#data-input-controls-and-audit.

[11] Additionally, user-facing applications may permit users to input personal data into a GAI application, among other application-level data inputs, such as geolocation. In the case of Google's Gemini, precise location information will be collected with user permission to improve Gemini's outputs, but is not persistently logged with the user's Gemini activity or used for model training purposes. And, if a user has not granted the additional permissions to process precise location, Gemini will attempt to answer such queries using the available coarse location information to the best of its ability. See Gemini Privacy Help Hub, Google, https://support.google.com/Gemini/answer/13594961#zippy=%2Cwhat-location-information-does-Gemini-collect-why-and-how-is-it-used (last visited Feb. 7, 2024).

[12] For example, modifying a dataset to remove all proper nouns and replace them with a symbol (e.g., "X") would make the following sentence very confusing: "X went to X's house while X was visiting X". Removing names of people, places, and organizations diminishes the utility of models.

[13] We have provided an overview of our work in this area. See Responsible Data Practices, Google, at 6 (2022), https://blog.google/documents/130/Responsible_Data_Practices.pdf ("Responsible Data Practices").

[14] We seek to lead not only in state-of-the-art AI technologies, but also in state-of-the-art AI responsibility. In 2018, we were one of the first companies to articulate AI Principles focusing on beneficial use, users, safety, and risk avoidance, and we have pioneered many best practices, like the use of model and data cards, now widely used by others. See AI Principles. We also regularly publish reports of our progress. See Marian Croak & Jen Gennai, An update on our work in responsible innovation (2022), https://blog.google/technology/ai/an-update-on-our-work-in-responsible-innovation; and AI Principles Progress Update (2022), https://ai.google/static/documents/ai-principles-2022-progress-update.pdf ("2022 AI Principles Progress Update").

[15] Responsible Data Practices; Perspectives on Issues in AI Governance, Google, https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf (last visited Feb. 8, 2024).

[16] Gemini Privacy Help Hub, Google, https://support.google.com/Gemini?p=privacy_help (last visited Feb. 7, 2024).

[17] See Access Transparency, Google Cloud, https://cloud.google.com/assured-workloads/access-transparency/docs/overview (last visited Feb. 7, 2024); Privacy, Google Cloud Help, https://support.google.com/googlecloud/answer/6056650 (last visited Feb. 7, 2024); How Google protects your organization's security and privacy, Google Workspace Admin Help, https://support.google.com/a/answer/60762 (last visited Feb. 7, 2024); Andrew Moore, Sharing our data privacy commitments for the AI era, Google Cloud (Oct. 14, 2020), https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-unveils-ai-and-ml-privacy-commitment; Access Transparency and Access Approval, Google Cloud, https://cloud.google.com/access-transparency (last visited Feb. 7, 2024).

[18] See the GAI Service terms as part of the Google Cloud Service Specific Terms. See Service Specific Terms, Google Cloud, https://cloud.google.com/terms/service-terms (last visited Feb. 7, 2024)

[19] See Cloud Data Processing Addendum (Customers), Google Cloud, https://cloud.google.com/terms/data-processing-addendum (last visited Feb. 7, 2024).

[20] As stated by the French data protection authority, the CNIL, "The principle of data minimization does not prevent, according to the CNIL, the training of algorithms on very large datasets. On the other hand, the data used must, in principle, have been selected to optimize the training of the algorithm while avoiding the use of unnecessary personal data. In any case, certain precautions to ensure data security are essential." Artificial intelligence: CNIL unveils its first answers for innovative and privacy-friendly AI, CNIL (Oct. 16, 2023), https://www.cnil.fr/en/artificial-intelligence-cnil-unveils-its-first-answers-innovative-and-privacy-friendly-ai.

[21] Manage & delete your Gemini activity, Gemini Help, https://support.google.com/Gemini/answer/13278892 (last visited Feb. 7, 2024).

22 See Guilherme Penedo et al., The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only, Falcon, at 4-7 (June 1, 2023), https://arxiv.org/abs/2306.01116 ("Falcon Report").

23 Katherine Lee et al., Deduplicating Training Data Makes Language Models Better (Mar. 24, 2022), https://arxiv.org/abs/2107.06499; see also Google, PaLM 2 Technical Report, at 21, https://ai.google/static/documents/palm2techreport.pdf (last visited Feb. 8, 2024) ("PaLM 2 Technical Report").

24 See PaLM 2 Technical Report at https://ai.google/static/documents/palm2techreport.pdf.

25 See 2022 AI Principles Progress Update at https://ai.google/static/documents/ai-principles-2022-progress-update.pdf.

26 See, e.g., Guidelines 4/2019 on Article 25 Data Protection by Design and by Default: Version 2.0, European Data Protection Board (Oct. 20, 2020), https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf.

27 Yury Pinsky, Gemini can now connect to your Google apps and services, The Keyword (Sept. 19, 2023), https://blog.google/products/gemini/google-bard-new-features-update-sept-2023/.

28 See, e.g., Safety Settings, Google AI for Developers, https://developers.generativeai.google/guide/safety_setting (last updated May 9, 2023).

29 See, e.g., Community Guidelines, YouTube, https://www.youtube.com/howyoutubeworks/policies/community-guidelines (last visited Feb. 7, 2024).

30 See, e.g., Sven Gowal and Pushmeet Kohli, Identifying AI-generated images with SynthID, Google DeepMind (Aug. 29, 2023), https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid.

31 Hema Budaraju, How we're responsibly expanding access to generative AI in Search, The Keyword (Sept. 28, 2023), https://blog.google/products/search/google-generative-ai-search-expansion.

32 Google has launched several initiatives across the globe to support kids and families to learn how to harness the full potential of generative AI while understanding the risks. 5 must-knows to get started with generative A, Google, https://services.google.com/fh/files/misc/google_youtube_5_must_knows_to_get_started_with_generative_ai.pdf; Experience AI, Google DeepMind, https://experience-ai.org; A Guide to AI in Education, Google for Education (2023), https://services.google.com/fh/files/misc/gfe_guide_to_ai_in_education.pdf.

33 Google has long championed smart, interoperable, and adaptable data protection regulations—rules that will protect the privacy rights of people and communities while enabling innovative services.

34 Strong examples of this exist already today, such as the UK Digital Regulation Cooperation Forum (DRCF), which was established to ensure greater cooperation on regulatory matters, including AI Members of the DRCF include the Competition and Markets Authority, the Information Commissioner's Office, the Office of Communications, and the Financial Conduct Authority.

35 Hamza Harhous et al., Hark: A Deep Learning System for Navigating Privacy Feedback at Scale, 2022 IEEE Symposium on Security and Privacy (2022), https://research.google/pubs/pub51307.

36 Fergus Hurley and Nia Castelly, Checks, Google's AI-powered privacy platform, The Keyword (May 3, 2023), https://blog.google/technology/ai/checks-googles-ai-powered-privacy-platform/.

37 For example, multiparty computation, which allows the combining of data sources of multiple parties without revealing the actual data, is computationally intensive and expensive. Federated learning involves greater processing on individuals' devices without data being processed in the cloud but can be hard to implement on devices that have less compute power. Synthetic data is good for distilling a smaller model from a bigger one, but it requires humans in the loop to maintain quality.