Google's secure AI framework

Previously, you learned that a **risk management framework** is a set of practices, processes, and technologies that help enable an organization to identify, assess, analyze, and manage risk within an organization. As artificial intelligence (AI) becomes more widely adopted, it's important for organizations to establish a risk management framework that addresses AI's complexities.

In this reading, you'll learn about Google's Secure AI Framework (SAIF), and explore examples of how this framework can be used as a tool to help secure AI technologies and processes.

Google's Secure AI Framework (SAIF)

The SAIF provides organizations with considerations to make when implementing AI systems and identifying how to mitigate risks. The SAIF offers six core elements:

1. Expand strong security foundations to the AI ecosystem
2. Extend detection and response to bring AI into an organization's threat universe
3. Automate defenses to keep pace with existing and new threats
4. Harmonize platform level controls to ensure consistent security across the organization
5. Adapt controls to adjust mitigations and create faster feedback loops for AI deployment
6. Contextualize AI system risks in surrounding business processes

SAIF core elements

The SAIF is a set of AI-focused best practices that security professionals can implement throughout the development lifecycles, from infrastructure to data. Organizations should tailor these best practices to meet their own unique business needs. Also, organizations should form teams with a variety of stakeholders to guide AI implementation, including security analysts, cloud engineers, developers, and responsible AI, ethics, and legal team members.

Expand strong security foundations to the AI ecosystem

Organizations should assess the existing security controls in place that secure their underlying infrastructure, data, and applications. Many, if not all of these controls also apply to securing AI systems, but some may need to be reconfigured to address the AI system's specific requirements. For example, data encryption might traditionally be used for authorization purposes, but when considering AI, organizations will also need encryption to help prevent data tampering or theft.

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

Organizations must be able to detect and respond to AI-related cyber threats and attacks. One way organizations accomplish this is by using threat intelligence. Threat intelligence is evidence-based threat information that provides context about existing or emerging threats. Imagine a system that's always up to date with how a virus, attack, or deception might compromise it. Then, that system can actively alert and initiate a response from analysts, all by using the power of AI-assisted threat intelligence.

To prepare for threats or cyber attacks, organizations should be familiar with training data used for AI systems, and the different components of the system. This means the organization should know how current their system is for training the data, and how its models work with other security toolsets. With this information, organizations can help identify vulnerabilities within their AI systems, which contributes to their threat intelligence, and helps inform a response to potential threats.

It's also important to monitor the input and output created by generative AI (GenAI) systems. For example, because of GenAI's ability to create a vast amount of content, the output should be closely monitored for negative or harmful output that can be generated for malicious purposes. Content safety policies can help prevent this risk of harmful content. This content might appear as GenAI-created malicious emails or misleading output from a company's AI-powered chatbot.

Automate defenses to keep pace with existing and new threats

Although AI technology may introduce new threats in the cyber space, AI can also be used as a tool to defend against those threats. AI innovates recognizing new patterns and attacks sooner than current toolsets. AI can help detect bias, security breaches, and malicious content. But, human input is necessary for larger and more important decision-making factors, like determining what constitutes a breach, and the appropriate response.

Harmonize platform level controls to ensure consistent security across the organization

AI should be regularly monitored to ensure the AI system, including training data and any related applications, are performing as expected. A comprehensive understanding of each component in the AI system's lifecycle provides organizations with a clear picture of the assets involved. With this information, the organization can identify any overlapping areas where controls or frameworks are duplicating efforts. Then, organizations can consolidate frameworks and controls, helping them save costs and become more efficient.

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

Threat actors execute new cyberattacks every day, and AI opens a new environment for these actors to exploit. To help combat evolving threats, organizations should constantly test their AI implementations to improve the security of the changing threat landscape. One way to test the durability of an AI system is to perform red team exercises. A **red team** is a group of ethical hackers who mimic potential adversaries in order to examine the security defenses of an organization. Red teams can be a team within an organization, or they can be hired as a third party.

Organizations, and especially security analysts, should stay up to date on the latest attack methods and techniques that threat actors use. Common AI vulnerabilities might include leaking data or forming wrong predictions. Organizations can use red teams or other testing measures to help identify attack vectors that could exploit these vulnerabilities.

Communication is a critical aspect for AI deployment. Any vulnerability should be shared with remediation teams, and also with the organization at large. With constant testing in the AI environment, there should also be constant feedback to stakeholders and to the training data used for the AI model.

Contextualize AI system risks in surrounding business processes

Organizations should create processes for managing risks related to AI systems. Risk assessments outline how the organization, third parties, and security controls interact with the AI model's components. The risk assessments should encompass the entire AI lifecycle, including training data and applications that use that data. For example, organizations can incorporate automated security checks as a control to monitor AI performance both internally, and with third-party applications. Understanding how and where AI is implemented can help analysts select the right risks to address first, from inadequately tuned chatbots, to infrastructure as code repeating AI-generated design mistakes.

Key takeaways

Risk management frameworks are an important tool to help reduce risks and improve security posture. With the emergence of AI technology, organizations need to integrate frameworks that address risks AI can impose. Google's SAIF is a tool that can help organizations navigate implementing AI and the many considerations involved in the process.

Resources for more information

For more information about securing AI frameworks, check out:

- Google's quick guide to implementing the SAIF