Gartner®

4 Ways Generative AI Will Impact CISOs and Their Teams

Published 29 June 2023 - ID G00793265 - 32 min read

By Analyst(s): Jeremy D'Hoinne, Avivah Litan, Peter Firstbrook

Initiatives: Cyber Risk; Artificial Intelligence; Build and Optimize Cybersecurity Programs

ChatGPT and large language models are the early signs of how generative AI will shape many business processes. Security and risk management leaders, specifically CISOs, and their teams need to secure how their organization builds and consumes generative AI, and navigate its impacts on cybersecurity.

Overview

Impacts

- A proliferation of overoptimistic generative AI (GenAI) announcements in the security and risk management markets could still drive promising improvements in productivity and accuracy for security teams, but also lead to waste and disappointments.

- Consumption of GenAI applications, such as large language models (LLMs), from business experiments and unmanaged, ad hoc employee adoption creates new attack surfaces and risks on individual privacy, sensitive data and organizational intellectual property (IP).

- Many businesses are rushing to capitalize on their IP and develop their own GenAI applications, creating new requirements for AI application security.

- Attackers will use GenAI. They've started with the creation of more seemingly authentic content, phishing lures and impersonating humans at scale. The uncertainty about how successfully they can leverage GenAI for more sophisticated attacks will require more flexible cybersecurity roadmaps.

Recommendations

To address the various impacts of generative AI on their organizations' security programs, chief information security officers (CISOs) and their teams must:

- Initiate experiments of "generative cybersecurity AI," starting with chat assistants for security operations center (SOCs) and application security.
- Work with organizational counterparts who have active interests in GenAI, such as those in legal and compliance, and lines of business to formulate user policies, training and guidance. This will help minimize unsanctioned uses of GenAI and reduce privacy and copyright infringement risks.
- Apply the AI trust, risk and security management (AI TRiSM) framework when developing new first-party, or consuming new third-party, applications leveraging LLMs and GenAI.
- Reinforce methods for how they assess exposure to unpredictable threats, and measure changes in the efficacy of their controls, as they cannot guess if and how malicious actors might use GenAI.

Strategic Planning Assumptions

By 2027, generative AI will contribute to a 30% reduction in false positive rates for application security testing and threat detection by refining results from other techniques to categorize benign from malicious events.

Through 2025, attacks leveraging generative AI will force security-conscious organizations to lower thresholds for detecting suspicious activity, generating more false alerts, and thus requiring more — not less — human response.

Introduction

The level of hype, scale and speed of adoption of ChatGPT has raised end-user awareness of LLMs, leading to uncontrolled uses of LLM applications. It has also opened the floodgates to business experiments and a wave of AI-based startups promising unique value propositions from new LLM and GenAI applications. Many business and IT project teams have already launched GenAI initiatives, or will start soon.

CISOs and security teams need to prepare for impacts from generative AI in four different areas (see also Figure 1):

1. **"Defend with"** generative cybersecurity AI: Receive the mandate to exploit GenAI opportunities to improve security and risk management, optimize resources, defend against emerging attack techniques or even reduce costs.

2. “**Attacked by**” GenAI: Adapt to malicious actors evolving their techniques or even exploiting new attack vectors thanks to the development of GenAI tools and techniques.

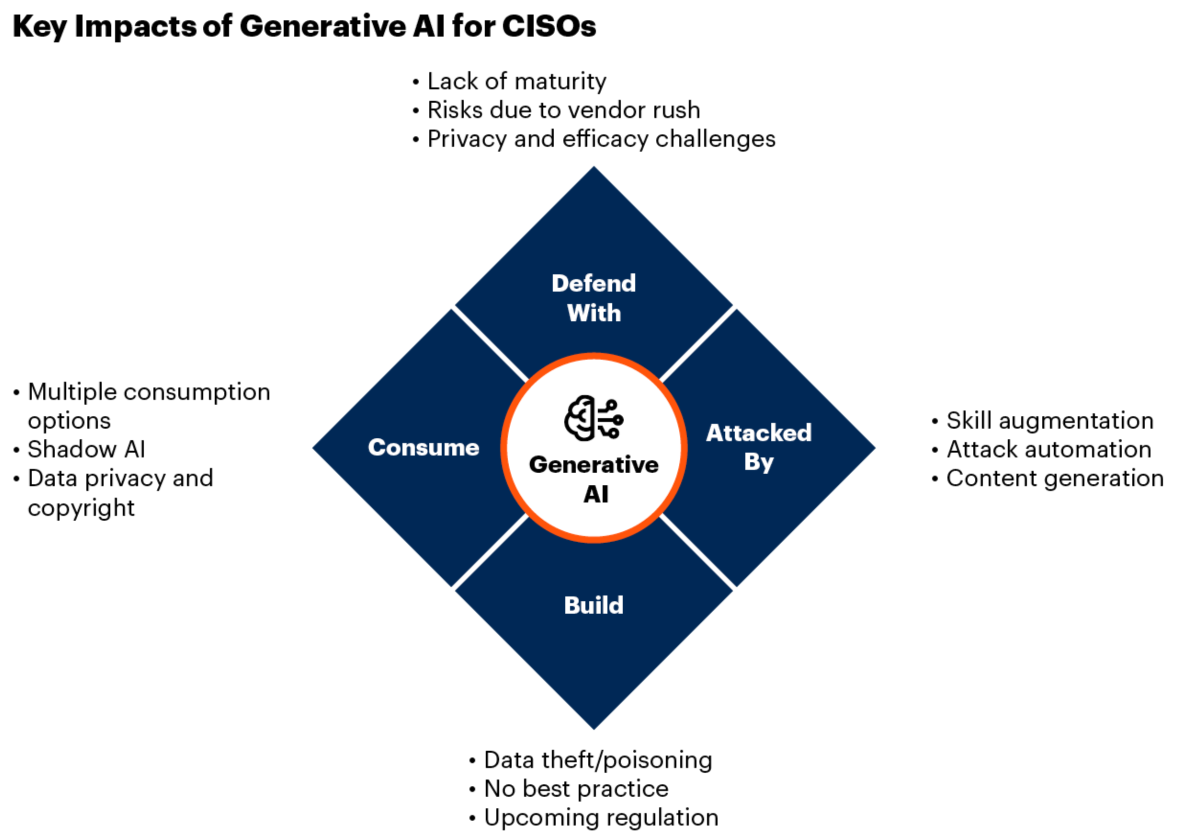
3. Secure enterprise initiatives to “**build**” GenAI applications: AI applications have an expanded attack surface and pose new potential risks that require adjustments to existing application security practices.

4. Manage and monitor how the organization “**consumes**” GenAI: ChatGPT was the first example; embedded GenAI assistants in existing applications will be the next. These applications all have unique security requirements that are not fulfilled by legacy security controls.

> Businesses will embrace generative AI, regardless of security.

This research provides clarity through the hype and gives actionable recommendations for each of these four impact areas, enabling CISOs and their teams to quickly adapt to this fast pace of change.

Figure 1: Key Impacts of Generative AI for CISOs



Source: Gartner
793265_C

Impacts and Recommendations

Overoptimistic GenAI Announcements in the Security and Risk Management Could Drive Improvements, but Also Lead to Waste and Disappointments

Gartner defines generative cybersecurity AI as GenAI techniques that learn a representation of artifacts from existing (cybersecurity) data or through simulation agents and then use it to generate new artifacts.

As its name suggests, generative cybersecurity AI derives its main use cases from GenAI (see Note 1). Gartner has reviewed more than 30 announcements from security providers. The two most frequent generative cybersecurity AI use cases are **secure application development assistants** and **security operations chatbots**.

Secure Application Development Assistants

Code generation tools (e.g., GitHub Copilot, Amazon CodeWhisperer) are embedding security features, and application security tools are already leveraging LLM applications. Example use cases for these secure code assistants include:

- **Targeting primarily application security teams:**
  - **Vulnerability detection:** Highlights issues in code snippets entered in the prompts or by performing a scan of the source code.
  - **False positive reduction:** Used as a confirmation layer for other code analysis techniques, the feature analyzes the alert and the related code and indicates in conversational language why it might be a false positive.
  - **Remediation assistant:** Suggests updates in the code to fix identified vulnerabilities as part of the finding summary that is generated.
- **Targeting primarily development teams, not security personnel:**
  - **Code generation:** Creates script/code from developers input including natural language comments or instructions, or acts as an advanced code completion tool. Some tools can also indicate whether the generated code resembles an open-source project, or can help validate that code (generated or human written) complies with standard industry security practices.
  - **Unit test creation:** Suggests a series of tests for a submitted function to ensure its behavior is as expected and is resistant to malicious inputs.
  - **Code documentation:** Generates explanations of what a piece of code does, or the impact of a code change.

While the use cases are easy to identify, it is still really early to get relevant qualitative measurement of these assistants. While some ad hoc evaluations of ChatGPT invite caution, [1] specialized providers are investing a lot of resources in these tools and Gartner anticipates broad adoption in the future.

Security Operation Tool Integration

GenAI utilities have made their way into a number of security operations tools from vendors like Microsoft, [2] Google, [3] SentinelOne, [4] CrowdStrike [5] and Cisco. These utilities have the potential to improve the productivity and skill set of the average administrator, and improve security outcomes and communication.

The first wave of GenAI implementation within SOC tools consists of conversational prompts replacing existing query-based search features, and as the front end for the users. This new interface reduces the skill requirements to use the tool, reducing the length of the learning curve and enabling more users to benefit from the tools.

> Security operation chatbots make it easier to surface insights from SOC tools. But experienced analysts are still needed to assess the quality of the outputs, detect potential hallucination and take appropriate actions according to the organization's requirements.

These GenAI features will be increasingly embedded in existing security tools to improve operator proficiency and productivity. The first implementations of these prompts support threat analysis and threat hunting workflows:

- **Interactive threat intelligence:** Access to technical documentation, threat intelligence from the security provider, or crowdsourced (using various methods separately or in combination, including prompt engineering, retrieval augmented generation and querying an LLM model using an API).
- **Alert enrichment**: Automatically add contextual information to an alert, including threat intelligence, or categorization in known frameworks.
- **Alert/risk scoring explanation:** Refines existing scoring mechanisms to identify false positives, or contribute to an existing risk-scoring engine.
- **Attack surface/threat summarization:** Aggregate multiple alerts and available telemetry information to summarize the content according to the target reader use case.
- **Mitigation assistants:** Suggest changes in security controls; new or improved detection rules.
- **Security engineering automation:** Generate security automation code and playbooks on demand, leveraging the conversational prompt. [6]
- **Documentation:** Develop, manage and maintain cohesive security policy documentation and best practices policies and procedures.

For people unfamiliar with the preexisting capabilities of security tools that add GenAI features, these features might look more impressive than they really are. The novelty might reside mostly in the interactivity, which has value, but the search and analysis capabilities are already available.

Short-Term Consequences for Security Teams

Generative cybersecurity AI will apply to many other initiatives. Content summarization and categorization could have a big impact on compliance and audit. Extending from what we see in application security testing (AST), vulnerability assessment and prioritization could also see improvements by leveraging GenAI for false positive reduction, alert enrichment and mitigation suggestions.

Proofs of concept also exist for red teaming automation, mainly tools suggesting "recipes" to perform penetration testing, such as the PentestGPT software, or for vulnerability assessment automation. [7]

But GenAI goes beyond generating new content. It can be used as part of an insight pipeline, or to analyze and classify data, which could improve malware and threat detection. [8] Generative cybersecurity AI on real-time content (e.g., data in transit, network traffic) might take more time to arrive, as it is likely to require specialized (and maybe smaller) models trained on such data.

> **Prepare for a flood of vendors with existing structured datasets (e.g., categorized emails or malwares) to release new GenAI capabilities for threat and anomaly detection.**

The Promise of Fully Automated Defense

In a second phase, starting in 2024, Gartner expects security providers to release features experimenting with the potential of composite AI — integrating multiple AI models and combining them with other techniques. As an example, a GenAI module could create a remediation logic in reaction to the detection of a new malware, and multiple autonomous agents would implement the actions from the recipe, which could include quarantining the host, deleting unopened emails, or sharing indicators of compromise (IOCs) with other organizations. The promise of automated action to prevent an attack, contain lateral movement or adapt a policy in real time is a differentiator too good for enterprises — if it works — and too good for technology providers to pass on.

> The main challenge with automated response is the accountability of stakeholders, not the absence of technical capabilities. The lack of transparency on GenAI mechanisms and data sources will increase security leaders' hesitancy to automate action directly from outputs of generative cybersecurity AI applications.

The inability to explain and document the generated response will limit (but not fully stop) response automation, especially in critical operations or with first-line operations personnel. Mandatory approval workflows and detailed documentation will be necessary until the organization gains enough trust in the engine to incrementally increase the level of automation. Good explainability could also become a differentiator for security providers or a legal mandate in some territories.

Key Challenges and Recommendations When Embracing Generative Cybersecurity AI

Examples and first demos of generative cybersecurity AI products and features appeal to many security professionals. However, inquiries with Gartner's CISO clients show that these benefits come with a few identified questions and challenges:

1. **Short-term staff productivity:** Will the alert enrichment reduce diagnosis fatigue or make it much worse by adding generated content? Junior staff might only get fatigued by the amount of data because they can't really determine whether it makes sense.

2. **Privacy and third-party dependencies:** As providers rush to release features, many of them leverage a third-party LLM, using an API to interact with a GenAI provider, or use third-party libraries or models directly. This new dependency might create privacy issues and third-party risk management challenges.

3. **Costs:** Many of the new generative cybersecurity AI features and products are currently in private beta or preview. There is little information on the impact these features will have on security solution prices. Commercial models are generally priced based on a volume of tokens used, and security providers are likely to make their clients pay for it. Training and developing a model is also expensive. The cost of using GenAI might be much higher than the cost of other techniques addressing the same use case.

4. **Overall quality:** For most of the early implementations of generative cybersecurity AI applications, organizations will aim at "good enough" and basic skill augmentation. Still, first evaluations of the secure code assistant outputs quality gives mixed results. [9] Similarly, threat intelligence and alert scoring features might be biased by the model's training set or impacted by hallucination (fabricated inaccurate outputs).

5. **Regression to the mean versus state of the art:** For specialized use cases, such as incident response against advanced attacks, the quality of the outputs issued by GenAI might not be up to the standard of the most experienced teams. This is because its outputs partially come from crowdsourced training datasets issued from lower maturity practices.

Generative cybersecurity AI will impact security and risk management teams, but security leaders should also prepare for "external/indirect" impact of GenAI on security programs, such as assisted RFP analysis, code annotation, and various other content generation and automation affecting compliance, HR and many other teams.

Recommendations:

- Run experiments with new features from existing security providers, starting with targeted and narrow use cases in the security operation and application security areas.
- Establish or extend existing metrics to benchmark generative cybersecurity AI against other approaches, and to measure expected productivity improvements.
- Determine your corporate position on providing feedback to the applications and improve their efficacy in the long run.
- Identify changes in data processing and supply chain dependencies and require transparency about data usage by your security providers.
- Adapt documentation and traceability processes to ensure internal knowledge augmentation and avoid only feeding the tools with your insights.
- Choose fine-tuned or specialized models that align with the relevant security use case or for more advanced teams.
- Advanced security use cases might not be met by consuming GenAI base models "as is."

- Prepare and train your team for dealing with direct (privacy, IP, AI application security) and indirect impacts (other teams using GenAI, such as HR, finance or procurements) coming from generative AI uses in the enterprise.

Consumption of GenAI Applications From Business Experiments and Unmanaged, Ad Hoc Employee Adoption Creates New Attack Surfaces and Risks

Since the launch of ChatGPT at the end of 2022, Gartner has dealt with multiple inquiries from CISOs who want to better understand the risks of the chatbot or other enterprise applications, and what they should do about it. In Quick Answer: How Can You Manage Trust, Risk and Security for ChatGPT Usage in Your Enterprise?, Gartner highlights immediate risks from unsanctioned used of LLM applications, such as:

1. **Sensitive data exposure:** Rules on what providers do with the data sent in the prompt will vary per provider. Enterprises have no method to verify if their prompt data remains private, according to vendor claims. They must rely on vendor license agreements to enforce data confidentiality.

2. **Potential copyright violations:** The responsibility for copyright violation coming from the generated outputs (based on training data that organizations can't identify) falls back on the users and enterprises using it. There might also be rules of usage for the generated output.

3. **Biased, incomplete or wrong responses:** The risk of "AI hallucinations" (fabricated answers) is real, [10] but answers might also be wrong due to relying on biased, fragmented or obsolete training datasets.

4. **LLM content input and output policy violations:** Enterprises can control prompt inputs using legacy security controls, but they need another layer of controls to ensure the crafted prompt meets their policy guidelines — for example, around the transmission of questions that violate preset ethical guidelines. LLM outputs must also be monitored and controlled so that they too meet company policies — for example, around domain-specific harmful content. Legacy security controls do not provide this type of domain-policy-specific content monitoring.

5. **Brand damage:** Beyond the clumsiness of "regenerate response" or "as an AI language model" found in customer-facing content, customers of your organization are likely to expect some level of transparency. [11,12]

The consensus from these inquiries is that employees and organizations want to leverage the benefits of GenAI for their daily tasks, and are not ready to wait for security teams to be ready.

> **Eighty-nine percent of business technologists would bypass cybersecurity guidance to meet a business objective. Organizations should expect shadow generative AI, especially from business teams whose main task is to generate content or code.[13]**

Consumption of GenAI applications takes four main forms (refer also to How to Deploy Generative AI — An Analysis of Various Approaches):

- **Third-party applications or agents** such as the web-based or mobile app versions of ChatGPT (out of the box).
- **Generative AI embedded in enterprise applications:** Organizations can directly use commercial applications that have GenAI capabilities embedded within them. An example of this would be using an established software application (see AI Design Patterns for Large Language Models) that now includes LLMs (like Microsoft 365 Copilot or image-generation capabilities like Adobe Firefly). See also Quick Answer: How to Make Microsoft 365 Copilot Enterprise-Ready From a Security and Risk Perspective.
- **Embed model APIs into custom applications:** Enterprises can build their own applications, integrating GenAI via foundation model APIs. Most closed-source GenAI models such as GPT-3, GPT-4 and PaLM are available for deployment via cloud APIs. This approach can be further refined by prompt engineering — this could include templates, examples or the organization's own data — to better inform the LLM output. An example would be searching a private document database to find relevant data to add into the foundation model's prompt, augmenting its response with this additional relevant, similar information.

- **Extend GenAI via fine-tuning:** Fine-tuning takes an LLM and further trains it on a smaller dataset for a specific use case. (Please note that not all LLMs can be fine tuned. See OpenAI on Fine-Tuning for some examples.) For instance, an insurance company could fine-tune a foundation model with its own policy documents, incorporate this knowledge into the model and improve its performance on specific use cases. Prompt engineering approaches are limited by the context window of the models, whereas fine-tuning enables a larger corpus of data to be incorporated.

The integration of third-party models and fine-tuning blur the lines between consumption and building your own GenAI application.

> **Custom code to integrate or privately host third-party models requires security teams to expand their controls beyond those required for consumption of third-party AI services, applications or agents, by adding the infrastructure and internal application life cycle attack surfaces.**

The ability to prevent unsanctioned uses is limited, especially since employees can access GenAI applications and commercial or open-source LLMs from unmanaged devices. Still, organizations can leverage five tactics to gain better control of GenAI consumption (see also Figure 2):

1. **Define a governance entity and workflow:** The immediate objective is to establish an inventory for all business and projects, and to define acceptable use cases and policy requirements. GenAI applications might require a specific approval workflow and continual usage monitoring where possible, but also periodic user attestations that actual usage conforms to preset intentions (see Applying AI — Governance and Risk Management).

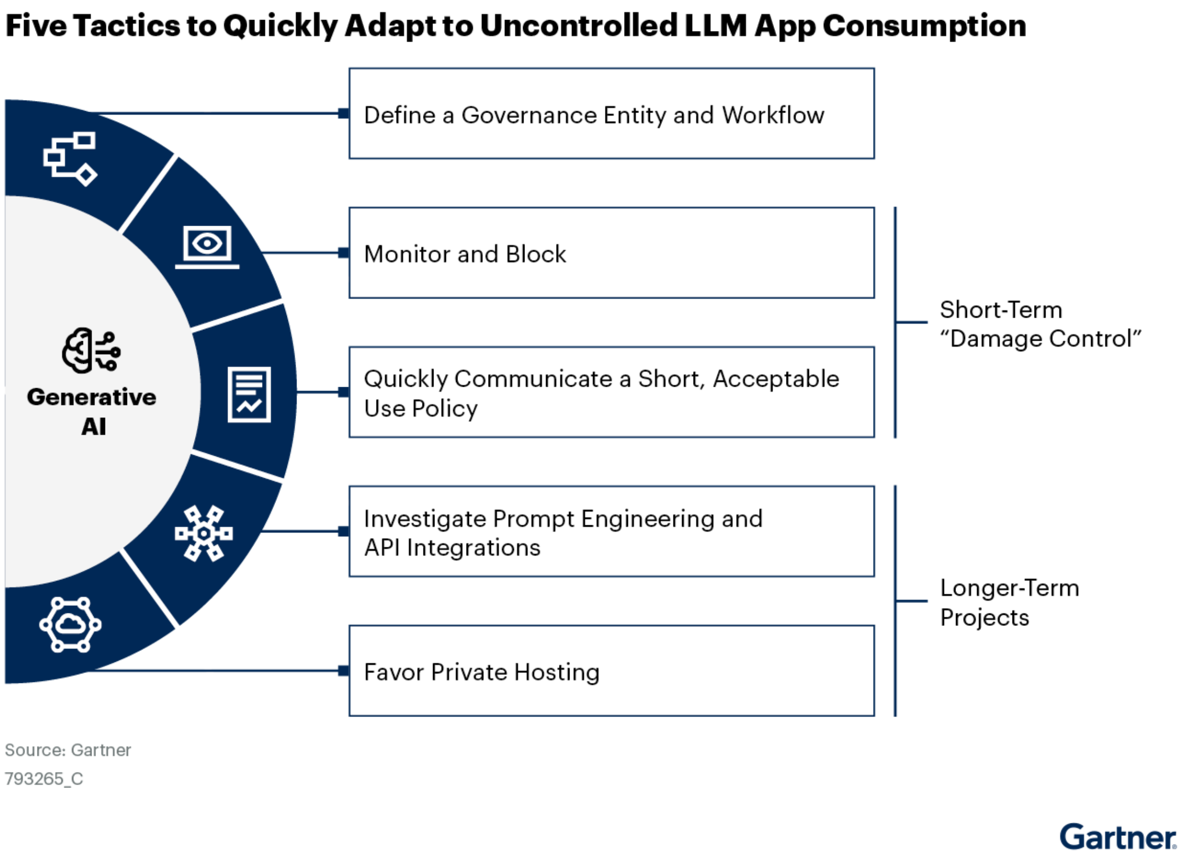
2. **Monitor and block:** When ChatGPT became public, many organizations decided to block access to OpenAI domains or apply some level of data leakage prevention, leveraging their existing security controls, such as security service edge (SSE) providers that can intercept web traffic to known applications. [14]

3. **Quickly communicate a short acceptable use policy:** Often, a one- or two-page policy document can be used to share internal contact for approval, highlight the risks of the applications, forbid the use of client data, and request documentation and traceability of outputs generated by these applications.

4. **Investigate prompt engineering and API integrations:** Intercepting inputs and outputs with a custom-made prompt might improve results but also enable more security controls. Larger organizations might investigate prompt augmentation techniques, such as retrieval augmented generation (RAG).

5. **Favor private hosting** options when available, which affords extra security and privacy controls.

Figure 2: Five Tactics to Quickly Adapt to Uncontrolled LLM App Consumption


Five Tactics to Quickly Adapt to Uncontrolled LLM App Consumption
Generative AI
Define a Governance Entity and Workflow
Monitor and Block
Quickly Communicate a Short, Acceptable Use Policy
Investigate Prompt Engineering and API Integrations
Favor Private Hosting
Short-Term "Damage Control"
Longer-Term Projects
Gartner

Source: Gartner
793265_C

Security leaders must acknowledge that blocking known domains and applications is not a sustainable or comprehensive option. It will also trigger "user bypasses," where employees would share corporate data with unmonitored personal devices in order to get access to the tools. Many organizations already shifted from blocking to an "approval" page with a link to the organization's policy and a form to submit an access request.

Generative AI Consumption Is a New Attack Surface

As with any innovation (consider the rise of social media and cryptocurrency as two recent examples), malicious actors will seek creative methods to exploit the immaturity of GenAI security practices and awareness in novel ways.

Perhaps the biggest risk of generative AI will be its potential to rapidly create believable fake content to influence popular opinion and provide seemingly legitimate content to add a layer of authenticity to scams (see the section on how attackers will leverage generative AI later in this document). When securing GenAI consumption, CISOs and their teams should anticipate the following changes in the attack surface:

- **Generative AI as a lure:** The popularity of GenAI will result in significant use of the topic as a lure to attract victims and as a new form of potential fraud. Expect counterfeit GenAI apps, browser plug-ins, applications and websites. [15,16]
- **Digital supply chain:** Attackers will exploit any weakness in GenAI components that will be widely deployed as microservices in popular business applications. These subroutines could be impacted by a number of machine learning (ML) attack techniques, such as training data manipulation and other attacks to manipulate the response. Anticipate critical vulnerabilities and patch management for these embedded components (i.e., Log4j).
- **Adversarial prompting:** The application direct and indirect prompts are a prominent attack surface. The notions of “prompt injections” and “adversarial prompting” emerge as threats when consuming third-party GenAI applications, or when building your own.

> “Prompt injection” is an adversarial prompting technique. It describes the ability to insert hidden instructions or context in the application’s conversational or system prompt interfaces, or in the generated outputs.

Early research work shows how to exploit application prompts. Security teams need to consider this new interface as a potential attack surface. [17] Gartner expects existing security tools to embed prompt security features and new providers to emerge with prompt security services, but gaps between attackers and security controls will remain for a few months, at least.

Audits and Regulations Will Impact Generative AI Consumption

Upcoming regulations are also a latent threat for organizations that are consuming (and building) AI applications. As laws might change requirements for the providers, organizations from heavily regulated industries, or in regions with more stringent privacy laws that are also actively seeking to enact AI regulations, might need to put a hold on or revert back the consumption of LLM applications.

The level of details might differ depending on the sensitivity of the content. The applications and services might include logging, but organizations might need to implement their own process for the most sensitive content, such as code, rules and scripts deployed in production.

Recommendations:

Governance:

In addition to the five tactics highlighted previously, CISOs and their teams should work together with relevant stakeholders and organizational counterparts to:

- Build a corporate policy that clearly lists governance rules, includes all necessary workflows, and leverages existing data classification to restrict uses of sensitive data in prompts and third-party applications.
- Formulate user policies, training and guidance to minimize unsanctioned uses of GenAI, privacy and copyright infringement risks.
- Require completion of the necessary impact assessments demanded by privacy and AI regulations, such as the EU's General Data Protection Regulation (GDPR) and upcoming Artificial Intelligence Act.
- Define workflows to inventory, approve and manage consumptions of GenAI. Include "generative AI as a feature," included during a software update of existing products.
- Classify use cases with the highest potential business impacts and identify the teams more likely to initiate a project quickly.

- Define new vendor risk management requirements for providers leveraging GenAI.
- Obtain and verify hosting vendors' data governance and protection assurances that confidential enterprise information transmitted to its large language model (LLM) — for example, in the form of stored prompts — is not compromised. These assurances are gained through contractual license agreements as confidentiality verification tools that run in hosted environments are not yet available.

See Applying AI — Governance and Risk Management.

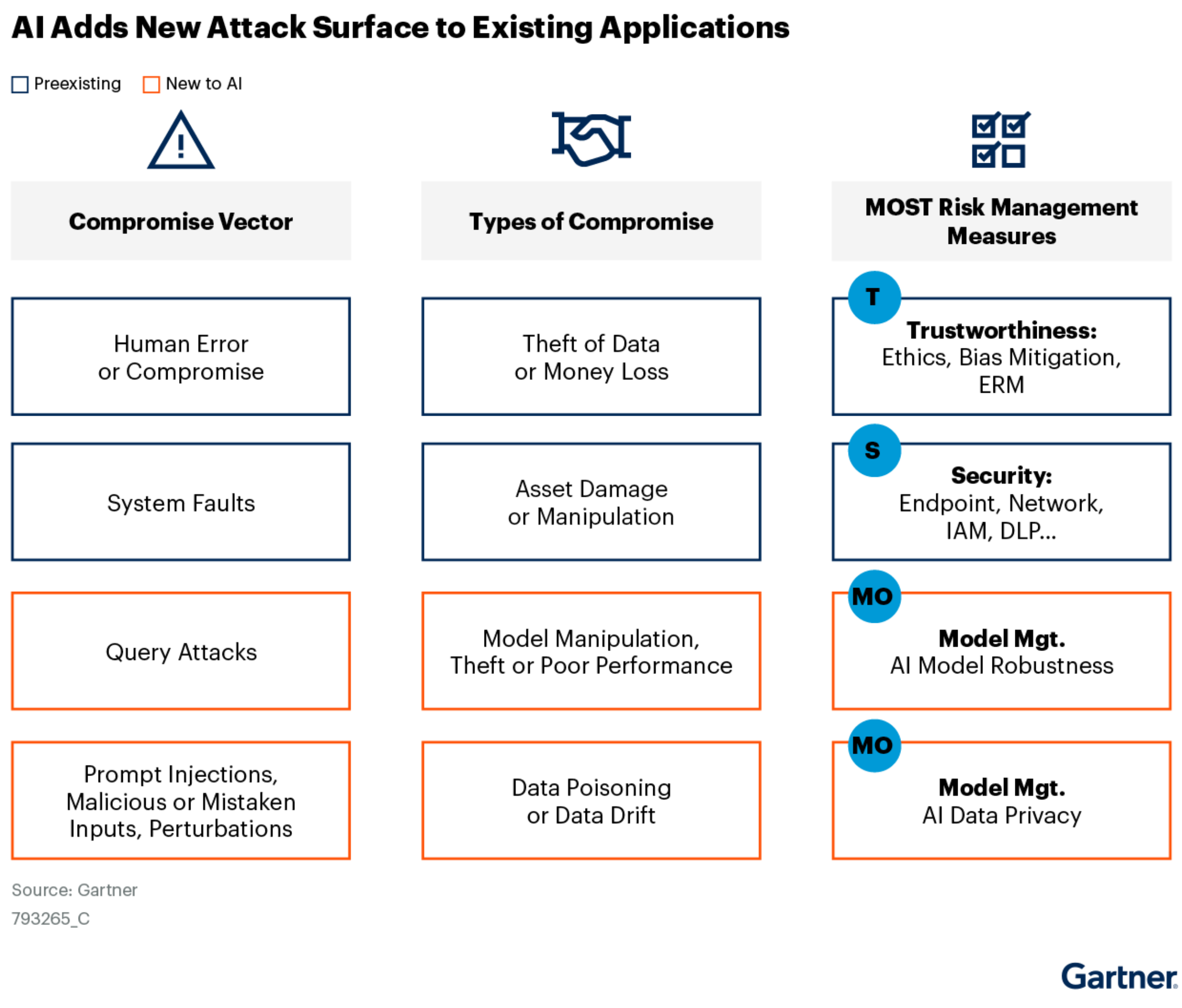
Security initiatives:

- Prioritize security resource involvement for use cases with direct financial and brand impacts, such as code automation, customer-facing content generation and customer-facing teams, such as support centers.
- Assess third-party security products for:
  - Non-AI-related controls (such as IAM, data governance, SSE functions)
  - AI-specific security (such as monitoring, controlling and managing LLM inputs)
- Prepare to evaluate emerging products enabling zero-code customization of prompts.
- Test emerging products that inspect and review outputs for potential misinformation, hallucinations, factual errors, bias, copyright violations, and other illegitimate or unwanted information that the technology might generate, which might lead to unintended or harmful outcomes. Alternatively, implement a temporary manual review process.
- Progressively deploy automated action and only with preestablished accuracy tracking metrics. Ensure that any automated action can quickly and easily be reverted back.
- Include LLM model transparency requirements when evaluating third-party applications. First tools don't include necessary visibility on users' actions.
- Consider the security advantages of private hosting of smaller or domain-specific LLMs, but work with the infrastructure and application teams to evaluate infrastructural and operational challenges.

Many Businesses, in Rushing to Capitalize on Their Own IP and Develop Their Own GenAI Applications, Create New Requirements for AI Application Security

In addition to the use of third-party applications and services, many organizations will build their own GenAI applications that will pose new risks, such as attacks on ML models and data poisoning (see Figure 3 below and Top 5 Priorities for Managing AI Risk Within Gartner's MOST Framework).

Figure 3: AI Adds New Attack Surface to Existing Applications



AI Adds New Attack Surface to Existing Applications

Compromise Vector	Types of Compromise	MOST Risk Management Measures
Human Error or Compromise	Theft of Data or Money Loss	T **Trustworthiness:** Ethics, Bias Mitigation, ERM
System Faults	Asset Damage or Manipulation	S **Security:** Endpoint, Network, IAM, DLP...
Query Attacks	Model Manipulation, Theft or Poor Performance	MO **Model Mgt.** AI Model Robustness
Prompt Injections, Malicious or Mistaken Inputs, Perturbations	Data Poisoning or Data Drift	MO **Model Mgt.** AI Data Privacy

Source: Gartner
793265_C

In Market Guide for AI Trust, Risk and Security Management, Gartner identifies four categories of requirements specific to the security and safety of AI applications: **explainability and model monitoring, ModelOps, AI application security**, and **privacy.** These requirements follow the progress of state-of-the-art AI and GenAI implementations, and will continue to evolve.

Implementing controls for GenAI applications will heavily depend on the AI model implementation. The scope of security controls will depend on whether an application includes:

- Its own model — trained and built in-house
- A third-party model trained on internal data
- Fine-tuning a third-party model
- Private hosting of an out-of-the-box model

When building and training your own model, security and development teams share the responsibility for protecting against the entire AI attack surface. When fine-tuning a model, organizations can use synthetic data. This might improve security but also negatively impact the accuracy of the applications. These techniques are not available when hosting an already pretrained out-of-the-box model.

But when including a third-party model, there is still some attack surface that must be addressed, such as adversarial direct and indirect prompt injections, and digital supply chain management challenges related to the model itself.

Then, even when hosting an out-of-the-box third-party application or model, CISOs and their teams will be responsible for infrastructure and data security, but will also have to be involved in managing the risks and vulnerabilities of the third-party AI application, agent or model.

In March 2023, OpenAI shared details about how it used “red teaming” to make GPT-4 more resistant to adversarial and other malicious inputs. Red teaming LLM applications, by combining manual and automated adversarial prompts, emerges as a necessary practice, even if it might be too complex to do for smaller teams. [18]

Gartner already sees specialized AI security vendors adding capabilities to help secure GenAI applications, notably “prompt security” services.

Recommendations:

- Adapt to hybrid development models, such as front-end wrapping (e.g., prompt engineering), private hosting of third-party models and custom GenAI applications with in-house model design and fine-tuning.

- Consider the data security options when training and fine-tuning models, notably the potential impact on accuracy or additional costs. This must be weighed against requirements in modern privacy laws, which often include the ability for individuals to request that organizations delete their data.

- Upskill your security champions as soon as training on secure GenAI coding is available. [19]

- Apply AI TRiSM principles and update security best practices to include AI applications requirements (see How to Securely Design and Operate Machine Learning).

- Add requirements for testing against adversarial prompts and prompt injections.

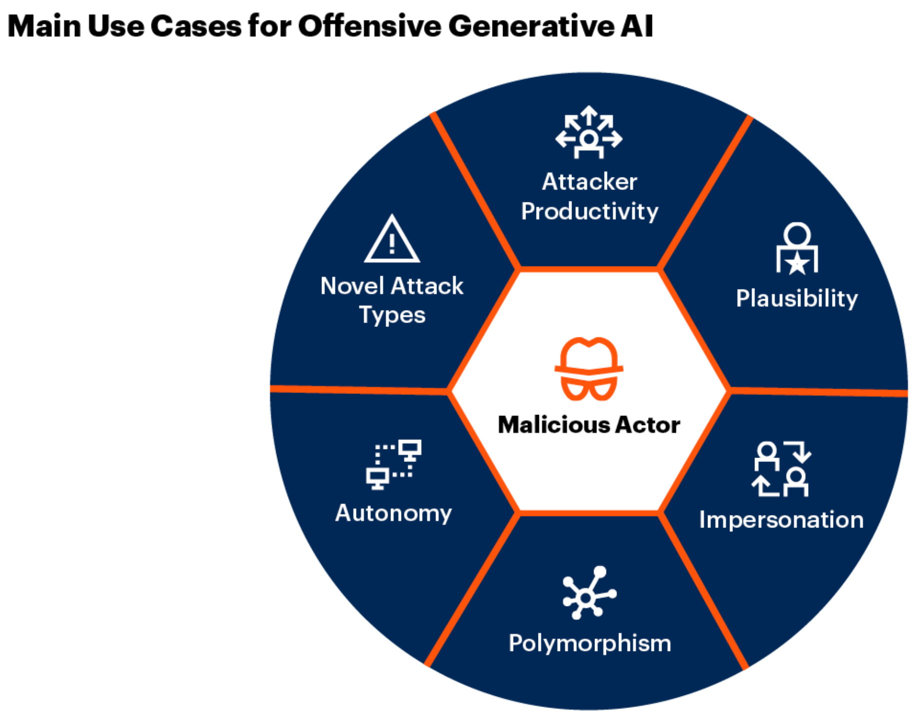
- Monitor model operation tool improvements.

Attackers Will Use Generative AI

CISOs and their teams must approach this threat without a strong fact base or direct proof of the full impact of adversarial use of GenAI. In How to Respond to the 2023 Cyberthreat Landscape, Gartner categorized "attackers using AI" as an uncertain threat. Uncertain threats can often be real but lack a direct and obvious immediate response from targeted enterprises.

GenAI is already being explored for creative use by attackers and if history is a guide, like any new technology, attackers will exploit it. Gartner expects attackers to gain the same benefits that most industries are expecting from GenAI: productivity and skill augmentation. GenAI tools will enable attackers to improve the quantity and quality of attacks at low cost. They will then leverage the technology to automate more of their workflows in areas such as scan-and-exploit activities.

Figure 4: Main Use Cases for Offensive Generative AI



Source: Gartner
793265_C

For malicious actors, the main benefits of using LLMs and generative AI include:

- **Attacker productivity:** GenAI can help attackers create more attacks through:
  - *Upskilling*: GenAI lowers the training required to write credible lure, and syntactically correct code and scripts.
  - *Automation*: Chaining tasks, providing recipes and integrating with external resources to achieve higher-level tasks.
  - *Scalability*: As a consequence of more automated content generation, attackers can rapidly develop useful content for most stages of the kill chain, or discover more vulnerabilities.
- **Plausibility:** GenAI applications can help discover and curate content from multiple sources to increase trustworthiness of a lure and other fraudulent content (e.g., brand impersonation).

- **Impersonation:** GenAI can create more realistic human voice/video (deepfakes) that appear to be from a trusted source and could undermine identity verification and voice biometrics.

- **Polymorphism:** GenAI can be used to develop multivarious attacks, which will be harder to detect than repacking polymorphism.

- **Autonomy:** Using LLM as a controller to achieve a higher level of autonomous local action decision or more automated command and control interactions (as the server component would leverage GenAI).

- **Novel attack types:** The worst-possible security threat from GenAI would be large-scale discovery of entirely new attack classes. Although the security industry will use this scenario to instill fear, there is no evidence that this has greater likelihood than discovery by human threat actors.

Attackers have started to exploit the generative content capabilities of the technology to craft better malicious and fraudulent content at scale and low cost. There are already lots of threat research examples of ChatGPT crafting more believable phishing lures, customized with public domain knowledge. [20]

It is quite clear that use of GenAI will enable attackers to generate better lures faster, but it's not clear that attack goals will diversify. Consequently, defensive solutions that focus on "the call to action" will retain value. Anti-spam solutions have been using ML for years. Bayesian spam filtering was deployed in commercial spam filters in the late 1990s. More recently, social graphing AI solutions from vendors such as Abnormal Security, Proofpoint and Tessian have evolved to detect impersonation attempts.

Although the intellectual exercise of anticipating what more advanced GenAI capabilities would do for malware is interesting, [21] GenAI's ability to produce code is primarily a skill augmentation function, enabling lower skill attackers to create malware. Although more malware and malware authors are unwelcome, existing techniques (i.e., ML antivirus detection and behavioral detection) are relevant to address this aspect of GenAI.

Polymorphic malware (rapid iteration of new code for the same attack) has been around for many years. [22] Existing detention technology, such as ML and "cloud lookup," are capable of managing polymorphism and increased volumes. When advanced cybercriminal groups start building more sophisticated malware that might rely on custom models, they'll probably deliver these malwares "as a service," as the examples of EvilProxy for multifactor authentication bypass and REvil for ransomware, shows. These more advanced malware, while more limited in number, could be more evasive and pretested by the malware creators against existing detection technologies. Generative AI is already being used to discover new vulnerabilities in code. [7] If GenAI becomes more proficient at this task than current approaches, zero-day attacks (i.e., attacks exploiting undisclosed vulnerabilities) will become more common. Gartner anticipates that this will accelerate the development of more supply chain attacks against widely deployed and privileged applications. This will likely be an area where rapid reactive response is necessary. Thus end-user organizations should develop playbooks for responding to supply chain compromise and evaluate tools and services to monitor their software dependencies.

The biggest stake for both attackers and defenders will be to find out if GenAI can repeatedly find novel attack types that can consistently bypass existing defenses.

The more significant risk is if GenAI can repeatedly find novel attack techniques that bypass existing behavioral defenses. This scenario is unlikely, at least in the near term. Creating entirely new attack techniques would require a motivated and well-funded group of smart humans to prompt the system effectively, probably a private model with appropriate training data, and output that is unique enough to be undetected by existing ML or behavioral systems in the wild. Given enough time and resources, however, GenAI could stumble upon a brand new adversarial technique. While it will be possible for an educated attacker's creative prompt input to initially evade behavioral-based detection with an unanticipated new technique, consistently generating new behavioral techniques will be difficult.

Meanwhile, as seen in previous sections, defenders are already exploring adversarial GenAI to understand its potential and develop new defenses. Overall, Gartner views GenAI primarily as a new productivity tool for already well-motivated and professional attackers, but there is no direct evidence that it will shift the balance of power to attackers.

Security leaders must avoid distraction from the hype and focus on monitoring microtrends for existing and emerging threat vectors. They then need to reinforce their investments in resilience and in reducing their exposure to categories of threats, rather than individual known attacks (see Implement a Continuous Threat Exposure Management (CTEM) Program).

Recommendations:

CISOs and their teams should adapt to the potential impact of GenAI uses by attackers:

- Consider that the right order for any investment in security is people, process and, only then, technology. Focus on the threat vectors that are tied closely to human interpretation impacted by generated content for which there are no existing technical controls.
- Monitor industry statistics to measure the impact of GenAI on the attacker landscape.
- Ensure you can measure drift in detection rate from existing controls.
- Elevate requirements for more adaptive behavioral and ML defenses in your existing security controls. Currently, only 50% of enterprise endpoints have behavioral-based detection logic.
- Challenge all existing and prospective security infrastructure vendors to outline how the product and research will evolve to address the emerging velocity of threats and tactics that are now possible due to GenAI. Beware of overstated claims in this area.
- Evaluate the business dependency of key digital supply chain software and develop playbooks for zero-day vulnerability attacks.
- Reduce the number of “blind spots” — assets, transactions and business processes that you cannot monitor for anomalies.
- Address the potential increased risk of fraudulent content and influence operations on the corporate brands.
- Add GenAI content to security awareness training and phishing simulations. Strengthen business workflow to become more resistant to realistic phishing campaigns and voice and video impersonations.

Additional contributions for this research came from Mark Horvath, Anthony Mullen and Wilco van Ginkel.

Evidence

[1] Security Code Review With ChatGPT, NCC Group.

[2] Introducing Microsoft Security Copilot, Microsoft.

[3] Introducing AI-Powered Investigation in Chronicle Security Operations, Google Cloud [Blog].

[4] Purple AI — Empowering Cybersecurity Analysts With AI-Driven Threat Hunting, Analysis & Response, SentinelOne [Blog].

[5] Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity, CrowdStrike [Blog].

[6] Improving Network Policy Enforcement Using Natural Language Processing and Programmable Networks, UKnowledge, University of Kentucky, 2022.

[7] How Generative AI Is Changing Security Research, Tenable.

[8] Introducing VirusTotal Code Insight: Empowering Threat Analysis With Generative AI, VirusTotal.

[9] Do Users Write More Insecure Code With AI Assistants?, arXiv, Cornell University.

[10] Lawyer Cites Fake Cases Invented by ChatGPT, Judge Is Not Amused, Simon Willison's Weblog.

[11] Vanderbilt University Staff Apologized for Using AI to Write an Email to Students About the Shooting at Michigan State, BuzzFeed News.

[12] 'As an AI Language Model': The Phrase That Shows How AI Is Polluting the Web, The Verge.

[13] Infographic: Build Business Technologists' Cyber Judgment to Improve Risk Decision Making.

[14] Netskope Next-Gen Secure Web Gateway Controls for ChatGPT, Netskope.

[15] The Malware Threat Landscape: NodeStealer, DuckTail, and More, Engineering at Meta.

[16] Fake Websites Impersonating Association to ChatGPT Poses High Risk, Warns Check Point Research, Check Point.

[17] Not What You've Signed Up for: Compromising Real-World LLM-Integrated Applications With Indirect Prompt Injection, arXiv, Cornell University.

[18] On the Impossible Safety of Large AI Models, arXiv, Cornell University.

[19] ChatGPT Prompt Engineering for Developers, DeepLearning.AI

[20] Generating Phishing Attacks Using ChatGPT, arXiv, Cornell University.

[21] Get Ready for the First Wave of AI Malware, SecurityWeek.

[22] Learning to Evade Static PE Machine Learning Malware Models Via Reinforcement Learning, arXiv, Cornell University.

Note 1: Generative AI Use Cases

Generative cybersecurity AI derives its use cases from three main GenAI benefits:

1. **Data analysis:** Large language models (LLMs) learn primarily from text-input only, but GenAI also leverages foundation models and uses multimodal data for its learning. A typical example for generative cybersecurity AI would include images and diagrams, but for security use cases, time series and binaries (e.g., software, malware) could be components of a training set.

2. **Conversational search:** Facilitates access to complex information using natural language and interactive ("multiple step conversation") queries.

3. **Content generation:** Summarizing, optimizing or generating new code, documentation or procedure.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Innovation Insight for Generative AI