SAP Global Security & Cloud Compliance White Paper

SAP Generative AI
Cybersecurity Strategy

Table of contents

Introduction

SAP anticipates that Generative AI (GenAI) will become one of the most disruptive technologies of this decade. It is increasingly transforming the way people collaborate, communicate, and perceive the world. GenAI offers tremendous potential to improve the way the world conducts business, but it also introduces new and complex security challenges. SAP sees secure and responsible use of GenAI as a market differentiator.

Our strategy for GenAI cybersecurity builds upon depth-in-defense principles extending our ability to identify, protect, detect, respond, and recover to the following major GenAI risk categories:

GenAI adoption risks
The cyber-attack surface of an organization increases with the technical complexity of automation. GenAI technologies are so immensely complex that understanding their operation is an active field of scientific research. Our responsible adoption considers GenAI as an unknown hostile element within the business environment.

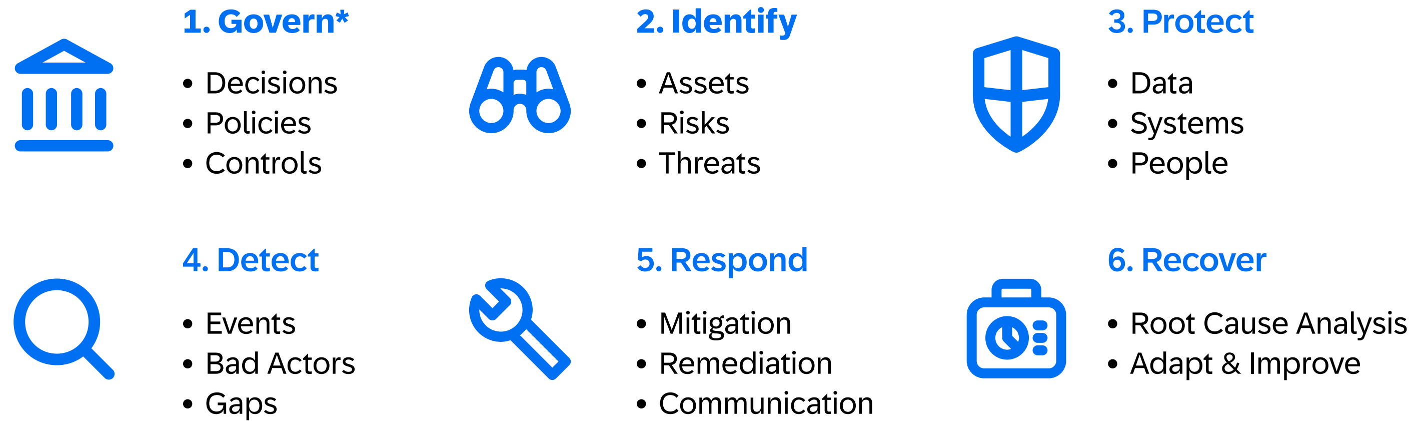
GenAI vulnerability risks
GenAI attempts to replicate human cognitive abilities, but in doing so it also replicates human cognitive weaknesses. Those weaknesses, combined with technical weaknesses, make GenAI an attractive new target for adversaries. Our strategy extends traditional cybersecurity defenses to consider both technical and cognitive weakness of GenAI.

GenAI weaponization risks
Any tool can become a weapon. GenAI is no exception and has potential to become an offensive tool used by adversaries. While the maturity of current GenAI technologies may not yet allow for practical attacks at scale, our strategy anticipates that GenAI technology will continue to improve, making GenAI-assisted attacks a likely threat in the future.

In addition to risk mitigation, SAP sees an opportunity to use GenAI-assisted security tooling and operations. By leveraging GenAI responsibly and appropriately, we believe that GenAI can enable new security automation capabilities, reduce risk, and help to improve the overall security posture of SAP and our customers.

GenAI Security Methodology

The National Institute of Standards and Technology Cybersecurity Framework (NIST CSF), as adopted by SAP, provides a suitable methodology with which to assess GenAI in the business context. By using a common and industry-wide taxonomy, we can express the risks and remediations of new technologies like GenAI in terms familiar to our employees and customers.



1. Govern*
- Decisions
- Policies
- Controls

2. Identify
- Assets
- Risks
- Threats

3. Protect
- Data
- Systems
- People

4. Detect
- Events
- Bad Actors
- Gaps

5. Respond
- Mitigation
- Remediation
- Communication

6. Recover
- Root Cause Analysis
- Adapt & Improve

Figure 1: NIST CSF's V2.0 six functions (adapted and abbreviated for GenAI)

These are some of our considerations in applying the NIST CSF's six functions to GenAI:

1. Govern*

The potential for vulnerabilities, human error, and misuse are important factors to consider in software development. SAP's Software Development and Operations Lifecycle (SDOL) mitigates these risks by applying accountable governance to key decisions with a clear assignment of responsibility. Combined with an AI ethics review process based on our guiding principles for AI ethics, our existing secure development disciplines have well-prepared SAP for a secure and responsible GenAI-enabled future.

2. Identify

GenAI foundational model training requires input of a vast amount of data. Fine-tuning a model for application-specific workloads requires input of domain-specific data. For these reasons, data and the data handling processes require careful protection measures, especially given the unique and inherit risks of GenAI.

3. Protect

Protection of assets against identified GenAI threats begins with determining and prioritizing defensive goals designed to mitigate risks. These goals are codified into policies and standards, enforced via technical and process controls, and taught as standard practice to SAP employees and suppliers. Any new and disruptive technology, such as GenAI, receives special emphasis on asset protection.

* Included in anticipation of NIST CSF v2.0 adoption

4. Detect

Proactive monitoring to detect violation of policy or intrusion of defenses is essential. GenAI requires new tools, processes, and techniques to perform important detection functions. For example, combining analysis of system and process events with GenAI “explainability”, can aid in detecting gaps in technical and process controls.

5. Respond

The unique nature of GenAI often requires unique threat response. For example, with the open-ended prompting often used with GenAI, injection flaws are more challenging to mitigate compared to traditional software inputs. Regardless of the threat, clear communication to all involved parties is essential to successful mitigation.

6. Recover

The threat landscape for GenAI technologies is constantly changing. When something unexpected occurs, we take proactive steps to analyze the cause and take measures to adapt and improve our defenses. These measures are recorded and monitored until fully adopted in our policies, standards, controls, and processes.

The NIST Cybersecurity Framework is an effective tool that relies upon people to make it successful. Not covered in the six functions is the importance of a strong security culture. SAP successfully navigates new security challenges thanks to internal security communities and working groups composed of enthusiastic security-minded individuals who wish to make a better future. Fostering healthy security culture is an important part of our methodology.

The High Stakes of GenAI Security

Data is one of the largest incentivized targets for adversaries. To an adversary data may have intrinsic value, or the value may be in the opportunity to manipulate data. Strong access control and monitoring are proven, effective defenses for traditional software. But in the case of GenAI, these approaches alone are insufficient given the sometimes-unpredictable behavior of GenAI models.

As the first step in data protection, SAP has defined strong data classification standards, and labeling requirements for data processing. These help us meet our regulatory and contractual obligations, and enable automation of security process and controls. While the specific types of data used with GenAI will vary by use case, some types of data deserve special consideration:

Customer data

Whether PII (Personal Identifiable Information), PI (Personal Information), or other customer-provided or owned data, protecting customer data is one of SAP's highest priorities. Given the potential for GenAI to "leak" data used during training, SAP follows strict data usage policies and GenAI tenancy controls.

Business data

Maintaining the confidentiality, integrity, and availability of internal SAP data and systems is important to SAP's business and to our customers who trust and depend upon us. GenAI technologies fall under SAP's existing enterprise data policies and controls that govern sensitive data such as contracts, finance, sales, and employee PII.

Intellectual property

Application source code, copyrighted media assets, and product specifications are all examples of traditional intellectual property (IP). SAP is extending the umbrella of IP protection and controls to mitigate model theft techniques unique to GenAI. Special consideration is also given to GenAI that may generate copyrighted works.

In addition to identifying the data assets that require protection, the business processes involved in data handling must also be hardened against GenAI risks and threats. This includes the data and software supply-chains, as well as the human elements involved. SAP has formal documentation and audit procedures in place for all key business processes.

GenAI Adoption Risks

Disruptive technologies are known to impact the security of existing environments in unknown ways. For this reason, SAP strives to balance of technological advancement with appropriate caution. We have identified three major adoption risk categories by which we base this strategy:

Human Risks
- Misplaced trust
- Irresponsible use

Technical Risks
- Forensic gaps
- Indeterminism

Exfiltration Risks
- Data mishandling
- Insider threats

Figure 2: The three major GenAI adoption risk categories

These risks are not exclusive to GenAI, but as with any new technology, it takes time for knowledge, skills, and defenses to mature. By identifying adoption risks early, SAP can better target mitigative efforts.

Human risks

For any give information system, the greatest potential for risk arises from simple human error. With GenAI, human error can extend to misplaced trust in generated output. Adversaries can take advantage of this trust whether GenAI outputs insecure code, AI hallucinations, or simply bad advice. SAP mitigates these human risk factors with comprehensive education combined with strict enforcement of security controls.

Technical risks

There are two major technical risks intrinsic to GenAI that affect secure adoption: lack of explainability, and lack of determinism. If these risks are left unmitigated, they will severely impede forensic analysis and security testing efforts. By implementing GenAI-specific logging standards, coupled with explainability-first GenAI design patterns, SAP can better recreate the conditions of generated output and understand the behavior.

Exfiltration risks

The development and fine-tuning of GenAI-enabled solutions require access to potentially sensitive data. Without strict data classification and access controls, data may accidentally become embedded within a GenAI model during training or development. As a result, GenAI training data is at risk of exfiltration by an adversary attacking the model or by an insider threat who has access to the training data. SAP's mature data handling practices gives us a foundation to mitigate such risks.

These examples are only a small set of the possible risks inherit to GenAI adoption. As a result, specific use cases may incur additional adoption risks besides the general risks described above. SAP has formed an AI Security Taskforce, involving AI leaders from across the company, to identify and mitigate situational adoption risks.

GenAI Vulnerability Risks

GenAI also brings many new vulnerabilities to the application environment. Some of these are different forms of traditional vulnerabilities, some are unique to GenAI, and still others are yet to be discovered. SAP takes an active approach to identifying vulnerabilities in new technologies. SAP Security Research and the global security community have identified several GenAI risks and vulnerabilities.

Below are the top three GenAI vulnerability risks related to Large Language Models (LLM) that have the greatest potential for damage:

Prompt injection

LLMs are a type of GenAI that processes natural language by predicting the completion of an initial prompt. Prompts are often complex and engineered to provide contextual knowledge and instructions. Therefore, applications often use a prompt template replaced with user-supplied input to guide the LLM response for the intended use-case. This opens LLMs to the risk of an adversary sufficiently manipulating the context and therefore behavior of the LLM. These attacks can be conducted directly or by placing adversarial prompts in web page content scraped by an autonomous agent.

Glitch tokens and adversarial examples

LLMs operate on tokens. Tokens are combinations of letters that represent words or parts of words. LLM behavior becomes unpredictable when presented with tokens or sequences of tokens that it rarely encountered during training. The effect can range from benign, nonsensical output, to potentially serious, such as revealing of training data. Adversaries can also use this and other weaknesses to engineer a more complex attack and trigger the LLM to exhibit behavior of their choosing – an advanced type of prompt injection.

Hallucinations and package squatting

As LLMs are predictive language models, their predictions can sometimes be incorrect. These are known generally as “AI Hallucinations” and the damage can range from flawed decision making by a human who is trusting the hallucinations as truthful, to an adversary using this weakness to their advantage. An example of the latter is “Package Squatting,” where an adversary anticipates hallucinated code that imports a fictional package name, registers that package name with a malicious payload, and a developer using an LLM to generate code trusting the hallucinated code, thus exposing their environment and application to the threat.

GenAI vulnerabilities are challenging to mitigate because of the open-ended nature of GenAI input. This presents a dilemma between constraining input for security, but at the cost of reduced capability. SAP manages this balance with responsible AI use guidelines combined with strong product security standards and threat modeling practices.

GenAI Weaponization Risks

As GenAI becomes more commoditized, it can and will fall into the hands of bad actors. SAP Security Research has determined that the practical threat of GenAI-assisted adversaries is currently uncertain. Given existing illegal marketplaces for malware, credentials, and for-hire services, many adversaries may choose not to directly utilize GenAI to develop attacks.

Nevertheless, as technology evolves, theoretical GenAI-enabled attacks may become practical. SAP therefore chooses to consider the most likely theoretical threats in our defensive strategy:

AI-Assisted Attack Development	**Intelligence Gathering**	**Increase in Attack Sophistication**
• Exploit generation	• Vulnerability hunting	• Spear phishing
• Prompt injection engineering	• Target reconnaissance	• Adaptive malware
• Attack scaling	• Deanonymization	• Multimodal attacks

Figure 3: GenAI weaponization risks

These risk categories can be further described as follows:

AI-Assisted attack development

Adversaries may use GenAI to develop attack code. This can be as simple as a coding assistant, or something more sophisticated, such as digesting a security research paper into a working exploit. Development of attack prompts is also a possibility, and will be accelerated if the adversary has access to the GenAI model under attack. Automation enabled by GenAI may also allow adversaries to scale targeted attacks much more easily, reducing the cost of attacks and increasing their incentive.

Intelligence gathering

Searching through vast amounts of data is a time-consuming effort. With GenAI's ability to summarize and identify connections between data, the cost for adversaries to research targets may become much lower. Adversaries are likely to discover new insights across social media accounts, emails, compromised password databases, and more – potentially leading to deanonymization of large datasets. Intelligence gathering also extends to software by discovering new security flaws in code repositories.

Increase in attack sophistication

GenAI may enable adversaries to create personalized attacks at far less cost in time and effort. We may therefore see a rise in spear phishing campaigns, especially if combined with other capabilities of GenAI such as intelligence gathering and multimodal generation. An unsuspecting victim may be easily convinced if targeted via multiple channels in a single coherent real-time and interactive attack.

Offensive GenAI use is likely to continue evolving with security defenses. This arms race is anticipated by SAP, and mitigated via red teams leveraging offensive GenAI techniques to identify potential weaknesses during penetration testing of product and services.

SAP's Core GenAI Security Principles

SAP GenAI security strategy is based on a set of core principles that guides, but does not constrain, our approach.

We see GenAI as an assistive technology that enables better productivity by relieving cognitive burden from our customers and employees in their daily activities. By ensuring that humans are always in the loop, GenAI usage has checks and balances to reduce risk.

But this is only effective if those who are using or integrating GenAI share the responsibility for the outcome. We see secure and responsible use of GenAI as a shared responsibility between development teams, customers, supplies, and end-users. Combined with technical and process controls, shared responsibility is a part of a larger set of proactive risk mitigations.

One of our most important assets to protect is our customer and business data. Strong data processing controls such as classification and anonymization are essential to preserving the integrity of GenAI against accidental or intentional poisoning. Combined with other defenses, SAP follows a defense-in-depth approach toward data confidentiality, integrity, and availability.

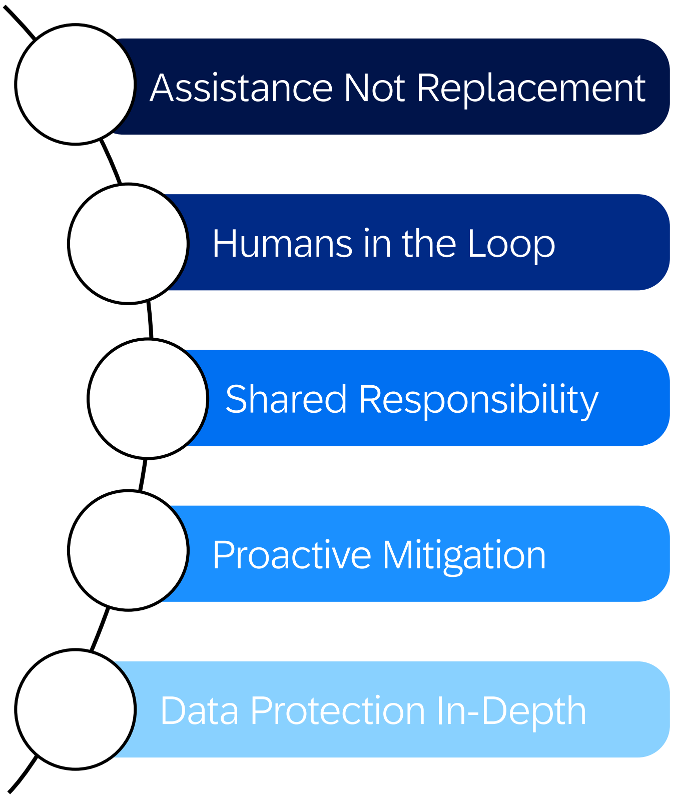


Figure 4: SAP's core GenAI security principles

> Risk reduction requires the **active participation** of all involved parties.

In Closing

SAP's continued investment in cybersecurity and prioritization of data protection for our customers has positioned us well in the marketplace for the next generation of GenAI-enabled products and services. SAP's existing software development and application security practices provide a strong foundation for future GenAI-enabled opportunities. As new advancements and applications emerge in the field of GenAI, SAP will be prepared to meet new cybersecurity challenges as they arise.

Our strategy toward GenAI security continues to evolve with continued security research, updates to our internal policies and standards, investigation of GenAI to support security operations, and exploring security dialog with suppliers, partners, and customers. We see our preparedness to leverage GenAI securely and responsibly as an invaluable differentiator to our customers, especially given potential future GenAI regulations they may need to follow.

Learn more
For more information, see the additional reading suggested below.

90449enUS (24/06) © 2024 SAP SE or an SAP affiliate company. All rights reserved. See Legal Notice on www.sap.com/legal-notice for use terms, disclaimers, disclosures, or restrictions related to this material.

Additional Reading

B. Jacques, S. Jeongmin, M. James, C. Michael and J. Raoul, “Notes from the AI frontier: Modeling the impact of AI on the world economy,” 4 September 2018. [Online]. Available: https://www.mckinsey.com/featured-insights/artificial-intelligencenotes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy.

M. Bezzi, “Assessing the Security of Large Language Models,” SAP Security Research, August 2023. [Online]. Available: https://www.sap.com/documents/2023/09/729b7ea1-8a7e-0010-bca6-c68f7e60039b.html.

B. Lake, T. Ullman, J. Tenenbaum and S. Gershman, “Building machines that learn and think like people,” Cambridge University Press, 24 November 2016. [Online]. Available: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/building-machines-that-learn-and-think-like-people/A9535B1D745A0377E16C590E14B94993.

M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel and et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” Future of Humanity Institute, 20 Feburary 2018. [Online]. Available: https://arxiv.org/abs/1802.07228.

SAP, “SAP’s Guiding Principles for Artificial Intelligence,” [Online]. Available: https://www.sap.com/products/artificial-intelligence/ai-ethics.html. [Accessed 13 September 2023].

NIST, “NIST Cybersecurity Framework 2.0,” [Online]. Available: https://csrc.nist.gov/pubs/cswp/29/the-nist-cybersecurity-framework-20/ipd. [Accessed 8 August 2023].

SAP, “SAP AI Core Security Documentation,” [Online]. Available: https://help.sap.com/docs/sap-ai-core/sap-ai-core-service-guide/security#multitenancy. [Accessed 5 October 2023].

SAP, “SAP Data Processing Agreements,” [Online]. Available: https://www.sap.com/about/trust-center/agreements/on-premise/data-processing-agreements.html. [Accessed 5 October 2023].

W. Knight, “It Costs Just $400 to Build an AI Disinformation Machine,” WIRED, 29 August 2023. [Online]. Available: https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/.

SAP, “The Secure Software Development,” 2020. [Online]. Available: https://www.sap.com/about/trust-center/security.asset-id-a248a699-627c-0010-82c7-eda71af511fa.html?pdf-asset=a248a699-627c-0010-82c7-eda71af511fa.