Google

Google’s Secure AI Framework

AI is advancing rapidly, and it’s important that effective risk management strategies evolve along with it. To help achieve this evolution, we’re introducing the Secure AI Framework (SAIF), a conceptual framework for secure AI systems. SAIF has six core elements:

1. Expand strong security foundations to the AI ecosystem

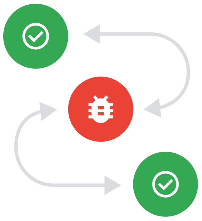
Leverage secure-by-default infrastructure protections and expertise built over the last two decades to protect AI systems, applications and users. At the same time, develop organisational expertise to keep pace with advances in AI and start ot scale and adapt infrastructure protections in the context of AI and evolving threat models. For example, injection techniques like SQL injection have existed for some time, and organisations can adapt mitigations, such as input sanitisation and limiting, to help better defend against prompt injection style attacks.

2. Extend detection and response to bring AI into an organisation’s threat universe

Detect and respond to AI-related cyber incidents in time by extending threat intelligence and other capabilities. For organisations, this includes monitoring input and output of generative AI systems to detect anomalies, and using threat intelligence to anticipate attacks. This effort typically requires collaboration with trust and safety, threat intelligence and counter abuse teams.

3. Automate defenses to keep pace with existing and new threats

Harness the latest AI innovations to improve the scale and speed of response efforts to security incidents, Adversaries will likely use AI to scale their impact, so it is important to use AI and its current and emerging capabilities to stay nimble and cost effective in protecting against them.

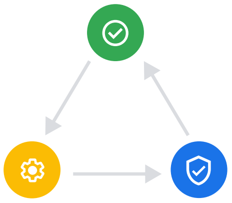


4. Harmonise platform level controls to ensure consistent security across the organisation

Align control frameworks to support AI risk mitigation and scale protections across different platforms and tools to ensure that the best protections are available to all AI applications in a scalable and cost-efficient manner. At Google, this includes extending secure-by-default protections to AI platforms like Vertex AI and Security AI Workbench, and building controls and protections into the software development lifecycle. Capabilities that address general use cases, like Perspective API, can help the entire organisation benefit from state of the art protections.

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

Constantly test implementations through continuous learning and evolve detection and protections to address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback, and involves steps such as updating training data sets, fine-tuning models to respond strategically to attacks, and allowing the software that is used to build models to embed further security in context (e.g. detecting anomalous behaviour). Organisations can also conduct regular Red Team exercises to improve safety assurance for AI-powered products and capabilities.



6. Contextualise AI system risks in surrounding business processes

Conduct end-to-end risk assessments related to how organisations will deploy AI. This includes an assessment of the end-to-end business risk, such as data lineage, validation and operational behaviour monitoring for certain types of applications. In addition, organisations should construct automated checks to validate AI performance.

Safer with Google