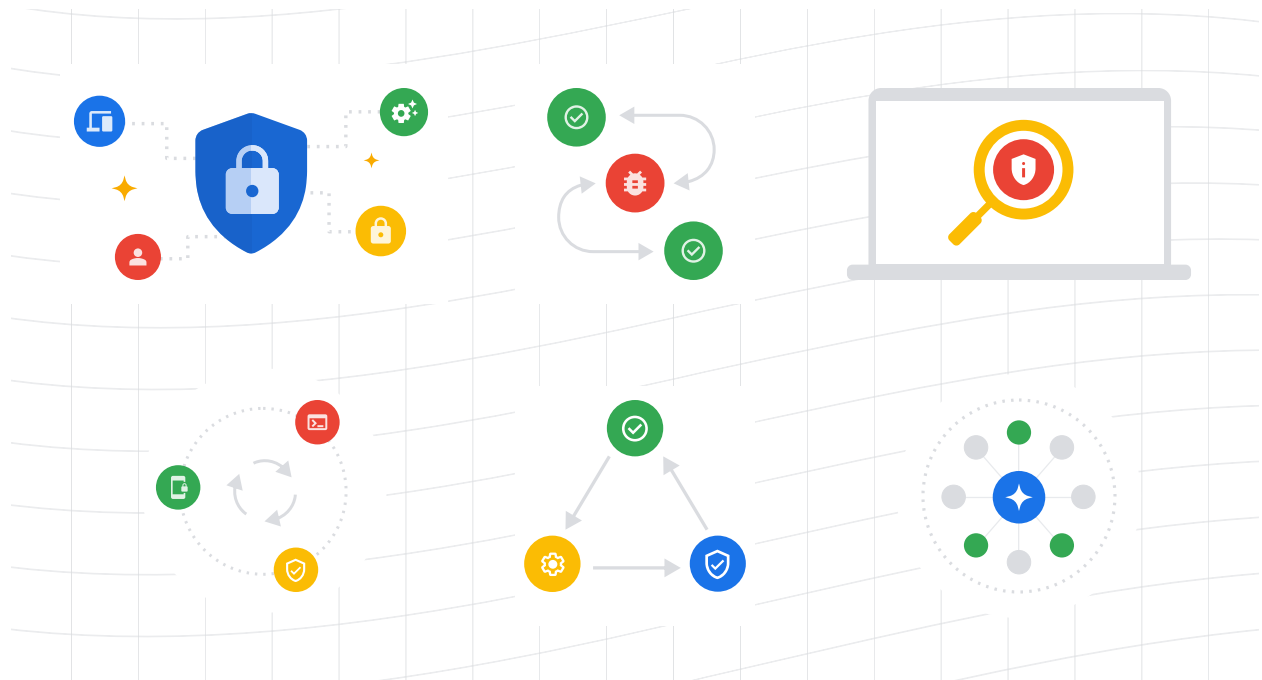




# Secure AI Framework Approach

A quick guide to implementing the Secure AI Framework (SAIF)



# Table of contents

Intro	2
Putting SAIF into practice	3
Step 1 - Understand the use	3
Step 2 - Assemble the team	3
Step 3 - Level set with an AI primer	4
Step 4 - Apply the six core elements of SAIF	4
Expand strong security foundations to the AI ecosystem	4
Extend detection and response to bring AI into an organization's threat universe	6
Automate defenses to keep pace with existing and new threats	7
Harmonize platform level controls to ensure consistent security across the organization	7
Adapt controls to adjust mitigations and create faster feedback loops for AI deployment	8
Contextualize AI system risks in surrounding business processes	9
Conclusion	11

# Intro

Secure AI Framework (SAIF) is a conceptual framework for secure artificial intelligence (AI) systems. It is inspired by security best practices — like reviewing, testing and controlling the supply chain — that Google has applied to software development, while incorporating our understanding of security mega-trends and risks specific to AI systems. SAIF offers a practical approach to address the concerns that are top of mind for security and risk professionals, such as:

- Security
  - a) Access management
  - b) Network / endpoint security
  - c) Application / product security
  - d) Supply chain attacks
  - e) Data security
  - f) AI specific threats
  - g) Threat detection and response
- AI/ML model risk management
  - a) Model transparency and accountability
  - b) Error-prone manual reviews for detecting anomalies
  - c) Data poisoning
  - d) Data lineage, retention and governance controls
- Privacy and compliance
  - a) Data privacy and usage of sensitive data
  - b) Emerging regulations
- People and organization
  - a) Talent gap
  - b) Governance / Board reporting

This quick guide is intended to provide high level practical considerations on how organizations could go about building the SAIF approach into their existing or new adoptions of AI. Further content will delve deeper into the topics - for now we focus on the priority elements that need to be addressed under each of the six core elements of SAIF:

- 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

# Putting SAIF into practice

## Step 1 - Understand the use

Many organizations are considering using AI for the first time, or expanding the AI solutions they have to take advantage of new Generative AI (Gen AI) capabilities. In all cases, understanding the specific business problem AI will solve and the data needed to train the model, will help drive the policy, protocols, and controls that need to be implemented as part of SAIF. For example, models designed to analyze or act on existing data, such as models that summarize an analyst report or detect fraud, may implicate fewer complex issues compared to models that, for example, are used to make predictions for consumer finances (e.g. credit risk models), which would raise additional challenges due to the potential impact on consumers and applicable consumer protection obligations. The context of how the models interact with end-users also plays an important role. For example, an AI model exposed externally, that takes in end-user input, will have different requirements for security and data governance compared to a model used for trading stocks. Along with that, using a pre-built model from a third party vs. developing and / or training your own will also have different implications related to securing the infrastructure and the development platform, monitoring model behavior and outcome, threat detection and protections.

Thus, thoroughly understanding the AI use case will ensure the implementation of the SAIF captures the complexities and risks of the particular deployment.

## Step 2 - Assemble the team

Developing and deploying AI systems, just like traditional systems, are multidisciplinary efforts and include similar elements, such as risk assessment, security / privacy / compliance controls, threat modeling, and incident response. Additionally, AI systems are often complex and opaque, have a large number of moving parts, rely on large amounts of data, are resource intensive, can be used to apply judgment-based decisions, and can generate novel content that may be offensive, harmful, or can perpetuate stereotypes and social biases. This expands the composition of the team to include stakeholders across multiple organizations, such as:

- Business use case owners
- Security
- Cloud Engineering
- Risk and Audit teams
- Privacy
- Legal
- Data Science teams
- Development teams
- Responsible AI and Ethics

Establishing the right cross-functional team ensures that security, privacy, risk, and compliance considerations are included from the start and not added after the fact.

## Step 3 - Level set with an AI primer

AI, especially [Gen AI](#), is a still-emerging and rapidly developing technology. As teams embark on evaluating the business use, the various and evolving complexities, risks, and security controls that apply - it is critical that parties involved understand the basics of the AI model development lifecycle, the design and logic of the model methodologies, including capabilities, merits and limitations. Starting with concepts like AI, machine learning (ML), Deep Learning, Gen AI, large language models (LLMs), etc., will level set and allow non-technical stakeholders to accurately capture and evaluate the risks and controls required to manage and deploy AI safely and responsibly.

## Step 4 - Apply the six core elements of SAIF

Once the use cases and context are known, the team has been assembled and primed on AI, you can start to apply the six elements of SAIF to address some of the concerns mentioned earlier. It should be noted that the elements are not intended to be applied in chronological order, rather levers that collectively guide the organizations to build and deploy AI systems in a secure and responsible manner.

### Expand strong security foundations to the AI ecosystem

#### **Review what existing security controls across the security domains apply to AI systems**

Existing security controls across the security domains apply to AI systems in a number of ways. For example, data security controls can be used to protect the data that AI systems use to train and operate. Application security controls can be used to protect the software that AI systems are implemented in; infrastructure security controls can be used to protect the underlying infrastructure that AI systems rely on; and operational security controls can be used to ensure that AI systems are operated in a secure manner.

The specific controls that are needed will vary depending on the use of AI, as well as the specific AI systems and environments.

#### **Evaluate the relevance of traditional controls to AI threats and risks using available frameworks**

Traditional security controls can be relevant to AI threats and risks, but they may need to be adapted to be effective, or additional layers added to the defense posture to help cover the AI specific risks. For example, data encryption can help to protect AI systems from unauthorized

access by limiting the access of the keys to certain roles, but it may also need to be used to protect AI models and their underpinning data from being stolen or tampered with.

**Perform an analysis to determine what security controls need to be added due to AI specific threats, regulations, etc.**

Using the assembled team, review how your current controls map to your AI use case, do a fit for purpose evaluation of these controls and then create a plan to address the gap areas. Once all of that is done, also measure the effectiveness of these controls based on whether they lower the risk and how well they address your intended AI usage.

**Prepare to store and track supply chain assets, code and training data**

Organizations that use AI systems must prepare to store and track supply chain assets, code, and training data. This includes identifying, categorizing, and securing all assets, as well as monitoring for unauthorized access or use. By taking these steps, organizations can help protect their AI systems from attack.

Ensure your data governance and lifecycle management are scalable and adapted to AI.

Depending on the definition of data governance you follow, there are up to six decision domains for data governance:

- Data quality
- Data security
- Data architecture
- Metadata
- Data lifecycle
- Data storage

AI data governance will become more important than ever. For example, a key underpinning of the effectiveness of AI models are the training sets of data. Ensure that you have a proper lifecycle management system when it comes to data sets, with a strong emphasis on security as part of the lifecycle (i.e. have security measures from creation of data to the ultimate destruction of data embedded throughout the lifecycle). Data lineage will also play a key part and help to answer questions with regards to privacy and intellectual property. If you know who created the data, where it came from, and what makes up the dataset, it is much easier to answer questions on the aforementioned topics.

As AI adoption grows, your organization's success will likely hinge on scaling these decision domains in an agile manner. To help support this effort, it is critical to review your data governance strategy with a cross functional team and potentially adjust it to ensure it reflects advances in AI.

## **Retain and retrain**

We are not talking about AI, but rather people. For many organizations, finding the right talent in security, privacy and compliance can be a multi-year journey. Taking steps to retain this talent can add to your success, as they can be retrained with skills relevant to AI quicker than hiring talent externally that may have the specific AI knowledge, but lack the institutional knowledge that can take longer to acquire.

## **Extend detection and response to bring AI into an organization's threat universe**

**Develop understanding of threats that matter for AI usage scenarios, the types of AI used, etc.**

Organizations that use AI systems must understand the threats relevant to their specific AI usage scenarios. This includes understanding the types of AI they use, the data they use to train AI systems, and the potential consequences of a security breach. By taking these steps, organizations can help protect their AI systems from attack.

**Prepare to respond to attacks against AI and also to issues raised by AI output**

Organizations that use AI systems must have a plan for detecting and responding to security incidents, and mitigate the risks of AI systems making harmful or biased decisions. By taking these steps, organizations can help protect their AI systems and users from harm.

**Specifically, for Gen AI, focus on AI output - prepare to enforce content safety policies**

Gen AI is a powerful tool for creating a variety of content, from text to images to videos. However, this power also comes with the potential for abuse. For example, Gen AI could be used to create harmful content, such as hate speech or violent images. To mitigate these risks, it is important to prepare to enforce content safety policies.

**Adjust your abuse policy and incident response processes to AI-specific incident types, such as malicious content creation or AI privacy violations**

As AI systems become more complex and pervasive, it is important to adjust your abuse policy to deal with use cases of abuse and then also adjust your incident response processes to account for AI-specific incident types. These types of incidents can include malicious content creation, AI privacy violations, AI bias and general abuse of the system.

## Automate defenses to keep pace with existing and new threats

**Identify the list of AI security capabilities focused on securing AI systems, training data pipelines, etc.**

AI security technologies can protect AI systems from a variety of threats, including data breaches, malicious content creation, and AI bias. Some of these technologies include traditional data encryption, access control, auditing which can be augmented with AI and newer technologies that can perform training data protection, and model protection.

**Use AI defenses to counter AI threats, but keep humans in the loop for decisions when necessary**

AI can be used to detect and respond to AI threats, such as data breaches, malicious content creation, and AI bias. However, humans must remain in the loop for important decisions, such as determining what constitutes a threat and how to respond to it. This is because AI systems can be biased or make mistakes, and human oversight is necessary to ensure that AI systems are used ethically and responsibly.

**Use AI to automate time consuming tasks, reduce toil, and speed up defensive mechanisms**

Although it seems like a more simplistic point in light of the uses for AI, using AI to speed up time consuming tasks will ultimately lead to faster outcomes. For example, it can be time consuming to reverse engineer a malware binary. However, AI can quickly review the relevant code and provide an analyst with actionable information. Using this information, the analyst could then ask the system to generate a YARA rule looking for these actions. In this example, there is an immediate reduction of toil and faster output for the defensive posture.

## Harmonize platform level controls to ensure consistent security across the organization

**Review usage of AI and life cycle of AI based apps**

As mentioned in Step 1, understanding the use of AI is a key component. Once AI becomes more widely used in your organization, you should implement a process for periodic review of usage to identify and mitigate security risks. This includes reviewing the types of AI models and applications being used, the data used to train and run AI models, the security measures in place to protect AI models and applications, the procedures for monitoring and responding to AI security incidents, and AI security risk awareness and training for all employees.

**Prevent fragmentation of controls by trying to standardize on tooling and frameworks**

With the above process in place, you can better understand the existing tooling, security controls, and frameworks currently in place. At the same time, it is important to examine whether your organization has different or overlapping frameworks for security and compliance controls to help



reduce fragmentation. Fragmentation will increase complexity and create significant overlap, increasing costs and inefficiencies. By harmonizing your frameworks and controls, and understanding their applicability to your AI usage context, you will limit fragmentation and provide a 'right fit' approach to controls to mitigate risk. This guidance primarily refers to existing control frameworks and standards, but the same principle (e.g. try to keep the overall number as small as possible) would apply to new and emerging frameworks and standards for AI.

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

### **Conduct Red Team exercises to improve safety and security for AI-powered products and capabilities**

Red Team exercises are a security testing method where a team of ethical hackers attempts to exploit vulnerabilities in an organization's systems and applications. This can help organizations identify and mitigate security risks in their AI systems before they can be exploited by malicious actors.

### **Stay on top of novel attacks including prompt injection, data poisoning and evasion attacks**

These attacks can exploit vulnerabilities in AI systems to cause harm, such as leaking sensitive data, making incorrect predictions, or disrupting operations. By staying up-to-date on the latest attack methods, organizations can take steps to mitigate these risks.

### **Apply machine learning techniques to improve detection accuracy and speed**

Although it is critical to focus on securing the use of AI, AI can also help organizations achieve better security outcomes at scale (see reference in Step 3 above). AI-assisted detection and response capabilities, for example, can be an important asset for any organization. At the same time, it is essential to keep humans in the loop to oversee relevant AI systems, processes, and decisions. Over time, this effort can drive continuous learning to improve AI base protections, update training and fine-tuning of data sets for foundation models, and the ML models used for building protections. In turn, this will enable organizations to strategically respond to attacks as the threat environment evolves. Continuous learning is also critical for improving accuracy, reducing latency and increasing efficiency of protections.

### **Create a feedback loop**

To maximize the impact of the above three elements, it is critical to create a feedback loop. For example, if your Red Team discovers a way to misuse your AI system, that information should be fed back into your organization to help improve defenses, rather than focusing solely on remediation. Similarly, if your organization discovers a new attack vector, it should be fed back into your training data set as part of continuous learning. To ensure that feedback is put to good use, it is important

to consider various ingestion avenues and have a good understanding of how quickly feedback can be incorporated into your protections.

## Contextualize AI system risks in surrounding business processes

### **Establish a model risk management framework and build a team that understands AI-related risks**

Organizations should develop a process for identifying, assessing, and mitigating the risks associated with AI models. The team should be composed of experts in AI, security, and risk management.

### **Build an inventory of AI models and their risk profile based on the specific use cases and shared responsibility when leveraging third-party solutions and services**

Organizations should build a comprehensive inventory of AI models and assess their risk profile based on the specific use cases, data sensitivity, and shared responsibility when leveraging third-party solutions and services. This means identifying all AI models in use, understanding the specific risks associated with each model, and implementing security controls to mitigate those risks along with having clear roles and responsibilities.

### **Implement data privacy, cyber risk, and third-party risk policies, protocols and controls throughout the ML model lifecycle to guide the model development, implementation, monitoring, and validation**

Organizations should implement data privacy, cyber risk, and third-party risk policies, protocols and controls throughout the ML model lifecycle to guide the model development, implementation, monitoring, and validation. This means developing and implementing policies, protocols, and controls that address the specific risks associated with each stage of the ML model lifecycle. Keep the fourth element of the framework above in mind to ensure you do not create undue fragmentation.

### **Perform a risk assessment that considers organizational use of AI**

Organizations should identify and assess the risks associated with the use of AI, and implement security controls to mitigate those risks. Organizations should also cover security practices to monitor and validate control effectiveness, including model output explainability and monitoring for drift. As referenced in Steps 1 and 2, it is important to create a cross functional team and build a deeper understanding of the relevant use cases to support this effort. Organizations can use existing frameworks for risk assessment to help guide their work, but will likely need to augment or adapt their approach to address new emerging AI risk management frameworks.

## **Incorporate the shared responsibility for securing AI depending on who develops AI systems, deploys models developed by model provider, tunes the models or uses off-the-shelf solutions**

The security of AI systems is a shared responsibility between the developers, deployers, and users of those systems. The specific responsibilities of each party will vary depending on their role in the development and deployment of the AI system. For example, the AI system developers are responsible for developing AI systems that are secure by design. This includes using secure coding practices, training AI models on clean data, and implementing security controls to protect AI systems from attack.

## **Match the AI use cases to risk tolerances**

This means understanding the specific risks associated with each AI use case and implementing security measures to mitigate those risks. For example, AI systems that are used to help make decisions that could significantly impact people's lives, such as healthcare or finance, will likely need to be more heavily secured than AI systems that are used for less urgent tasks, such as marketing or customer service.

## Conclusion

AI has captured the world's imagination and many organizations are seeing opportunities to boost creativity and improve productivity by leveraging this emerging technology. At Google, we've been bringing AI into our products and services for over a decade, and we remain committed to approaching it in a bold and responsible way.

SAIF is designed to help raise the security bar and reduce overall risk when developing and deploying AI systems. To ensure we enable secure-by-default AI advancements, it is important to work collaboratively. With support from customers, partners, industry and governments, we will continue to advance the core elements of the framework and offer practical and actionable resources to help organizations achieve better security outcomes at scale.