

Delivering trusted and secure AI

By: Marina Kaganovich,
Rohan Kanungo, and
Heidi Hellwig

March 2025

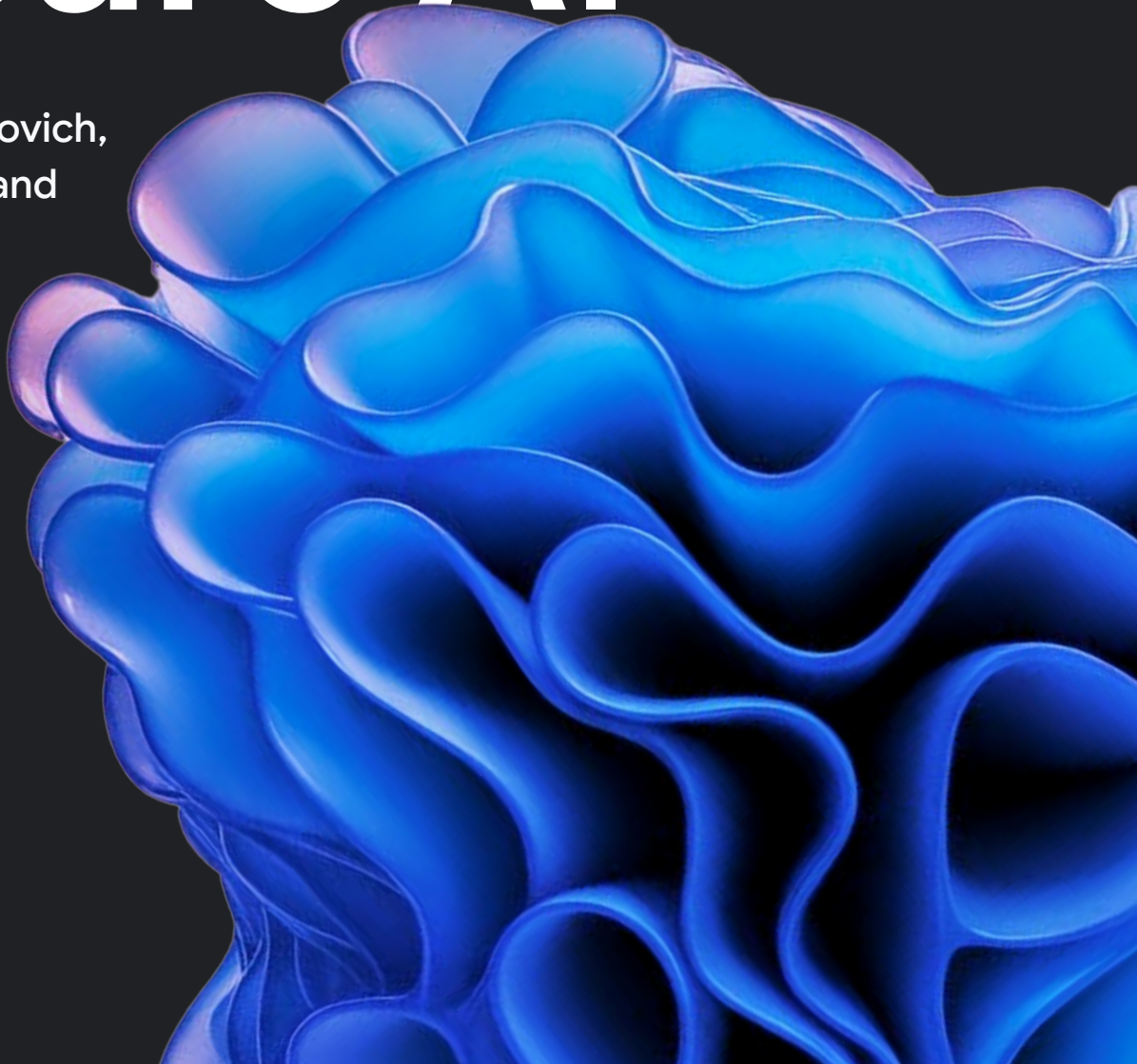




Table of Contents

Introduction	3
---------------------	----------

Chapter 01: Responsible Innovation	5
Risk Assessments	7
Data Governance	9
Grounding Gen AI in Enterprise Truth	10
Privacy	12
Security	15
Compliance	20
Open Cloud & Portability	23
Environmental Impact	24

Chapter 02: Shared Fate	26
--------------------------------	-----------

Chapter 03: Best Practices	28
Governance	29
Acceptable Use	29
Security	29
Privacy and Data Governance	30
Staying on Top of Gen AI Developments	31

Conclusion	32
-------------------	-----------

Disclaimer: The content contained herein is correct as of February 2025 and represents the status quo as of the time it was written. Google Cloud's security policies and systems may change going forward, as we continually improve protection for our customers.

Introduction

Enterprises today face a critical challenge: delivering AI to production while ensuring accuracy, safety, and data security. Google Cloud's approach to generative AI prioritizes enterprise readiness with built-in mechanisms for robust data governance, privacy controls, IP indemnification, and responsible AI practices. We provide the tools and services necessary to secure AI and offer data sovereignty options, giving customers the confidence to deploy models at scale.

To maximize the potential of AI innovation, while minimizing risks, enterprises are right to be asking questions around AI's governance, safety, and fairness. Many are also expressing concerns around privacy, transparency, and trustworthiness.



It's why responsible AI matters. It brings tremendous opportunity to:



Enhance brand reputation



Improve customer engagement



Improve long-term profitability



Strengthen product value and trust



Prepare for AI regulations



Increase customer loyalty and trust

In this paper, we explore how Google Cloud helps enterprises realize these benefits. It unpacks how we build enterprise-grade [gen AI](#) responsibly, and how we approach AI data governance, privacy, security, and compliance when developing [gen AI](#) through the [Vertex AI platform](#). For context as used throughout this paper, gen AI refers to the use of AI to create new content such as text, images, music, audio, and video, or some variation thereof as enabled by [multimodal gen AI](#).

Gen AI works by using an ML model to learn the patterns and relationships in the provided dataset(s). It then uses the learned patterns to generate new content. Gen AI is powered by foundational models (large AI models) that can multi-task and perform out-of-the-box tasks, including summarization, Q&A, classification, and much more.

The Vertex AI platform is a machine learning platform that enables you to train and deploy machine learning models and AI applications, and customize large language models (LLMs, a form of foundation models) for use in your AI-powered applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling your teams to collaborate using a common toolset and scale your applications using the benefits of Google Cloud. In addition, you can also adapt foundation models for targeted use cases with minimal training and very little example data.

Our approach to developing and harnessing the potential of AI is grounded in our founding mission — to organize the world’s information and make it universally accessible and useful. We believe [our approach](#) to AI must be both bold and responsible. Bold in rapidly innovating and deploying AI in groundbreaking products used by and benefiting people everywhere, contributing to scientific advances that deepen our understanding of the world, and helping humanity address its most pressing challenges and opportunities. And responsible in developing and deploying AI that addresses both user needs and broader responsibilities, while safeguarding user safety, security, and privacy.

We’ve also infused these values into the [Google Cloud Platform Acceptable Use Policy](#) and the [Generative AI Prohibitive Use Policy](#) so that they are transparent and clearly communicated. In addition, when it comes to AI, we recognize the need for both good individual practices and shared industry standards. We’ve continued evolving our practices, conducting industry-leading [research](#) on AI impacts and risk management, and [assessing proposals](#) for new AI research and applications to ensure they align with our principles. We continuously iterate and [reassess](#) how to build accountability and safety into our work, and publish on our progress to encourage collaboration and advancements in this field.



Core pillars for trust in AI:

01. Security

Know that your AI implementations are protected by the comprehensive security controls and best practices in place at Google Cloud.

02. Privacy

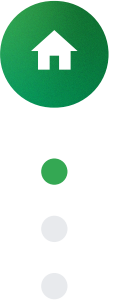
Uphold best-in-class privacy standards with Google Cloud’s transparent, fully accountable privacy-by-design approach to AI.

03. Compliance

Keep pace in an evolving and complex compliance landscape, with the assurance that Google Cloud can support your regulatory efforts – no matter where you operate around the world.

04. Data Governance

Control how and where your data is used, safe in the knowledge that Google Cloud will never use your data to train our models.



Responsible Innovation

With AI moving so fast, how can we strike the right balance between innovation, reliability, and risk mitigation? In this chapter, learn how Google Cloud embeds a responsible approach to building AI technologies by carefully considering everything from risk assessments and data governance to privacy, security, and compliance; as well as portability and emissions reduction.



As with any transformational and new technology, gen AI comes with complexities and risks, and these need to be managed as part of a comprehensive risk management framework and governance structure. AI presents critical questions and we are [working to build AI responsibly](#) to benefit both our customers and the wider societies in which we operate. The challenge is to do so in a way that is proportionately tailored to mitigate risks and promote reliable, robust, and trustworthy AI applications, while still enabling innovation and the promise of AI for societal benefit.

Responsible AI is woven into the fabric of our work. As part of our [principled](#) approach to building AI technologies, we commit to developing and applying strong safety and security practices, and incorporate our privacy principles in the development and use of AI. Recognizing that [rigorous evaluations](#) are critical to building successful AI, we engage specialized teams in analyses and risk assessments for the AI products we build and for early-stage customer co-development opportunities.

[Responsible product development](#) practices span multiple dimensions. Some are technical, involving evaluations of data sets and models for bias; some pertain to product experiences; and some are around policy, informing what we will and won't offer from a product perspective. We have developed a four-phase process (consisting of Research, Design, Govern, and Share) to review projects against the AI Principles and work with subject matter experts on privacy, security, and compliance, to name a few. The initial Research and Design phases foster innovation, while the Govern and Share phases focus on risk assessment, testing, monitoring, and transparency.

Our research draws on in-house expertise, including computer scientists, social scientists, and user experience researchers. We also regularly publish on the [progress we're making](#) to enable transparency into our work, support safer and more accountable products, earn and keep our customers' trust, and foster a culture of responsible innovation.

Our approach to building AI responsibly is guided by our AI Principles and builds upon Google's previous experience with keeping users safe on our platforms. As we build gen AI services, our technical approaches to enforce policies at scale include techniques like fine-tuning and [reinforcement learning from human feedback](#) (RLHF). Other layered protections are invoked both when a person inputs a prompt and again when the model provides the output. Policy improvements are informed by ongoing user [feedback](#) and monitoring. Responsibility by design also involves building security into our products from the very beginning. We've codified this approach in our [Secure AI Framework](#) (SAIF). Applying SAIF, we build on our existing security knowledge and adjust mitigations to these new threats, as further discussed below.

Risk Assessments

For every organization, the decision to leverage the power of gen AI hinges on a myriad of questions, one of the most salient being: How can I help my organization harness the power of AI while minimizing risk? Google Cloud helps customers answer this question in a number of ways.

01. Comprehensive reviews during AI product development

We identify and assess potential risks at both the model level, and the point of their integration into a product or service. Our socio-technical approach considers how AI will interact with the world and existing social systems, and assesses the potential impacts and risks that may be posed both at the initial release and as time goes on. Reviewers understand that potential risks and impacts might be different at the model level and at the application level, and consider mitigations accordingly. We draw from various sources, including academic literature, external and internal expertise, and our in-house ethics and safety research.

02. Private releases of models

This allows our product teams to gather valuable feedback before we make them generally available. Once feedback is incorporated, we update our product documentation to account for any changes. This documentation generally provides known limitations of the model and may include service-specific terms to further advise customers on proper use of our products. Google Cloud continues to invest in tools to support our customers including: Vertex's [Explainable AI](#), [Model Fairness](#), [Model Evaluation](#), [Model Monitoring](#), and [Model Registry](#) to support data and model governance.

03. Mitigation strategies to address potential risks

These apply to any risks identified prior to releasing the product for general availability (GA), and can take various forms. For instance, for gen AI products, mitigations may draw on technical approaches to evaluate and improve models during development, establish policy-driven safety guardrails, or may be enabled by tooling customers can leverage in their own projects for further safety efforts. Policy restrictions are typically guided by the relevant Acceptable Use Policy, Terms of Service, and privacy restrictions, as further discussed in the AI Data Governance and Privacy section below.



Enterprises can further mitigate the risk of AI adoption using:

Customizable technical controls such as [safety filters](#), which can block model responses that violate policy guidelines, for instance, around child safety. A customer could create safety filters leveraging [safety attributes](#), which include “harmful categories” and topics that may be considered sensitive, such as “drugs” or “derogatory.”

Google’s [Responsible Generative AI Toolkit](#), which provides guidance and tools to create safer AI applications with these new open models, including how to set safety policies and methodologies for building robust safety classifiers.

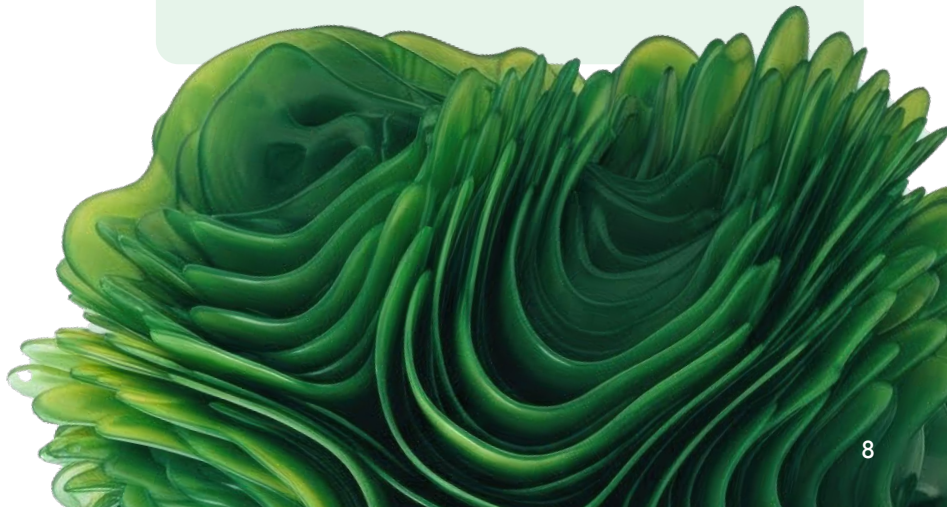
[Explainable AI](#) tools and frameworks to help understand and interpret predictions made by machine learning models, natively integrated with a number of Google Cloud products and services.

Adaptable thresholds for blocked responses to help enterprises control content based on their own business needs and policies. For instance, safety settings can be configured based on both probability and severity scores.

Tools to better understand and control AI models. For instance, using models similar to those in our safety filters, customers can use our [text moderation](#) service to scan the entire corpus of training set data for terms that fall within predefined “harmful categories” and topics that may be considered sensitive, enabling ongoing compliance.

[Model evaluation](#) on Vertex AI, which includes metrics to understand model performance and [evaluate potential bias](#) using common data and model bias metrics. These tools can promote fairness by evaluating data and model outputs during training and over time, highlighting areas of concern and providing suggestions for remediation.

In addition, documentation on our [API models](#), [open models](#), and [large language foundation models](#) is available on our [model card hub](#), articulating our model’s strengths and limitations.



Data Governance

Every enterprise wants the assurance that its differentiated data remains private and protected. Yet using gen AI in a business setting can pose various risks around accuracy, privacy and security, regulatory compliance, and intellectual property. As with other forms of digital innovation, creating a programmatic, repeatable structure can help achieve a consistent approach to evaluating AI use cases.

To ensure data privacy commitments are considered when developing and deploying gen AI, we've implemented robust data governance reviews. One of the questions we are frequently asked is whether our foundation models are trained on customer data, and by extension, whether customer data may as a result be exposed to Google Cloud, Google Cloud's other customers, or the public. To address this question, we outline some key aspects of our [model tuning and deployment](#) and [data governance practices](#) below. More information can be found in our [paper on foundation model adaptation](#).



Google Cloud never uses customer data to train our models

While we process customer data to provide our services, we do not use customer data to train our foundation models without the customer's prior permission or instruction.



Customers always have control over how and where their data is used

The foundation models on Vertex AI are developed to handle general use cases. Customers can customize foundation models for specific use cases by tuning them using our tuning APIs. This approach combines our research and product development expertise to enable world-class AI without compromising customers' control over their data.



Input data, including prompts and adapter weights, are kept private

These are considered customer data and are stored securely at every step along the way – encrypted at rest and in transit. Customers can control the encryption of the stored adapter weights by using customer-managed encryption keys (CMEK) and can delete their adapter weights at any time. Customer data used to train adapter models will not be logged or used for improving the foundation model without the customer's permission.

✦✦ Grounding Gen AI in Enterprise Truth

To fully unleash the power of gen AI, businesses must ground foundation models in enterprise systems and fresh data. At Google Cloud, we call this real-time information “[enterprise truth](#).”

Grounding a foundation model in enterprise truth will significantly improve response accuracy and completeness, unlocking unique use cases and paving the way for advanced AI agents. It helps to:



Build trust through data authenticity

Using genuine, reliable data vastly improves the trustworthiness of a model’s output – which in turn helps build users’ trust and confidence in its abilities. This earned trust unlocks more sensitive, mission-critical use cases and lays the groundwork for the next generation of AI agents.



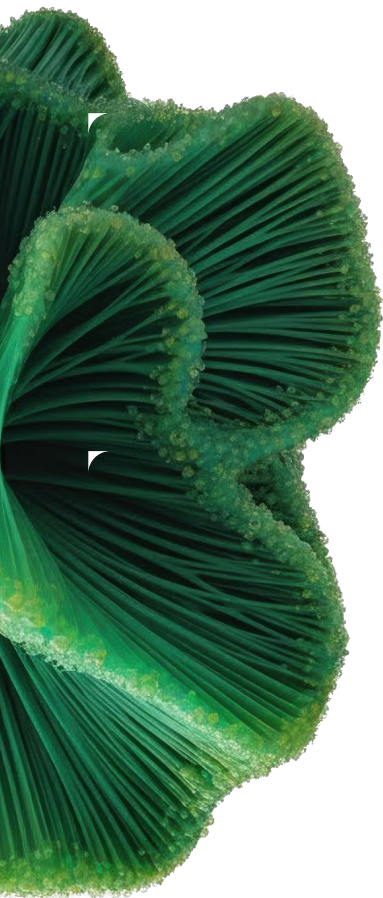
Deliver more data relevance

The model can deliver more informative and nuanced output, directly relevant to the specific context. This is particularly useful for applications that require specificity or detail.



Minimize hallucinations

Generative models learn from statistical relationships, and can sometimes generate plausible-sounding outputs that are factually incorrect. Grounding ensures the model cross-references its responses against verifiable facts. Google Cloud customers are protected with an industry-first indemnification, which means that if they are challenged on copyright grounds for our use of training data or their generated output, we will assume responsibility for the potential legal risks involved.





Use gen AI for real-world tasks

Models are trained on data that is just a snapshot in time. To apply gen AI to use cases that require precise, up-to-date information, it's critical that the model has access to fresh information.

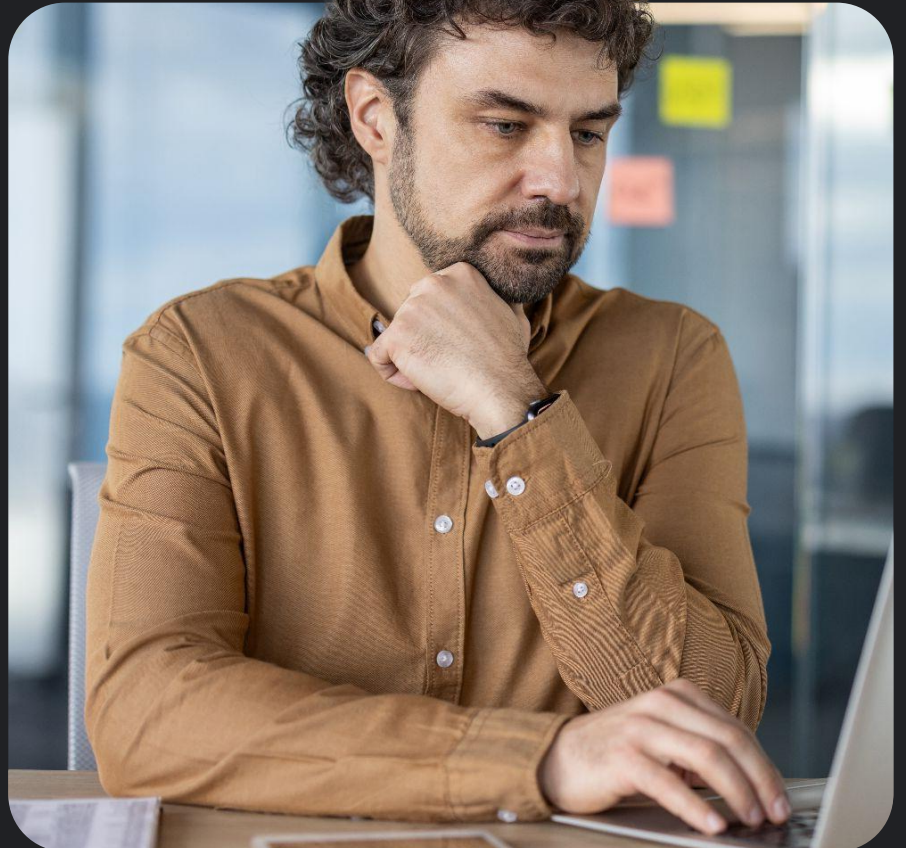
To help enterprises ensure their agents deliver accurate, factual, and reliable results, Google Cloud offers the most comprehensive approach to grounding AI in enterprise truth. We are the only cloud provider to combine the power of grounding with Google Search (which powers 99% of the world's search data), enterprise data like databases and data warehouses, third-party enterprise applications like ERP, CRM, and HR systems, and specialized third-party data from providers like Moody's, MSCI, Thomson Reuters, and Zoominfo. This multi-pronged grounding approach minimizes hallucinations in model responses, empowering organizations to unlock the full potential of their data and accelerate AI innovation.





Privacy

In this section, we provide an initial set of considerations for how we apply foundational privacy principles, such as accountability, transparency, and data minimization to enable the responsible collection and use of data for model training and safeguards for model outputs. Our approach includes incorporating privacy design principles, designing architectures with privacy safeguards, and providing appropriate transparency and control over the use of data. When bringing new offerings to the market, we incorporate these principles throughout the product lifecycle and design architectures with comprehensive privacy safeguards such as data encryption.





Accountability

AI governance at Google is built around our [AI Principles](#), which we released in 2018, and our AI privacy practices. These help ensure that the way we develop AI is aligned with core human values.

Google's responsible development and deployment helps guide our privacy practices in the development of AI technologies. To carry out this principle in gen AI, we use structured product launch processes to ensure that, wherever appropriate, gen AI 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 other specialists in privacy, legal, and safety charged with taking reasonable steps to assess relevant privacy risks. 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

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. In addition, Google Cloud does not use data provided to it by its customers to train its own models without

the customer's permission. Our teams assess products for compliance with data privacy and transparency requirements. These reviews also consider adherence to the [commitments](#) made to our customers regarding data privacy and protection – specifically, the ability to control how customer data is accessed, used and processed – as articulated in the [Google Cloud Platform Terms of Service](#) and [Cloud Data Processing Addendum](#). In addition, we provide customers with the option to see [who can access their data and why](#).

User Controls

Empowering users to manage their personal data helps establish a foundation of trust and provides tools for people to exercise their rights. Google Cloud's enterprise customers have the ability to control their data. On our Vertex AI Platform, customer data prompted into a model and the output generated by Vertex AI from those prompts is "Customer Data," and Google processes Customer Data only according to the customer's instructions.

Vertex AI already [provides robust data commitments](#) to ensure customer data and models are not used to train our foundation models without permission or leaked to other customers.





However, there are two additional concerns that organizations have: (1) protecting data, and (2) reducing the risk of customizing and training models with their own data that may include sensitive elements such as personal information (PI) or personally identifiable information (PII). Often, this personal data is surrounded by context that the model needs so it can function properly. To effectively segment, anonymize, and help protect data, we have a robust set of tools and services that we are continuously optimizing, including:

Our [Sensitive Data Protection](#) service, which provides sensitive data detection and transformation options such as masking or tokenization to add additional layers of data protection throughout a generative AI model's lifecycle, from training to tuning to inference.

[VPC Service Controls](#), which allows for secure deployment within defined data perimeters. With VPC SC, you can define perimeters to isolate resources, control and limit access, and reduce the risk of data exfiltration or leakage.

We are committed to preserving our customers' privacy with our Cloud AI offerings and to supporting their compliance journey.

As a global cloud provider, Google Cloud has a long-standing commitment to [GDPR compliance](#) and has compiled comprehensive [DPIA Resource Center documentation](#) to support customers in their data protection impact assessment efforts.

We also enable certain AI/ML services to be configured to meet [data residency requirements](#) as noted in our [Service Terms](#).



Security

Privacy is tightly intertwined with security, and both are primary design criteria for all products built on Google Cloud.

As AI technologies advance, we have more opportunities to enhance how we identify, address, and reduce security risks. We've taken a three-pronged approach to secure, scale, and evolve the security ecosystem by:

01.

Supporting customers in their AI implementation with controls, best practices, and capabilities

02.

Continuing to launch cutting-edge, AI-powered products and services to help organizations achieve better security outcomes at scale

03.

Continuously evolving to stay ahead of threats

Google Cloud's AI products benefit from our globally distributed, scalable, and redundant infrastructure, and [inherit the platform's foundational controls](#). Security is strengthened in a number of ways:

Defense in depth

Instead of relying on any single technology to make our infrastructure secure, our technology stack builds security through progressive layers that deliver defense in depth, at scale, and by default. It means data and systems are protected through multiple layered defenses using policies and controls that are configured across identity and access management (IAM), encryption, networking, detection, logging, and monitoring.

[A secure-by-design foundation](#)

Operational controls include in-depth security reviews, vulnerability scanning, ongoing threat monitoring, and intrusion detection mechanisms that enable secure service deployment and safeguard customer data. There are also security controls specific to [Vertex AI and gen AI](#).

Solutions designed to safeguard AI workloads and data

[AI Protection](#) can help teams comprehensively manage AI risk by: discovering AI inventory in your environment and assessing it for potential vulnerabilities; securing AI assets with controls, policies, and guardrails; and managing threats against AI systems with detection, investigation, and response capabilities.





Encryption and anonymization

Building an AI/ML system requires a large corpus of data to appropriately train models. Often, the data is sensitive – and securing it is critical. All data stored within Google Cloud is encrypted at rest using the same hardened key management systems that Google uses for our own encrypted data. These provide strict key access controls and auditing, and encrypt user data at rest using AES-256 encryption standards, with no setup, configuration, or management is required. Alternatively, you can use customer-managed encryption keys (CMEK) in the Cloud Key Management Service (Cloud KMS), which offer more control around key generation, key rotation frequency, and key location. With added control comes added responsibility in appropriately managing the keys, so we recommend assessing whether the default encryption is sufficient for your compliance needs. For more information, see how to [meet compliance requirements for encryption at rest](#).

Protecting software supply chains

Extending existing [software supply chain](#) solutions is an effective way to counter many of the risks associated with AI software supply chains. Rather than creating new solutions, we can approach AI models like traditional software. Across Google’s first-party and open source AI development ecosystems, we’re adopting the SLSA framework and format to sign model provenance. This metadata document cryptographically binds a model to the service account (an identifying account that represents an application rather than a human user) that was used to train it. It also enables Google to verify all of its models against the expected signing keys, such that an insider cannot overwrite or change the model (including the weights that determine its behavior) without detection. Google invests significantly in securing the open source software world, including support for SLSA and the Sigstore project. In 2023, we [open sourced](#) our work to apply existing solutions such as SLSA and Sigstore to AI.



As well as securing our AI products, we strive to support customers in [using AI to bolster their security](#) capabilities. Our [use of AI in security offerings](#) combines world-class threat intelligence with point-in-time incident analysis and threat detections and analytics to help:



Prevent new infections



Make security more understandable while helping to improve its effectiveness



Reduce the number of tools organizations need to secure their vast attack surface areas




Empower systems to secure themselves

Customers using Vertex AI can also benefit from [Sensitive Data Protection](#), which enables the identification of sensitive data such as email addresses, phone numbers, and job titles, to name a few, based on a pattern or a list, and then automatically hides or transforms that data by using methods such as masking or tokenization. This tool can also be used to redact sensitive data, such as social security numbers, from images before ingesting it into a machine learning training environment.


In addition to building our AI products on a secure platform, we've also developed the Secure AI Framework ([SAIF](#)), which is a conceptual framework for securing AI systems across the four dimensions that make up an AI system: data, infrastructure, application, and model.

SAIF is inspired by the security best practices – like reviewing, testing and controlling the supply chain – that we apply to software development, while incorporating our understanding of [security megatrends](#) and risks specific to AI systems. It offers a practical approach to address the concerns that are top of mind for customers, including security, AI/ML model risk management, privacy, compliance and others. Customers may wish to consider SAIF as they define and refine their approach to adopting AI.


A framework like [SAIE](#), which spans the public and private sectors, is essential for safeguarding the technology that supports AI advancements, so that when AI models are implemented, they're secure-by-default. The [six core elements](#) of SAIF include:




Expand strong security foundations to the AI ecosystem




Extend detection and response to bring AI into an organization's threat model




Automate defenses to keep pace with existing and new threats



Harmonize platform level controls to ensure consistent security across the organization



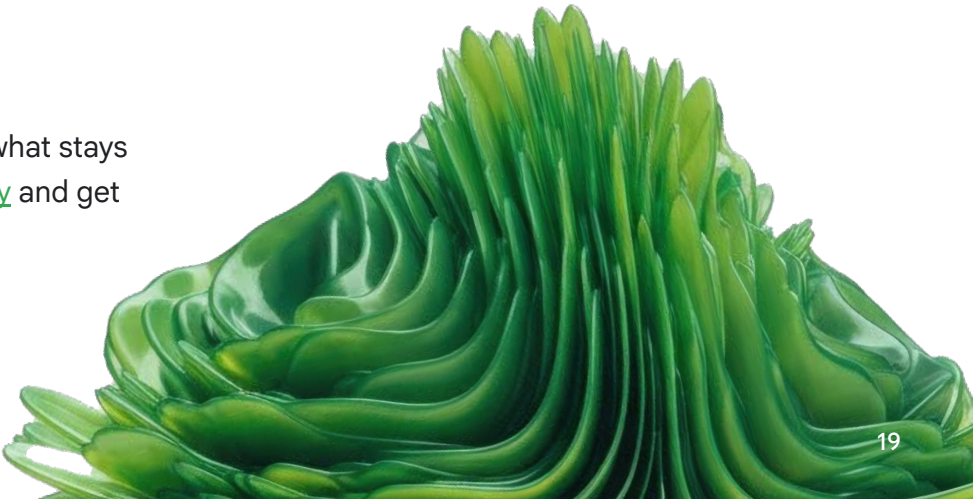
Adapt controls to adjust mitigations and create faster feedback loops for AI deployment



Contextualize AI system risks in surrounding business processes

These steps aren't simply conceptual. Rather, it's a framework that works for all. It's also important to consider that while there are novel aspects to securing AI, many of the current approaches to developing, deploying, and utilizing AI systems can be adjusted to account for these core elements rather than requiring a completely new approach.

To learn more, explore what changes and what stays the same when it comes to [AI cybersecurity](#) and get best practices on [securely deploying AI](#).



In addition, [Google's AI red team](#) focused on testing for security and privacy risks. Red teaming, also referred to as adversarial testing, is a technique where “ethical hackers” intentionally violate policies for the purpose of discovering and addressing vulnerabilities which could harm users. With the rise of gen AI, it has become a useful tool to help teams systematically improve models and products, and to inform launch decisions.

To expand on these efforts to address content safety risks, we’ve built a new team to use adversarial testing techniques to identify new and unexpected patterns on gen AI products. We also offer [Mandiant AI Security Consulting Services](#) to help organizations safeguard the use of AI systems as well as utilize AI to enhance cyber defenses.

This team explores innovative uses of AI to augment and expand existing testing efforts. Applying the adversarial approach we use during pre-launch responsibility evaluations to post-launch evaluations helps us improve model performance based on user feedback and helps us identify emerging risks.

Across the development and deployment lifecycle of our AI technologies, we use robust security and safety controls, which we adapt to risks for specific products and users. This is important as it enables early inclusion of prevention and detection controls, augmented by adversarial testing and red teaming. We also use threat intelligence to stay abreast of novel attacks. Our models are developed, trained, and stored within Google’s infrastructure, supported by our global teams of security engineers.

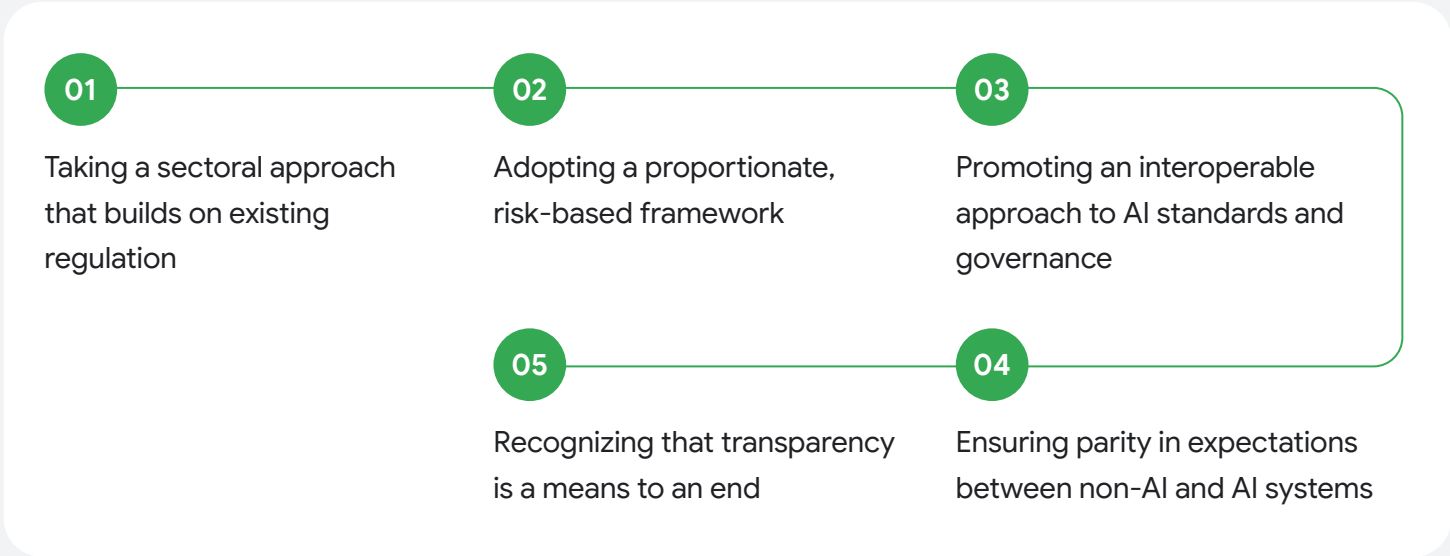
Compliance

Security and privacy in cloud computing are subject to legal, regulatory compliance, and risk management requirements. This is particularly the case in regulated industries such as financial services and healthcare, and for certain critical service or critical infrastructure providers. Organizations running workloads and storing data on Google Cloud rightfully seek assurances as to the platform’s controls posture, frequently requiring documentation from an independent third party to validate their existence and efficacy. To provide transparency into its controls, Google Cloud makes compliance documentation, certifications, control attestations, and independent

audit reports [readily available](#) to satisfy regional and industry-specific requirements and support customers in their compliance validation efforts of the Google Cloud platform, as well as their assessment of [Vertex AI’s compliance and security controls](#).



The rapid advance of AI has captured regulators’ attention worldwide, who are increasingly interested in understanding how current regulatory frameworks address AI and what new measures might be necessary in order to ensure AI is developed and deployed in a way that respects laws, norms, and human rights. We believe that AI is too important not to regulate, and too important not to regulate well, and thus advocate for risk-based frameworks that reflect the complexity of the AI ecosystem by building on existing general concepts. We previously published [recommendations for regulating AI](#) which outlined a general approach and some key implementation practicalities for policymakers to consider in developing practical AI regulations. Our top line recommendations included:



Since then, we’ve further clarified our position, publishing [A Policy Agenda for Responsible Progress in Artificial Intelligence](#), [Applying Model Risk Management Guidance to Artificial Intelligence/Machine Learning-based Risk Models](#), and [Generative AI Risk Management in Financial Institutions](#). In addition, our teams have adopted a risk assessment process to help:

- ✓ Identify, measure, and analyze ethical risks throughout the life of an AI-powered product
- ✓ Map these risks to appropriate mitigations
- ✓ Develop clearer standards of acceptable risk

Google Cloud is a trusted voice in the international and regional standards development community. We actively provide feedback and shape the regulations, standards, and framework.

We also closely track, monitor, and actively support industry standards such as the recently published ISO/IEC 42001 AI Management System Standard, NIST AI Risk Management Framework (RMF), as well as global regulatory developments to ensure we continue to develop and deliver tools that serve our customers’ needs.

We understand that AI comes with complexities and risks, and to ensure our readiness for the future landscape of AI compliance we proactively benchmark ourselves against emerging AI governance frameworks. To put our commitments into practice, we invited [Coalfire](#), a respected leader in cybersecurity, to examine our current processes, measure alignment and maturity toward the objectives defined in the National Institute of Standards and Technology (NIST) [Artificial Intelligence Risk Management Framework](#) (AI RMF) and International Organization for Standardization (ISO) [ISO/IEC 42001 standard](#). Coalfire's [assessment](#) provided valuable insights, allowing us to enhance our security posture as we continuously work to uphold the highest standards of data protection and privacy. We believe that an independent and external perspective offers critical objectivity, and we are

proud to be among the first organizations to perform a third-party AI readiness assessment.

Likewise, we closely monitor regulatory developments, such as the EU AI Act. The AI Act is a legal framework that establishes obligations for AI systems based on their potential risks and levels of impact. It will come into effect in phases and include bans on certain practices, general-purpose AI rules, and obligations for high-risk systems. Google is [actively preparing](#) for AI Act compliance. Internally, our AI Act readiness program is focused on ensuring our products and services align with the Act's requirements while continuing to deliver the innovative solutions our customers expect. This is a company-wide initiative that involves collaboration among a multitude of teams, including:



Legal and Policy

Thoroughly analyzing the AI Act's requirements and working to integrate them into our existing policies, practices, and contracts.



Risk and Compliance

Assessing and mitigating potential risks associated with AI Act compliance, ensuring robust processes are in place.



Product and Engineering

Ensuring our AI systems continue to be designed and built with the AI Act's principles of transparency, accountability, and fairness in mind and constantly improving the user experience, incorporating the AI Act's requirements for testing, monitoring, and documentation.



Customer engagement

Working closely with our customers to understand their needs and concerns regarding the AI Act, and providing guidance and support as needed.

Open Cloud & Portability

Google Cloud offers both first-party and third-party AI models in the Vertex AI Model Garden. This is part of our open philosophy which extends far beyond just the ability to utilize a variety of AI models. It encompasses a core belief in giving customers maximum choice without forced lock-in. This means allowing easy connection to existing on-premises systems, other clouds, SaaS applications, and even supporting businesses' proprietary models. Vertex AI offers a unified platform to manage, monitor, and continuously improve models regardless of their origin, whether that's open-source, partner-developed, or Google's in-house models.

This approach in turn supports portability that enables customers to take custom code, OSS code and containers anywhere. Model Garden simplifies the process of selecting (and switching between) the most options of models from any model provider. Vertex AI integrates with MLOps tools for streamlining the movement of models – be they first-party, third-party, or open-source – from development into production and ensure that changes are well-tracked. Model Garden encourages a modular approach, making it easier to swap out individual components within your ML pipelines, promoting experimentation and faster iteration. Together, this empowers you to make informed decisions based on your specific requirements and risk tolerance.



Environmental Impact

AI models and services can consume vast amounts of energy which raises the responsibility for managing the carbon footprint resulting from the computing power required to train and run foundation models. As part of our ongoing work, we have identified four best practices that reduce energy and carbon emissions significantly, which we refer to as the “4Ms,” all of which are being used today and are available to anyone using Google Cloud services. These four practices, each of which is briefly noted below, can when implemented together, reduce energy by 100x and emissions by 1000x.



Model

Selecting efficient ML model architectures, such as sparse models, can advance ML quality while reducing computation by 3x–10x



Machine

Using processors and systems optimized for ML training, versus general-purpose processors, can improve performance and energy efficiency by 2x–5x.



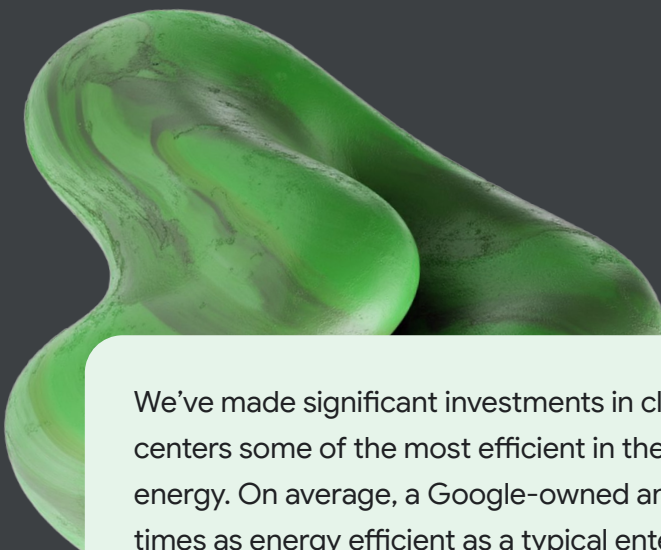
Map Optimization

Moreover, the cloud enables customers to pick the location with the cleanest energy, further reducing the gross carbon footprint by 5x–10x.



Mechanization

Computing in the Cloud rather than on premise reduces energy usage and therefore emissions by 1.4x–2x. Cloud-based data centers are new, custom-designed warehouses equipped for energy efficiency for 50,000 servers, resulting in very good power usage effectiveness (PUE). On-premise data centers are often older and smaller and thus cannot amortize the cost of new energy-efficient cooling and power distribution systems.



We’ve made significant investments in cleaner cloud computing by making our data centers some of the most efficient in the world and sourcing more carbon-free energy. On average, a Google-owned and -operated data center is more than 1.5 times as energy efficient as a typical enterprise data center and, compared with five years ago, we now deliver approximately three times as much computing power with the same amount of electrical power. Put simply, we can do more with less energy.



We're constantly looking for new ways to build products, design out waste and pollution, and keep materials and resources in use for as long as possible. We aim to maximize the reuse of finite resources across our operations, products, and supply chains and to enable others to do the same. We're helping our customers make real-time decisions to reduce emissions, and mitigate climate risks with data and AI. For example, Google Cloud customers can reduce their cloud footprint with a feature called Active Assist, which uses machine learning to identify unused (and potentially wasteful) workloads that could reduce carbon emissions if removed.

AI and machine learning workloads are quickly becoming larger and more capable, raising concerns about their energy use and their impact on the environment. With AI at an inflection point, predicting the future growth of energy use and emissions from AI compute in our data centers is challenging. Historically, research has shown that as AI/ML compute demand has gone up, the energy needed to power this technology has increased at a much slower rate than many forecasts predicted.

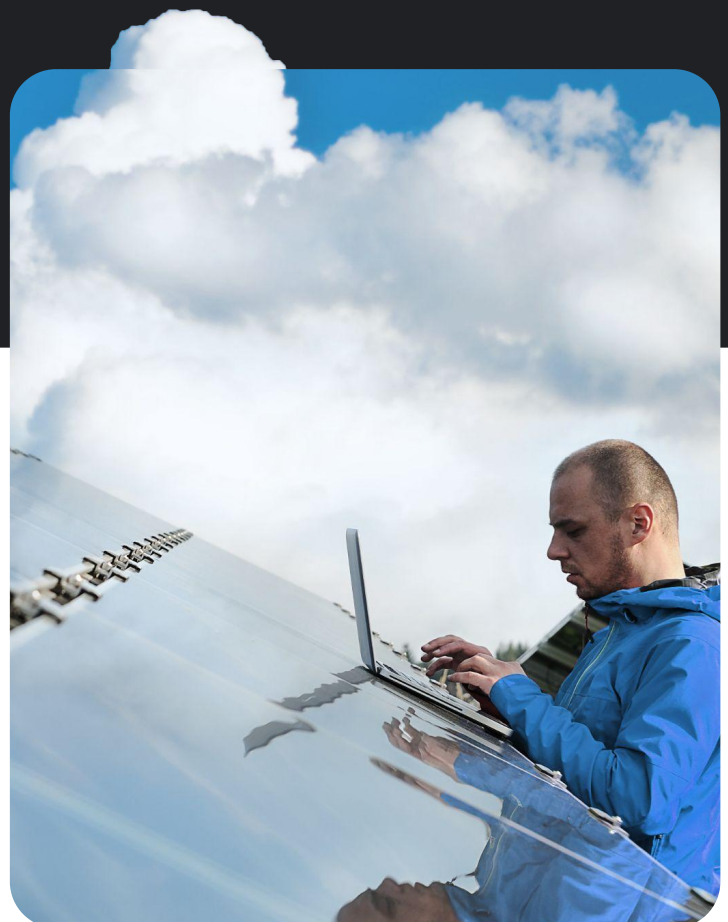
We have used tested practices to reduce the carbon footprint of workloads by large margins; together these principles have reduced the energy of training a model by up to **100x** and **emissions by up to 1,000x**. We plan to continue applying these tested practices and to keep developing new ways to make AI computing more efficient. Google data centers are designed, built, and operated to maximize efficiency – even as computing demand grows.

Environmental footprint

Leveraging AI to optimize our own operations, and working to reduce energy use and emissions from AI computing in our data centers.

We are committed to operating carbon-free by 2030 and replenishing 120% of the water we consume by 2030.

We're also dedicated to raising our standard of water stewardship, improving water quality and security, and restoring the health of ecosystems in the communities in which we operate.





Shared Fate

As we continue to develop our AI platform, systems, and foundational models, our belief in shared fate and our experience in using these technologies guides us to invest in end-to-end governance tools, opinionated guidance, and best practices to help our customers keep their data and AI models safe.



At Google Cloud, we are committed to helping enterprises develop effective AI risk management strategies to be able to use the full potential of gen AI. While risk profiles are often complex, this is especially true for gen AI because of how intricate the models can be. Importantly, [risk management can vary](#) depending on how the organization has chosen to use AI – Is it developing its own AI applications, using AI applications developed by a third party (including those developed by Google Cloud), or a mix? How enterprise-ready those services are is also a factor.



Both Google and the customer play essential roles in [safeguarding AI systems](#). Understanding and exercising [your key responsibilities](#) is critical for securing AI workloads on Google Cloud. Our shared fate approach ensures that you have tools and guidance from Google to help along the way.



Looking through the lens of how a customer might participate in the AI ecosystem, we see [four basic scenarios](#) that require different risk management strategies: build it yourself, customize the model to your needs, integrate the model as-is, or consume the model out of the box. A key difference between these four scenarios is the level of direct control an organization has over the AI model, as compared to what is outsourced to an external provider, such as Google Cloud. Across all four scenarios, customers can rely on Google Cloud to uphold our strong [AI privacy commitments](#) and to [protect customers' data](#), enabling them to pursue data-rich use cases while [complying](#) with relevant regulations and laws.



Chapter

03

Best Practices

How can your organization build trust into every AI application you use? In this chapter, discover the role that governance, acceptable use policies, security, and privacy play in successfully advancing AI in your organization; and see how you can stay on top of the latest gen AI developments.

Governance

Organizations can successfully implement AI by following [best practices](#) like identifying stakeholders, defining principles, utilizing frameworks, documenting policies, articulating use cases, leveraging data governance, collaborating with relevant departments, establishing escalation points, providing status visibility, and implementing an AI training program. These practices help organizations navigate AI implementation challenges and ensure responsible integration.

Acceptable Use

Organizations that want to use AI in a safe, secure, dependable, and robust way should devise their own “building code” for gen AI through an internal [Acceptable Use Policy](#) (AUP). It’s important to align the use of Gen AI to an organization’s overall goals and values, as well as the broader regional and industry requirements that may apply. An AUP can be an important, multifaceted guide in shaping any organization’s governance structure and its relationship to gen AI because it ties into other organizational governance pillars, including broad-scale awareness campaigns, training, and ongoing monitoring for compliance.

Security

While gen AI does represent a new security world, it’s not the end of the old security world, either. [Securing AI](#) does not magically upend security best practices, and much of the wisdom that security teams have learned is still correct and applicable. We believe that many of the security principles and practices that apply to traditional systems also apply to AI systems. By understanding the [differences between securing a traditional enterprise software system and an AI system](#), organizations can develop a more comprehensive security strategy to protect their AI systems from a variety of security threats. Now is also the time to [take steps](#) to prevent potential attacks from happening in the first place. When securing AI systems, it is important to think like an attacker. Consider known weaknesses and identify the ways that an attacker could exploit a system. Work with other teams in the organization – including data science, engineering, and security – to develop a comprehensive security.

Privacy and Data Governance

ML models learn from training data and make predictions on input data. Sometimes the training data, input data, or both can be quite sensitive. Although there may be benefits to building a model that operates on sensitive data, it's essential to consider the potential privacy implications in using sensitive data.

This includes not only respecting the legal and regulatory requirements, but also considering social norms and typical individual expectations. It's essential to offer users transparency and control of their data.

Fortunately, the possibility that ML models reveal underlying data can be minimized by appropriately applying various techniques, some of which include:



Identify whether your ML model can be trained without the use of sensitive data, e.g., by utilizing non-sensitive data collection or removing sensitive data from the training set.



If it is essential to process sensitive training data, strive to minimize the use of such data. Handle any sensitive data with care: e.g., comply with required laws and standards, provide users with clear notice and give them any necessary controls over data use, follow best practices such as encryption in transit and rest, and adhere to privacy principles such as the ones found on the [Google Cloud Privacy Resource Center](#).

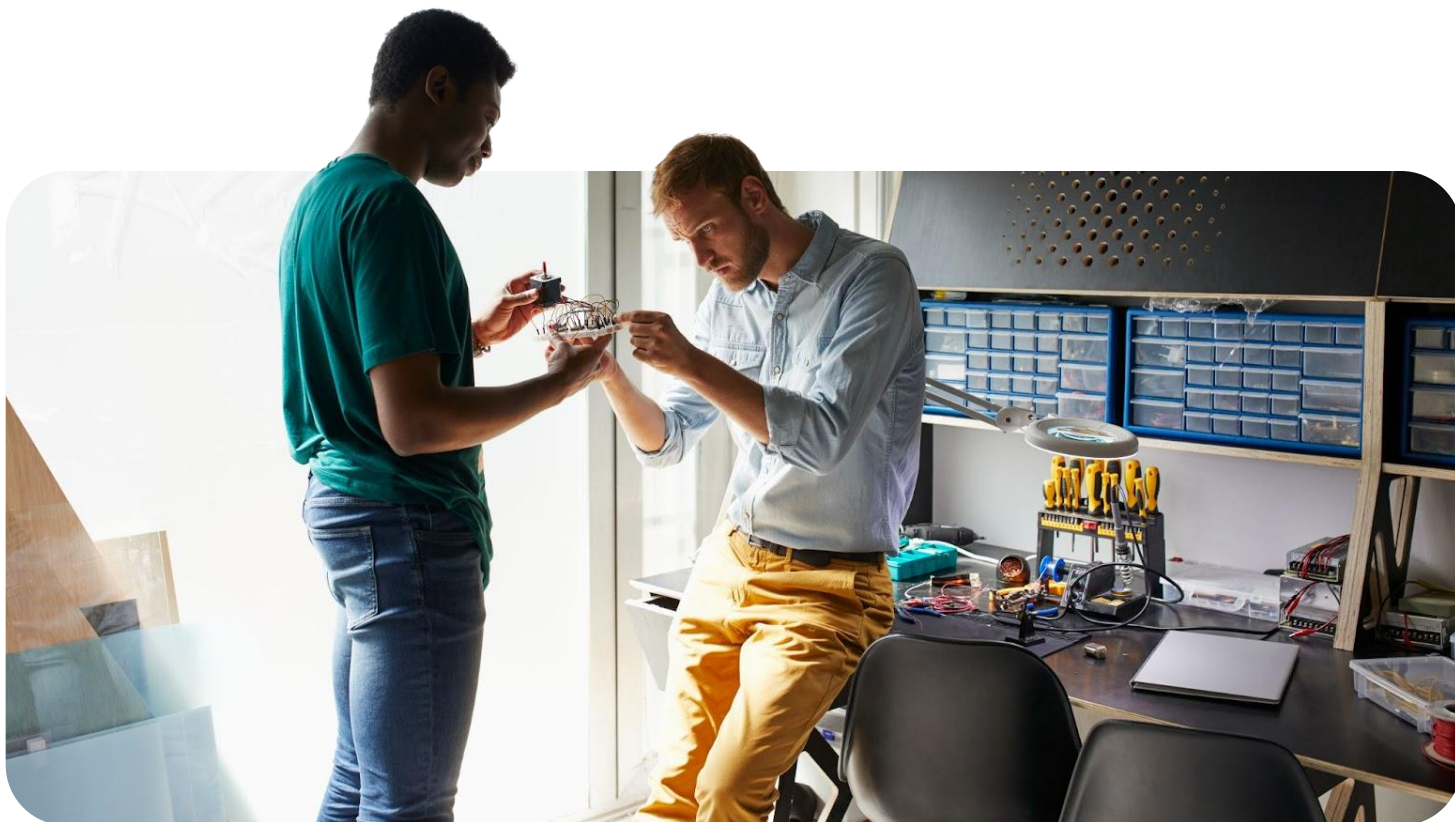


Anonymize and aggregate incoming data using best practice data-scrubbing pipelines: e.g., consider removing personally identifiable information (PII) and outlier or metadata values that might allow de-anonymization, for example by using the [Cloud Data Loss Prevention](#) API to automatically discover and redact sensitive and identifying data.

⚡ Staying on Top of Gen AI Developments

We're often asked how to [stay on top of AI developments](#), both technological and regulatory, and how to empower teams with the knowledge, skills, and an understanding of the risks in using gen AI. When approaching how to enable your workforce for gen AI adoption, it's important to recognize it isn't just about technology, but about investing in your people. By demystifying gen AI, focusing on strategic skills building, and creating a culture that values continuous learning, your enterprise can unlock the full potential of gen AI as a transformative technology and prepare for the future of work. It is critical that both IT and business teams understand how gen AI works, how these risks materialize, and what to do about them.

We believe the best way to learn gen AI is to actually use the models – experiment with them, spend time with them, and apply them in your work. [Learning paths](#) are also available to provide customers with a wide range of upskilling offers for different roles and expertise levels on, for example, gen AI concepts, fundamentals of large language models, and responsible AI Principles. In addition, our [recommended practices for AI](#) are a helpful guide to follow when designing, developing, testing and using AI systems with a focus on fairness, interpretability, privacy, safety and security, including relevant examples and documentation for the implementation of each.



Conclusion



We aim to be at the forefront of advancing AI through our deep research to develop more capable and useful AI. We're pursuing innovations that will help unlock scientific discoveries and tackle some of humanity's greatest challenges. From this research and development, we are bringing innovations into the world to assist people and benefit society everywhere through our infrastructure, tools, products, and services, and by enabling and working with others to benefit society.

To further the dialogue, we publish educational content, research and other forms of documentation to enable transparency and support our customers. These include [Responsible AI Guides](#) with best practices to assist customers in defining AI use cases and assessing their impact, and [AI research](#) on a range of topics including machine intelligence, natural language processing, and many others that maximize both scientific and real-world impact.

Contact us

Resources

Introduction

- [Generative AI Examples | Google Cloud](#)
- [Introduction to Vertex AI | Google Cloud](#)
- [Multimodal AI | Google Cloud](#)
- [AI Principles](#)
- [Multimodal AI | Google Cloud](#)
- [Google Cloud Platform AUP](#)
- [Generative AI Prohibited Use Policy](#)
- [Responsible AI](#)
- [Our views on AI policy – Google AI](#)
- [AI Principles Progress Update 2023](#)

01. Responsible innovation

- [Responsible AI Progress Report](#)
- [Our views on AI policy – Google AI](#)
- [End-to-end responsibility](#)
- [Fulfilling the Voluntary Industry Commitments on AI](#)
- [Machine Learning Glossary: Reinforcement Learning | Google for Developers](#)
- [Feedback + Control](#)
- [Google's Secure AI Framework \(SAIF\)](#)

🔍 Risk Assessment

- [Introduction to Vertex Explainable AI](#)
- [Introduction to model evaluation for fairness | Vertex AI | Google Cloud](#)
- [Improving model quality at scale with Vertex AI Model Evaluation | Google Cloud Blog](#)
- [Introduction to Vertex AI Model Monitoring | Google Cloud](#)
- [Vertex AI Model Registry | Google Cloud Blog](#)
- [Responsible AI | Generative AI on Vertex AI | Google Cloud](#)
- [Responsible Generative AI Toolkit | Google AI for Developers](#)
- [Moderate text | Cloud Natural Language API](#)
- [Model evaluation in Vertex AI | Google Cloud](#)
- [Introduction to model evaluation for fairness | Vertex AI | Google Cloud](#)
- [About the API - Model Cards](#)
- [Gemma model card | Google AI for Developers](#)
- [Gemini: A Family of Highly Capable Multimodal Models](#)
- [Google Model Cards](#)
- [Google's Secure AI Framework](#)

Data Governance

- [Generative AI beginner's guide | Generative AI on Vertex AI | Google Cloud](#)
- [Generative AI and data governance | Generative AI on Vertex AI | Google Cloud](#)
- [Adaptation of Large Foundation Models](#)
- [Ground gen AI outputs in Google Search and enterprise data](#)

Privacy

- [Generative AI and Privacy - Policy Recommendations Working Paper](#)
- [Designing for privacy in an AI world](#)
- [AI Principles](#)
- [Generative AI, Privacy, and Google Cloud](#)
- [Google Cloud Platform/SecOps Terms of Service](#)
- [Cloud Data Processing Addendum](#)
- [Access Transparency in Vertex AI | Google Cloud](#)
- [Vertex AI access control with IAM | Google Cloud](#)
- [How Sensitive Data Protection can help secure generative AI workloads | Google Cloud Blog](#)
- [Protect your resources with VPC Service Controls](#)
- [GDPR and Google Cloud](#)
- [Data Protection Impact Assessment \(DPIA\) | Google Cloud](#)
- [Google Cloud Platform Services Data Residency](#)
- [Service Specific Terms | Google Cloud](#)


Security

- [How AI can improve digital security | Google Cloud Blog](#)
- [Secure your AI with Google Cloud](#)
- [Google Cloud security best practices center](#)
- [Announcing AI Protection: Security for the AI Era | Google Cloud Blog](#)
- [The Defender's Advantage: Using AI in Cyber Defense](#)
- [Google's Secure AI Framework \(SAIF\)](#)
- [Trusting your data with Google Cloud](#)
- [Trusted Cloud Infrastructure \(IaaS\)](#)
- [Security controls for Vertex AI | Google Cloud](#)
- [Securing the AI Software Supply Chain](#)
- [Increasing transparency in AI security](#)
- [Supercharge security with AI](#)
- [Sensitive Data Protection](#)
- [Google's Secure AI Framework \(SAIF\)](#)
- [9 megatrends drive cloud adoption—and improve security for all | Google Cloud Blog](#)
- [Securing AI: Similar or Different?](#)
- [Best Practices for Securely Deploying AI on Google Cloud](#)
- [Google's AI Red Team: the ethical hackers making AI safer](#)

Compliance

- [Cloud Compliance - Regulations & Certifications | Google Cloud](#)
- [Introduction to Vertex AI | Google Cloud](#)
- [Recommendations for Regulating AI](#)
- [A Policy Agenda for Responsible Progress in Artificial Intelligence](#)
- [Applying model risk management guidance to artificial intelligence/ machine learning-based risk models](#)
- [Generative AI Risk Management in Financial Institutions](#)
- [Coalfire](#)
- [AI Risk Management Framework | NIST.](#)
- [ISO/IEC 42001:2023 - AI management systems](#)
- [Coalfire Partners with Google Cloud to Assess AI Governance and Security Risks against NIST AI RMF and ISO/IEC 42001](#)
- [AI Risk Management Framework | NIST.](#)
- [ISO/IEC 42001:2023 - AI management systems](#)
- [Coalfire Partners with Google Cloud to Assess AI Governance and Security Risks against NIST AI RMF and ISO/IEC 42001](#)
- [Navigating the EU AI Act: Google Cloud's proactive approach](#)

Environmental Impact

-  [Good News About the Carbon Footprint of Machine Learning Training](#)
- [Constructing Transformers For Longer Sequences with Sparse Attention Methods](#)
- [Power usage effectiveness - Wikipedia](#)
- [Efficiency - Data Centers - Google](#)
- [A Circular Google](#)
- [Realizing a carbon-free future: Google's Third Decade of Climate Action](#)
- [Restoring Ecosystems through Water Stewardship - Google Sustainability](#)

02. Shared fate

- [Best Practices for Securely Deploying AI on Google Cloud](#)
- [Vertex AI shared responsibility | Google Cloud.](#)
- [Best Practices for Securely Deploying AI on Google Cloud](#)
- [From turnkey to custom: Tailor your AI risk governance to help build confidence | Google Cloud Blog](#)
- [Generative AI, Privacy, and Google Cloud](#)
- [Security | Google Cloud](#)
- [Cloud compliance and regulations resources](#)

03. Best Practices

Governance

- [Gen AI governance: 10 tips to level up your AI program | Google Cloud Blog](#)

Acceptable use

- [How to craft an Acceptable Use Policy for gen AI \(and look smart doing it\) | Google Cloud Blog](#)

Security

- [Securing AI: Similar or Different?](#)
- [The Prompt: Insights from our AI Red Team's first report \(Q&A\) | Google Cloud Blog](#)

Privacy and Data Governance

- [Privacy Resource Center | Google Cloud](#)
- [Cloud Data Loss Prevention](#)

Staying on Top of Gen AI Developments

- [Staying on top of AI Developments - Google Cloud Community](#)
- [Machine Learning Engineer Learning Path | Google Cloud Skills Boost](#)
- [AI Principles](#)

Conclusion

- [Responsible AI | Generative AI on Vertex AI | Google Cloud](#)
- [Cloud AI](#)