Office of the CISO

SAIF in the real world

Key considerations in applying the Secure AI Framework (SAIF) through the AI development lifecycle

Authors
Anton Chuvakin
Security Advisor, Office of the CISO, Google Cloud

John Stone
Director, Office of the CISO, Google Cloud

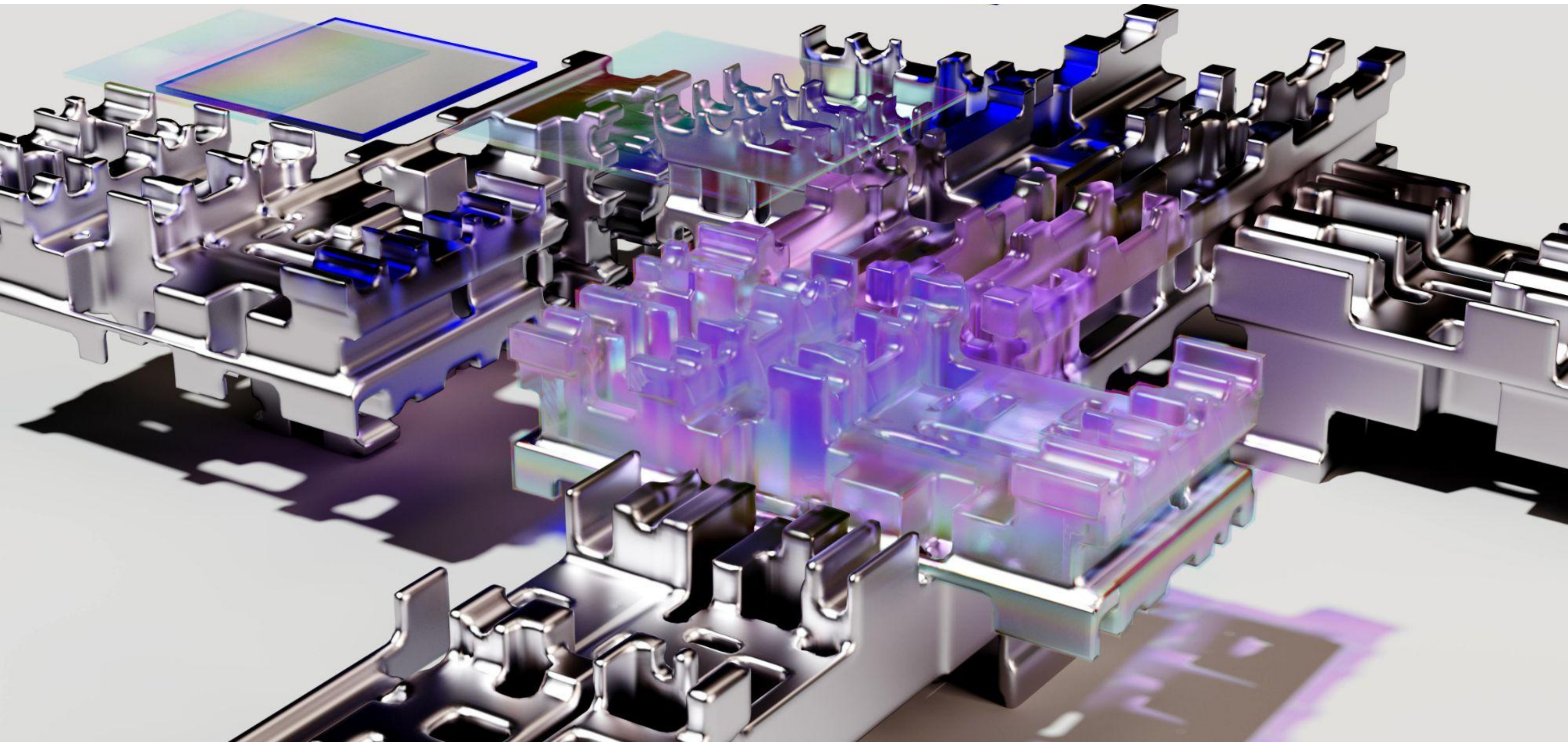


Table of contents

00 Intro 3

01 SAIF and industry frameworks 4

02 Categories and dimensions 5

Data 6
Infrastructure 6
Application 6
Model 7
When dimensions intersect 7
To build or buy (CSP vs. BYOAI) considerations 7
A word about threats 8
Model lifecycle refresh 8
Just in case it is all about the use case 9

03 Where are you with SAIF today? 10

Stage 1: Opportunity discovery and problem definition 11
Stage 2: Data collection and preparation 12
Stage 3: Model design and development 14
Stage 4: Model training, fine-tuning, and testing 17
Stage 5: Model deployment and integration with end-product 19
Stage 6: Model behavior/outcome monitoring and adjustment 21

05 Enter the matrix 23

06 What's next? 24

Intro 00

Many organizations today aim to take advantage of the large-scale promise of AI. But it is imperative that this is done in a responsible, ethical, and safe way. To achieve this, securing AI usage plays a big role in responsible, safe AI use cases. AI and security can largely be organized into three buckets:

01 Securing the use of AI

02 Using AI to secure other systems

03 Countering the attacker armed with AI

We've seen the rapid progress in AI – and specifically generative AI – going from simple chatbots to systems that can answer complex questions and generate new content. We can draw a parallel between its amplification and speed of change. What do we mean with that statement? The rate at which AI systems are being built, trained, and deployed means that bad security practices are amplified, and quickly turn into scenarios where the development process slows and leaves us with less secure states.

In turn, robust security practices and hygiene are fit for purpose to accelerate the building, training, and usage of AI systems in your organization. This greatly underscores the need for security by design and "shifting left."

Google's Secure AI Framework (SAIF) launched in June 2023 and the follow-on SAIF Risk Map and Risk Assessment present Google's year of securing AI systems in a practical framework for organizations to use in securing AI systems and addressing AI risks. Now Google Cloud is building around SAIF principles with AI security products like AI Protection that are designed around securing the data, model, infrastructure, and application.

This paper follows the previous publication, Best Practices for Securely Deploying AI on Google Cloud, and focuses on applying the SAIF principles in securing the use of AI throughout the model lifecycle. Future research papers by CoSAI (an OASIS Open project co-founded by Google) will further explore dimensions of secure supply chain for AI, AI security risk governance, and other AI for defenders aspects.

SAIF and industry frameworks

01

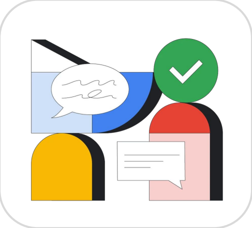
AI security is a rapidly developing field where new frameworks and standards are being created as we speak. Frequently asked questions include: How does SAIF map to other frameworks or standards? How is it different from other frameworks such as NIST AI Risk Management Framework or CISA/NCSC guidelines for secure AI system development?

The answer to this question is to think of SAIF as an umbrella framework. Let’s expand on that. SAIF pillars can map to other frameworks. For example, take the National Cyber Security Center (NCSC) guidelines, which are divided into four key areas. Consider number two, secure development. which states the following:

- **Secure your supply chain**
- **Identify, track, and protect your assets**
- **Document your data, models, and prompts**
- **Manage your technical debt**

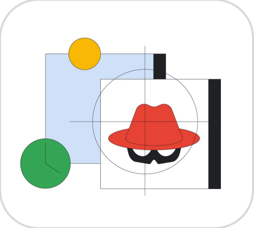
The Secure AI Framework (SAIF) consists of the following six pillars:

01



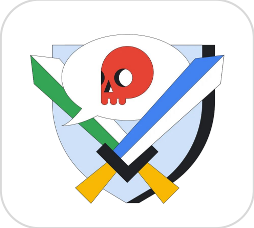
Expand strong security foundations to the AI ecosystem

02



Extend detection and response to bring AI into an organization’s threat universe

03



Automate defenses to keep pace with existing and new threats

04

Harmonize platform level controls to ensure consistent security across the organization

05

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

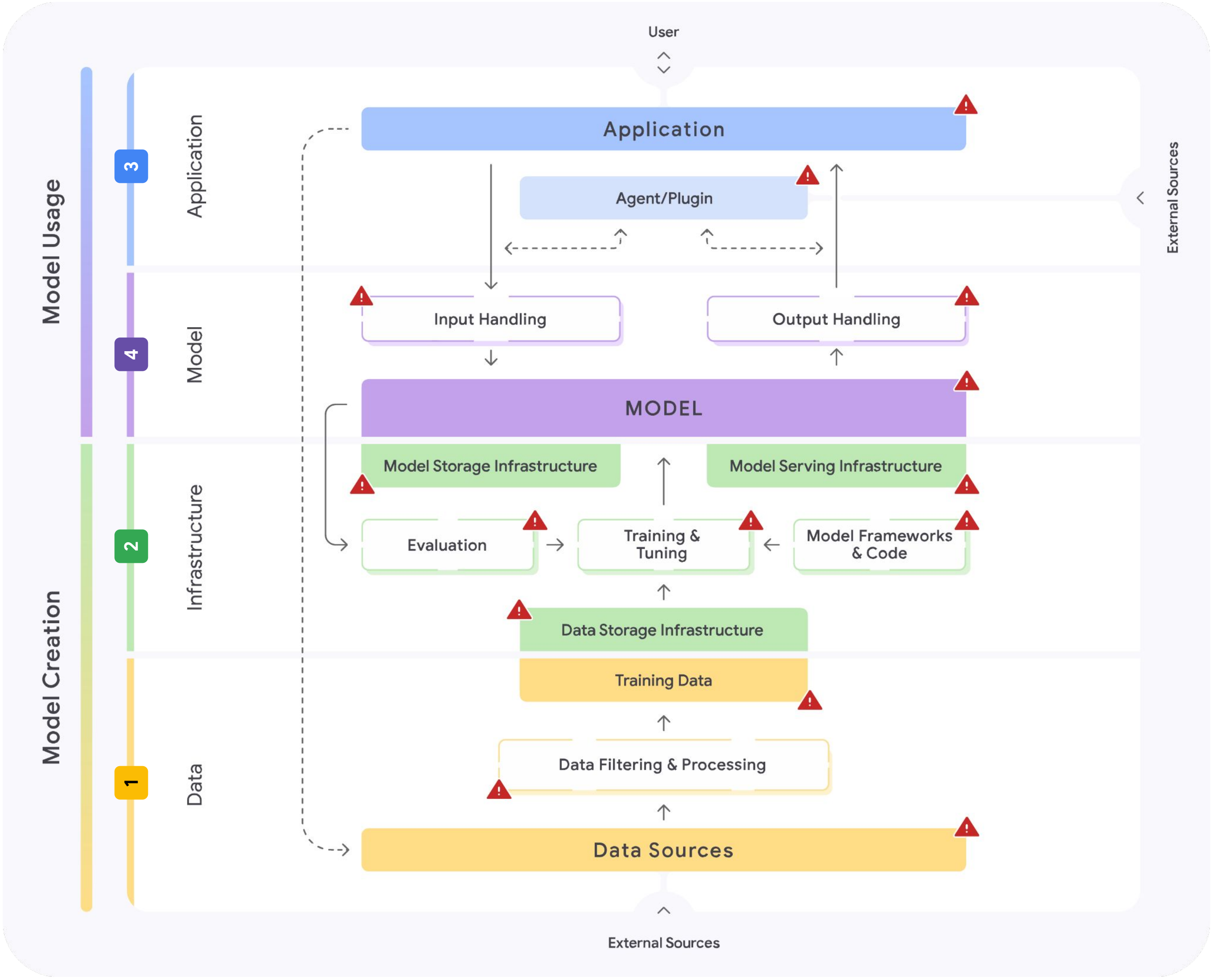
06

Contextualize AI system risks in surrounding business processes

Categories and dimensions

In order for us to apply the SAIF principles or frameworks such as NIST AI Risk Management Framweork, we need to distill the core components of an AI model. These could largely be grouped into four dimensions, giving us another aspect needed to overlay the SAIF pillars.

These four dimensions are: Data, Infrastructure, Application, and Model.



1

Data

At first glance, you might think of data being encapsulated in the Model element. That’s true, as data is one of the most essential components of a model. But there is also a precursive component, in the form of data governance that yields benefits upstream.

For example, is the data you need to build and train your model allowed for that purpose? This is an important first question.

Data governance practices similar to how we look at securing the supply chain are extremely important. For example, starting with knowing where sensitive data is and having a data bill of material along with the lineage of that data ensures that data is used in accordance with the purpose intended, but also maps out the associated risk and provides downstream benefits, including detection baselines on data tampering.

2

Infrastructure

Models are still composed of software needing to ultimately run on infrastructure or be served from infrastructure. This can take the form of AI on devices such as mobile phones, AI models built on your own infrastructure, or AI models provided in the cloud by a cloud service provider. In each of these cases, access to the infrastructure, and the infrastructure from which these models are served or built upon, needs to be secured such as via AI posture controls. You can think of this as a subdimension of the infrastructure. Whichever one of these three scenarios you encounter, or a combination of them, you will need to factor into your overall defense in depth, and secure accordingly.

3

Application

First we need to provide some context for “application”, as applications can mean different things. Here, we are referring to applications in two specific contexts:

- Development of AI systems related to software development and associated practices, such as writing the code for the systems and deploying the code to production
- Presenting the AI system in the form of an application, like an LLM chatbot-style application or an AI agent

Because AI models still need to be deployed into production, they require a software deployment pipeline. They are, in essence, still composed of software/code and will be presented in a form of application to be utilized for its intended purpose. That means application security is vital, all the way from secure coding practices and supply chain security (via SLSA levels) approaches, down to protecting the model when presented as an application.

A significant and growing area of AI application development is the creation of agentic AI. These AI systems are designed to perform actions in the world, either digital or physical, based on their analysis of data and their programmed goals. Agentic applications can take various forms, from simple bots that automate tasks to sophisticated systems that interact with complex environments. The security considerations for agentic AI are broader than for traditional applications.

4

Model

The final component to consider is the model itself, including how you secure the model from inception, all the way to when the model is running and being presented in the form of an application. This also encompasses data governance to make sure that training data is properly protected and not tampered with from the start.

When dimensions intersect

With the above four dimensions/components come cases of intersection. As an example, an LLM model that is ready for production needs to be deployed to infrastructure where this will run, and then needs to be presented to the end user for use. For the sake of this example, we are assuming that secure development was done and is now being securely deployed into infrastructure where it will run – and that protections and safety measures are in place when the model is used by the end user. We also assume that the presentation layer presenting the model in application form to the end user is protected. Now you can see all four previously mentioned dimensions intersecting.

To build or buy (CSP vs. BYOAI) considerations

Before we get deeper into SAIF to real-world considerations, consider another aspect: Are you going to be creating and training your own model? And if so, on your own infrastructure or on the infrastructure of a cloud service provider (CSP)? Or will you use an already provided model from a model provider, like Vertex AI? This consideration is an important one, as it will impact everything from cost, compute, and infrastructure to security. Even in this consideration, SAIF is still applicable. It's the context that matters.

For example, take access control and look at it through the lens of pillar 1 of SAIF, **"**Expand strong security foundations to the AI ecosystem."If you have a mature security process in your organization, chances are you have spent significant time in building Identity Access Management (IAM) access permissions. This can now be expanded and made AI specific. Depending on the use case, you will more than likely need to create some roles and groups related to the personas required to either build, train, or tune your model. However, when you use a model provided by the CSP, they may already have curated roles for you. So instead of you having to create new ones, you can use the preset ones and adapt as needed – versus if you have to create your own from scratch in your own infrastructure if you did not have such a need before. This is one small example of how SAIF can be applied to the "build or buy" scenario. You can also see how that decision can impact seemingly nascent things such as IAM access permissions.

A word about threats

As security practitioners, leaders, and security-minded people in general, we should be creating threat models to help us understand risks. We can then use that understanding to make risk-based decisions, which in turn drives decisions around mitigations, preventions, detections, and controls, and in some cases even acceptance for a particular risk. At a high level, this is, in theory, what we should be doing. Often, however, we get caught up in the threat itself and lose sight of the outcome and context.

As AI security is relatively newer than, say, traditional network security, there are new areas in attack and defense thereof. With "newer" things comes a high chance of fixating on certain threats without understanding them in context of the intended use case. In short, threats should be considered in context across the dimensions, and not just around the newer types of threats, such as prompt injection.

Model lifecycle refresh

There are a few examples of the model lifecycle with various different takes, but in general, the six stages below apply. We will use these phases and overlay SAIF principles across them.

01	Opportunity discovery and problem definition	• Identify the business problem or opportunity that you want to solve with AI. • Evaluate for regulatory relevance, possible bias, etc. • Choose the right data, algorithms, and evaluation metrics.
02	Data collection and preparation	• Collect (or generate) the data that you will need to train your AI model. • Analyze, label, transform, and ingest the data. • Establish end-to-end data governance based on regular risk assessment and threat modeling.
03	Model design and development	• Design and develop the AI model that will address the business opportunity that you have defined. • Build in mechanisms to assess and mitigate potential risks. • Audit model performance and screen output. • Build in the facilities for explainability and human intervention.
04	Model training, fine-tuning, and testing	• Train the AI model on the data that you have collected. • Test the model to make sure that it is performing as expected. • Analyze outcomes to compare with expected results. • Implement security measures throughout the training process.
05	Model deployment and integration with end-product	• Deploy the AI model production. • Make the model available to users so that they can solve the problem that you have defined. • Implement runtime security safeguards.
06	Model behavior/outcome monitoring and adjustment	• Monitor the behavior and outcomes of the AI model to make sure that it is performing as expected. • Understand how users may be using the model to identify signs of badness. • Adjust the model over time to account for the changes in the data or the environment. • Implement output filtering measures.

Just in case it is all about the use case

Understanding the use of the intended AI service or model will help you set the context that is needed to understand the risks overlaid with DIAM. This allows us to start applying SAIF pillars. As a quick example, let's take everyone's favorite threat at the moment: prompt injection. This use case matters as prompt injection is mainly targeted toward large language models (LLMs), which are a form of generative AI. If your use case does not involve an LLM, then the risk of a prompt injection threat scenario is fairly low. To underscore even further why the use case matters, let's assume your use case will involve an LLM.

Let's take three use cases where an LLM can be used. These will not be exhaustively detailed on how these use cases would work, but rather used to illustrate the importance of the use case context.

Use case 1

As a new employee at company Y, to speed up your onboarding process, you interact with the in-house internal-only LLM called Boardy. You might ask Boardy things such as, "Hey Boardy, please summarize in five steps how I would create a git repository in my company's repo?" Boardy then summarizes the internal documentation for you in five steps that you start following.

Use case 2

Company X launches an LLM chatbot assistant, "Speaky," that is external and open to the internet, allowing anyone to interact with the LLM to perform various different tasks.

Prompt injection exists in both use cases. However, as one might imagine, the contextual risk for use case 2 is higher for this threat than in use case 1.

Use case 3

Company Z, a rapidly growing e-commerce business, is struggling to keep up with increasing customer support requests. They launch "SupportBot 3000," an agentic AI designed to handle a wide range of customer service interactions. This is not a simple chatbot providing information. SupportBot 3000 is designed to perform actions on behalf of customers, cancel orders, make replacement purchases, and call APIs to test the product performance.

Prompt injection exists in the agentic case as well, but the risk is even higher due to real-world consequences of these actions.

Callout: Agentic AI refers to a type of AI system that can perform tasks, make decisions, and achieve goals in a complex, dynamic environment with a degree of independence and proactivity. It goes beyond simply responding to commands or analyzing data. It is a rapidly developing area currently, but one can already start to think about risks such as mentioned above in use case 3, where the agent has too much agency over what it has access to and what actions it can take. This should be considered in stage 1 as part of your use case. Future papers will delve further into agentic AI systems and how SAIF plays a role in ensuring the desired outcome.

Where are you with SAIF today?

Here, we are making the assumption that you have already assembled the correct multidisciplinary team as mentioned in our SAIF approach paper, and that everyone has had an AI primer to bring them up to speed.

Stage 1 Stage 2 Stage 3 Stage 4 Stage 5 Stage 6

Opportunity discovery and problem definition

At this first stage, you or your organization is looking at a problem definition or opportunity where an AI model can be beneficial. The most important dimension would be data, and the SAIF pillar that would apply would be pillar 6.

Applicable Dimension: Data

At this stage, it is important to make sure that you have a data governance process in place and that it takes into account AI use cases.

In the context of an AI system, data governance encompasses the practices, processes, and technologies implemented to ensure that the data used in AI systems is secure, private, accurate, available, and usable throughout the entire data lifecycle.

Key aspects of data governance in AI include:

- **Data classification** – Categorizing data based on its sensitivity, value, and potential risks.
- **Data lineage** – Tracking the origin, movement, and transformations of data to understand its context and ensure accountability.
- **Metadata and data catalogs** – Creating comprehensive descriptions of data assets to facilitate discovery, understanding, and management.
- **Data quality** – Ensuring the accuracy, completeness, and consistency of data used for AI model training and decision-making.
- **Data lifecycle management** – Defining and automating policies for data retention, archival, and deletion to comply with regulations and optimize storage resources.
- **Data access management** – Implementing controls to grant appropriate access privileges and monitor data usage to prevent unauthorized access and misuse.
- **Data privacy and security** – Protecting sensitive information through encryption, de-identification, and other security measures to maintain confidentiality and compliance.
- **Data policies and standards** – Establishing clear guidelines and rules for data handling, usage, and sharing to ensure consistency and ethical practices.

Applicable SAIF pillar: 6 - Contextualize AI risk in surrounding business processes

Knowing what the use case will be can help you understand the associated risk. This allows risk-based decision-making and also documenting that risk or risks. This, in turn, will drive our decision-making in regards to risk acceptance, reduction (mitigation), and, in some cases, not pursuing it due to the risk being too great.

To contextualize AI risk in surrounding business processes, you should first establish a model risk-management framework and build a team that understands AI-related risks. Then, evaluate the relevance of traditional controls to AI threats and risks using available frameworks.

Next, perform a risk assessment that considers organizational use of AI, and match the AI use cases to risk tolerances. Finally, review how existing security controls across the security domains apply to AI systems.

Summary

Focus on **data security** and especially **data governance**. During this stage it is crucial to establish or update your data governance process with AI in mind. This includes aspects like data classification, data lineage, data quality, and data policies and standards. AI-focused posture analysis tools help with the infrastructure side of securing AI.

Additionally, it is essential to establish a model risk-management framework and assemble a team who understands AI-related risks. This involves evaluating the relevance of traditional controls to AI threats and risks using available frameworks, performing a risk assessment that considers organizational use of AI, matching AI use cases to risk tolerances, and reviewing existing security controls across security domains to determine their applicability to AI systems.

Helpful questions

- What type of data will the model need?
- Can you use the type of data you want to use? Do you have permission to use the data for this purpose?
- What will the model do? That is, what is the use case?
- Will it be internal or external facing?
- Will we use an existing model/model from a service provider, or will we create one from scratch?
- Where will the model be hosted? On prem, CSP, or as a provider SAAS offering?

Data collection and preparation

We have now established the use case and contextualized the risk for our organization. You can now start to collect data and prepare for use in training the intended model to support the use case. As such, data will play a key role, as will the infrastructure where the data will be hosted and stored. The picture becomes more clear as the SAIF pillars start to apply practically across the DIAM dimensions.

Applicable Dimensions: Data, Infrastructure

Data dimension

- **Maintain provenance of data, secure training data supply chain** – Having a data "bill of materials" and inventory will simplify the task of creating provenance.
- **Ensure data tagging and classification** – This is also a good time to validate whether your classification policy is still consistent with new use cases.
- **Various data access control measures** – These are dependent on the tools of your business but also the types of data, including access control lists, encryption, endpoint controls, and in some cases, limiting physical access to data stored on servers.

Infrastructure dimension

- **Refine and define access control policies** for data repositories.
- **Apply security controls to the entire data pipeline** including data collection, cleaning, and retention. These would include things such as encryption for data at rest and in transit, but also details such as role-based access control to data sets, based on your adjusted policies.
- **Prepare to store and track** supply chain assets, code, and training data.
- **Create data access violations and anomalies detection** based on your intended usage to support the use case. Detections would include unapproved access to data, but also things such as unexpected altering or mislabeling of data.

Applicable SAIF pillars: 1, 2, 4

At the second phase of the lifecycle, these SAIF pillars play a prominent role.

- Expand strong security foundations to the AI ecosystem (SAIF1)
- Extend detection and response to bring AI into an organization's threat universe (SAIF2)
- Harmonize platform level controls to ensure consistent security across the organization (SAIF4)

Summary

Data dimension

- Maintaining provenance of data and securing the training data supply chain.
- Ensuring data tagging and classification, and validating that the classification policy is consistent with new use cases.
- Implementing various data access control measures, including access control lists (ACLs), encryption, and endpoint controls.

Infrastructure dimension

- Refining and defining access control policies for data repositories.
- Applying security controls to the entire data pipeline, including data collection, cleaning, and retention. This includes encryption for data at rest and in transit, and role-based access control (RBAC) to data sets.
- Preparing to store and track supply chain assets, code, and training data.
- Creating data access violation and anomaly detection based on intended usage and the use case.

Note: If you are using multi-cloud or hybrid cloud by combining on prem with the cloud or multiple projects in a GCP organization, ensure you establish control outcomes in order to make sure you are able to harmonize platform level controls to ensure consistent security across the organization (SAIF4).

Helpful questions

- What are the organization’s data collection and storage capabilities?
- How will the organization ensure the quality and integrity (and confidentiality, where necessary) of the data collected?
- How will the data be anonymized or de-identified to protect privacy?
- How will the organization ensure compliance with relevant data protection regulations?
- What security measures will be implemented to protect the data from unauthorized access or tampering?
- How will the organization track and document the data preparation process? What is the plan for data versioning and management?
- What list of existing security controls can be adjusted to be utilized as-is to start adding protections to secure the infrastructure dimension?

Model design and development

At this third stage, the model is being developed based on the training data and intended use case. Three dimensions feature more prominently: Application, Model, and Infrastructure. A big part of this stage is the actual software development lifecycle. Secure software development practices, along with supply chain security, are two of the larger components you need to think about.

From a model perspective, you need to think of initial safety measures and mitigations for risks such as abuse of the model, and lastly you need to apply appropriate security measures to the infrastructure where the development and training of the model will take place.

Applicable Dimensions: Application, Model, Infrastructure

Application dimension

- **Frameworks are essential:** Make sure you choose the appropriate machine learning framework, such as TensorFlow, PyTorch, or JAX. These frameworks provide pre-built components and abstractions that simplify the development process. They handle complex tasks like automatic differentiation, GPU acceleration, and model deployment – but they are still software and, as such, may at times have bugs, so you need to consider vulnerability management.
- **Software supply chain security:** As you will be using machine learning frameworks that are, in a lot of cases, open source software, you need to consider supply chain attack risk as well.
- **Focus on data pipelines:** A significant portion of the code deals with data preprocessing and pipeline management. This involves loading data, transforming it, and feeding it to the model in an efficient manner. Frameworks provide utilities for building these pipelines. Make sure that the pipelines are adequately protected and access is limited to where needed, and that this is consistent with what you defined for stage

Model dimension

- **Treat the model as code:** Use all the relevant software supply chain security safeguards. As with previous points, think about how vulnerability management impacts the model, and sanitize against potential malicious or unwanted input.
- **Output handling:** Output handling would have similar components to input handling that would protect against unwanted, unexpected, or dangerous output.
- **Audit:** Audit the performance of your input and output handling controls to ensure they are fit for purpose and aligned with your use case requirements.

Infrastructure dimension

- Apply security controls to the systems where the model is being trained.
- Limit access to raw models representations and weights.
- Ensure detections are in place to look for anomalies regarding data access, as well as cases of potential data poisoning which can be validated against your data bill of materials established earlier.
- Ensure the infrastructure housing the data is not accidentally exposed. For example, you would not want a storage bucket containing your private training data to be made publicly available.

Applicable SAIF pillars: 1, 2, 3, 4

- Expand strong security foundations to the AI ecosystem (SAIF1)
- Extend detection and response to bring AI into an organization's threat universe (SAIF2)
- Automate defenses to keep pace with existing and new threats (SAIF3)
- Harmonize platform level controls to ensure consistent security across the organization (SAIF4)

Summary

Application dimension

- For the **Application** dimension, it is essential to select the appropriate machine learning framework and consider vulnerability management and software supply chain security. It is also important to focus on data pipelines and ensure they are adequately protected.

Model dimension

- For the **Model** dimension, the model should be treated as code, with all the relevant software supply chain security safeguards. It is also important to start building controls for input and output components and audit the performance of input and output handling controls.

Infrastructure dimension

- For the **Infrastructure** dimension, security controls should be applied to the systems where the model is being trained. Access to raw models should be limited, and detections should be in place to look for anomalies regarding data access and potential data poisoning.

? Helpful questions

- What kind of access control will be used to restrict access to the training environment?
- What security measures will be implemented to protect the model from tampering or theft at the development stage?
- Can you identify where and when sensitive data is used?
- How will the organization ensure that the model is used in a safe and responsible manner?
- What code development pipeline will be used. and are you using public repositories or private repositories?
- Do you have an inventory of software libraries used in your development process?
- How will you deal with vulnerabilities in your chosen ML framework?

Model training, fine-tuning, and testing

In the fourth stage, we are now in a critical point of the model which is the training, fine-tuning, and the validation of the model through testing. Depending on the model size, this can be an expensive and resource-intensive undertaking.

Applicable Dimensions: Model, Data

Model dimension

- Treat models as software code and extend existing software supply chain integrity frameworks to cover AI/ML. This includes securing and tracking the provenance of the emerging model, its training data, and any code used in its development.
- Continuously monitor and evaluate the model's performance during training and fine-tuning to identify any anomalies or unexpected behaviors.
- Subject the model to various adversarial attacks and stress tests to evaluate its robustness against potential exploits and manipulations.
- Maintain detailed records of any changes made to the model during fine-tuning, including the rationale for the changes, the impact on model performance, and any potential security implications.

Data dimension

- Implement strict access controls to protect the training data and any additional datasets used for fine-tuning. This includes ensuring that only authorized personnel have access to the data and that it is protected from unauthorized modification or exfiltration.

Applicable SAIF pillars: 1, 2, 5

- Expand strong security foundations to the AI ecosystem (SAIF1)
- Extend detection and response to bring AI into an organization's threat universe (SAIF2)
- Adapt controls to adjust mitigations and create faster feedback loops for AI deployment (SAIF5)

Summary

Model dimension

- For the Model dimension, it is crucial to treat models as software code and extend existing software supply chain integrity frameworks to cover AI/ML. This includes securing and tracking the provenance of the emerging model, its training data, and any code used in its development. It is also important to continuously monitor and evaluate the model's performance during training and fine-tuning to identify any anomalies or unexpected behaviors, subject the model to various adversarial attacks and stress tests to evaluate its robustness, and maintain detailed records of any changes made to the model.

Data dimension

- For the Data dimension, it is vital to implement strict access controls to protect the training data and any additional datasets used for fine-tuning, ensuring that only authorized personnel have access to the data and that it is protected from unauthorized modification or exfiltration.

Helpful questions

- What adversarial attacks or stress tests will be used to evaluate the model's robustness?
- How will the model's operation be monitored and evaluated during training and fine-tuning?
- What data security measures will be implemented to protect the training data during this stage?
- How will the organization ensure the privacy and security of any sensitive data used for fine-tuning?
- What processes will be in place to track and document changes made to the model during fine-tuning?

Model deployment and integration with end-product

Stage 5 is the integration of the model into the intended end application after the training and fine-tuning have been completed. This stage also introduces potential risks through insecure integration so you will also need to consider if there are unique requirements specific to your AI model's integration into the end product. For example, does your integration have the potential to leak data? Does it have unrestricted access to a model along with your more traditional application security considerations?

Applicable Dimensions: Application, Infrastructure

Application dimension

- For the Application dimension, Secure Coding Practices come to the forefront. Implement traditional secure coding practices to minimize vulnerabilities in the application code that interacts with the AI model. At the same time, secure APIs to integrate the AI model with the end product, including proper authentication, input validation, and rate limiting.

Infrastructure dimension

- For the Infrastructure dimension, there is no magic. Enhanced Traditional Security plays a key role.
- Organizations need to secure the network with access control and treat the AI workload system as a valuable asset, enabling secure boot, minimizing services, and applying security updates consistently.
- Similarly, Strict Access Control is still central to this, especially for that model. Only authorized personnel with strong authentication (MFA is a must) should be granted access. If APIs are implemented, implement robust security measures such as keys, tokens, and rate limits to restrict unauthorized access.
- Detection and response is, unsurprisingly, also key. Constant monitoring is essential. Implement extensive logging mechanisms to track model usage, performance, and any anomalies. Utilize a security information and event management (SIEM) tool to assist in monitoring efforts. Configure anomaly detection systems to identify and address potential threats promptly.

Applicable SAIF pillars: 3, 4, 5, 6

- Automate defenses to keep pace with existing and new threats (SAIF3)
- Harmonize platform level controls to ensure consistent security across the organization (SAIF4)
- Adapt controls to adjust mitigations and create faster feedback loops for AI deployment (SAIF5)

Summary

Application dimension

- For the Application dimension, it is crucial to implement secure coding practices to minimize vulnerabilities in the application code which interact with the AI model and secure APIs used to integrate the AI model with the end product, including proper authentication, input validation, and rate limiting.

Infrastructure dimension

- For the Infrastructure dimension, it is essential to enhance traditional security measures by securing the network with access control and treating the AI workload system as a valuable asset. It is also critical to implement strict access control to the model, continuously monitor model usage and performance, and use anomaly detection systems to promptly identify and address potential threats.

Helpful questions

- How will the organization ensure the AI model is deployed securely into the production environment?
- How will the integration of the AI model with the end product be secured, especially regarding data flow and API security?
- How will the organization ensure the AI model's continued integrity and security throughout its lifecycle in the production environment?
- What measures will be taken to ensure the AI model's compatibility and secure integration with existing systems and infrastructure?

Model behavior/outcome monitoring and adjustment

Applicable Dimensions: Model, Application

Model dimension

- **Model monitoring and drift detection:** This will be announced at a later time.

Application dimension

- **Input validation and sanitization:** Garbage in, garbage out, and possibly a security hole. Never trust user input! Validate rigorously, sanitize meticulously, and be on the lookout for those sneaky injection attacks trying to poison your model.
- **Output monitoring:** Keep a close eye on what your model is spitting out. Is it within expected bounds? Are there any biases creeping in? Unexpected outputs could be a sign of trouble – either a flaw in the model or an attack in progress.
- **Auditing and logging:** Who did what, when, and to what effect? Log all access to the model, all changes made, and all significant decisions. This is not just for security, but also for debugging and understanding model behavior. A good audit trail is a security professional's best friend. Then use the logs to train detectors to spot deviations from this norm. Sudden spikes in activity, weird input patterns, or unexpected outputs could be red flags. Remember, attackers love to exploit the unexpected.
- **Log retention policies:** Define clear log retention policies that balance security needs with data minimization principles, and sanitize private data. Retain logs for as long as necessary for security analysis and incident response, but delete them when they are no longer needed.
- **Regular security testing:** The world changes, and attackers become smarter. Regularly test the application for vulnerabilities. Penetration testing, vulnerability scanning, and red teaming are your allies in this fight.

Applicable SAIF pillars: 2, 3, 5, 6, 4

- Extend detection and response to bring AI into an organization's threat universe (SAIF2)
- Automate defenses to keep pace with existing and new threats (SAIF3)
- Adapt controls to adjust mitigations and create faster feedback loops for AI deployment (SAIF5)
- *MAYBE* Harmonize platform level controls to ensure consistent security across the organization (SAIF4)
- Contextualize AI system risks in surrounding business processes (SAIF6)

Summary

Model dimension

- Stage 6 is the **Model behavior/outcome monitoring and adjustment phase**. This stage involves monitoring the model's behavior and outcomes in the production environment and making necessary adjustments to maintain its performance, security, and fairness.

- Key aspects for the **Model** dimension include monitoring the model and detecting drift to ensure it remains effective and secure.

Application dimension

- Key aspects for the **Application** dimension include input validation and sanitization, output monitoring, auditing and logging, and regular security testing.

Helpful questions

- What steps will be taken to protect the model from unauthorized access, tampering, or data exfiltration during deployment?
- What monitoring and logging mechanisms will be implemented to track the model's performance and detect potential security issues in the production environment?
- What processes will be in place to manage and respond to security incidents or vulnerabilities discovered after deployment?

Enter the matrix

04

Starting here, we are making the assumption that you have already started to follow the three pre-SAIF steps we laid out in the previous SAIF AI Framework Approach publication. Here is a quick recap of the four steps:

Step 1
Understand the use

Step 2
Assemble the team

Step 3
Level set with an AI primer

Step 4
Apply the six core elements of SAIF

**Note: Not counting step 4 as pre-step as step 4 is to apply SAIF.*

In order for you to start applying SAIF to secure your AI models, a matrix-style approach illustrates how this can be applied. For some, the imagery of scrolling green code and dodging bullets will appear in your minds when the word "matrix" is mentioned. Even though at times it may feel like that is the case, this means applying a matrix-type grid approach. Here we will take the six SAIF elements overlaid onto the six model lifecycle steps, use the DIAM four dimensions to break it down into components, and throw in some threats for good measure. We hope that once we illustrate further down, it will become clear how these things intertwine.

SAIF **down**, lifecycle **right**	Opportunity discovery and problem definition	Data collection and preparation	Model design and development	Model training, fine-tuning, and testing	Model deployment and integration with end-product	Model behavior/outcome monitoring and adjustment
1 Expand strong security foundations to the AI ecosystem (SAIF1)	○	○	○	○	○	○
2 Extend detection and response to bring AI into an organization's threat universe (SAIF2)		○	○	○	○	○
3 Automate defenses to keep pace with existing and new threats (SAIF3)		○	○		○	○
4 Harmonize platform level controls to ensure consistent security across the organization (SAIF4)		○	○	○	○	○
5 Adapt controls to adjust mitigations and create faster feedback loops for AI deployment (SAIF5)		○		○	○	○
6 Contextualize AI system risks in surrounding business processes (SAIF6)	○	○				

What's next?

As you continue to develop your AI use cases, review and implement the following recommendations:

- Understanding the use case of an AI service or model is critical to securing it. What is the purpose of the AI system? What data will it be used on? How will it be used?
- Consider three main dimensions when securing AI: Data infrastructure, application, and model. All three need to be covered by the controls to have a secure AI system.
- Threat modeling can help security practitioners understand the risks associated with AI and make informed decisions about how to mitigate those risks. Such a threat model needs to be broad, and not just focus on a headline threat spotted this week.
- The SAIF framework can be used to help organizations implement secure AI practices.

Additional action items that your organization can take to secure your AI deployments include:

- Creating a risk management plan for AI
- Implementing security controls for AI systems
- Monitoring AI systems for security vulnerabilities and threats
- Continually evaluating and improving their AI security posture
- Expanding the community role in security AI with CoSAI