CYBER SHERPAS

The CyberSherpas guide on Generative AI Security

It's a Generative AI World

How vCISOs, MSPs and MSSPs Can Keep their Customers Safe from Gen AI Risks

Introduction:

ChatGPT has captured the popular imagination. It became the fastest app to reach 100 million users –managing it in a couple of weeks. It is the most visible form of a huge explosion in the use of generative AI.

- According to McKinsey[1], 79% of organizations have had at least some exposure to gen
- AI. 35% of businesses have adopted gen AI.
- 22% of employees say they harness gen AI in their own work.
- Analysts predict that AI will contribute $15.7 trillion to the global economy by 2030.
- 60% of organizations with reported AI adoption are using gen AI.

In other words, AI is almost everywhere and generative AI is leading the charge. There is no holding back enterprise and consumer enthusiasm for using gen AI.

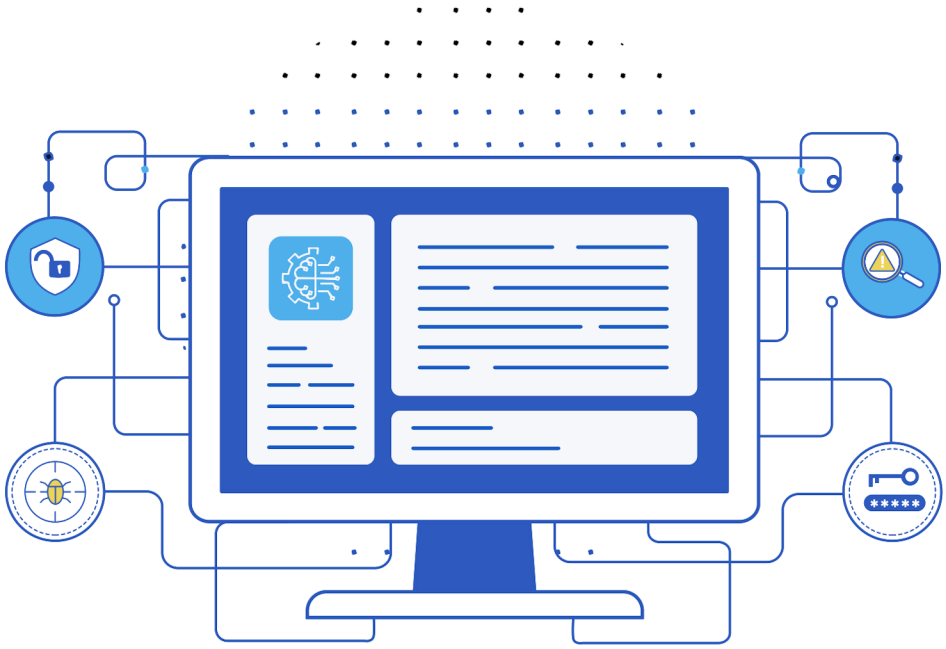
The reasons are not hard to fathom. McKinsey estimates that tech companies add 9% in revenue courtesy of gen AI adoption. Banking, pharmaceuticals and healthcare can expect a 5% boost in revenue, while education should see a 4% increase. The highest performing companies now attribute at least 20% of their earnings to gen AI and other forms of AI.

While there are obvious benefits to the use of gen AI, there are also serious risks. Unchecked AI usage in organizations can lead to:

- → Major data breaches
- → Compromised identities
- → Loss of intellectual property (IP)
- → Lawsuits for plagiarism
- → Data privacy violations

There are cases on record of gen AI-related data breaches that compromised identities, exposed sensitive information, or infringed on IP or privacy rights. The purpose of this guide, therefore, is to provide vCISOs, MSPs and MSSPs with an understanding of the risks posed by gen AI, how they can assess cybersecurity challenges in customer environments and how to establish practices on safe use of generative AI in organizations to minimize the chance of negative outcomes. With the right security tools and policies in place, service providers can shield their clients from unforeseen consequences of gen AI implementations.

[1]The state of AI in 2023: Generative AI's breakout year | McKinsey, 2023



Top Gen AI Security Risks

Generative AI is a form of AI that is designed to generate content such as text, images, video and music. It uses algorithms to analyze patterns in datasets to be able to mimic style or structure and replicate different types of content. This is the technology behind Deepfake videos and altered voice messages. It is also widely used in chatbots and search engines to provide a fast response to queries and requests.

Gen AI combines complex algorithms, deep learning neural network techniques and large language models (LLMs) to generate content and responses based on the patterns it observes in other content. Although it is classified as original material, it applies machine learning and other AI techniques to the work of others. It taps into massive repositories of content and uses that information to mimic human creativity.

Unfortunately, gen AI usage in organizations is happening far in advance of efforts to implement safeguards and constraints on possible misuse and glaring security challenges. Three primary concerns are associated with generative AI: the data employed in gen AI scripts, the outcomes produced by these tools, and the risks involved with utilizing third-party Generative AI tools.

Data Employed in Gen AI Scripts:

Data inputs and prompts into gen AI engines may inadvertently include sensitive, confidential or private data. A sales manager, for example, may use summarized or actual customer data from the CRM database as part of an input. As the person is probably using a gen AI engine in the public domain, this poses a data leakage or exposure risk.

Similarly, users inputting their own or other's healthcare data into prompts, or other personally identifiable information, are asking for trouble if those inputs are going outside of the organization. This sometimes happens when people are training models for their own purposes. If sensitive data is used to train a gen AI model, it must be masked or anonymized. Those in the EU, for instance, must ensure that their data does not go beyond national borders in violation of GDPR. Security professionals and gen AI users must be alert to such dangers.

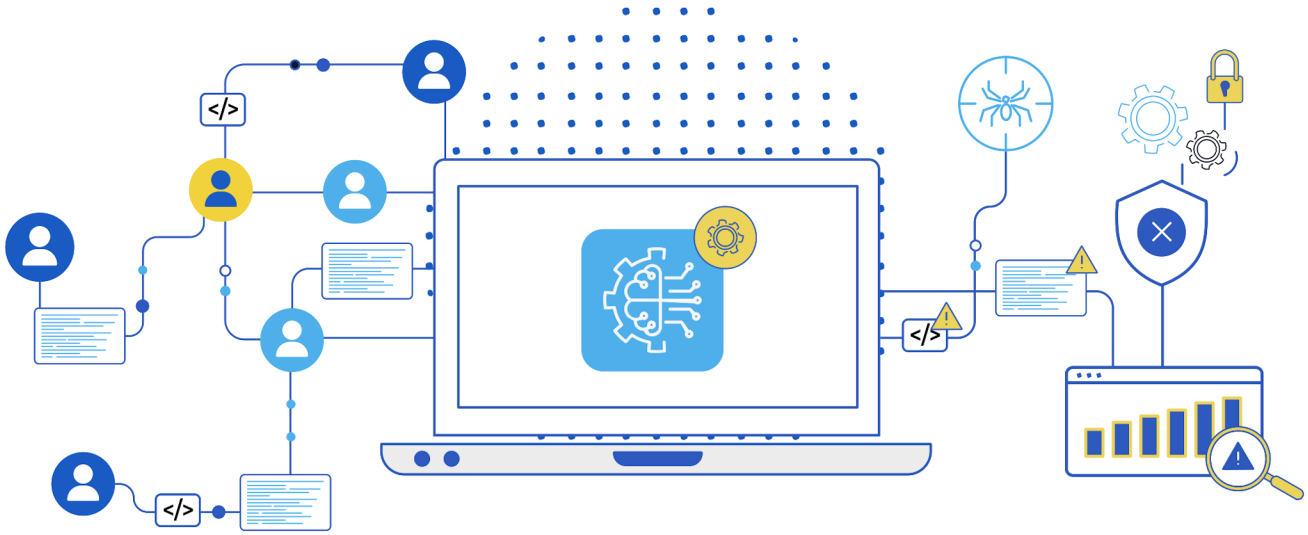
Gen AI Outcomes:

When a prompt is sent to a gen AI engine, the user receives an output in the form of a response, conclusion or answer. These outputs may contain sensitive information, bias, hallucination, proprietary information or plagiarized content.

It is quite possible for a gen AI engine to serve up sensitive information that someone else has posted on the web. The use of this data opens the user and the engine to potential legal or privacy liabilities. After all, it is not uncommon for proprietary or copyrighted information to be included in gen AI responses. A recent lawsuit by authors George Martin, John Grisham and others claims that OpenAI illegally copied and used data from their works in its responses. Comedian Sarah Silverman, Getty Images and a variety of image owners have also filed suits against misuse of their work in gen AI engines. Plagiarism is a big risk that users must be alert to.

Bias is another thorny issue. If gen AI engines utilize the web, sources such as Wikipedia and others are known to possess certain biases. Think about how the conclusions might differ if the answers came from CNN versus Fox News. Each would probably offer conflicting answers to political queries. Employees, too, can set up a model or query in such a way as to bring about bias. Anyone using gen AI responses publicly or even within the corporate intranet had better watch out for bias and avoid passing it on.

In addition, hallucination must be watched for. Gen AI engines are programmed to come up with something. That something might be completely imaginary or thoroughly embroidered, although the instances of this are relatively small. Nevertheless, if gen AI tells you that 60% of your users are of below average intelligence or states something potentially inflammatory, it might not be wise to publicize that information.



Use of Third-Party Gen AI Tools:

As well as risk related to inputs and outputs, there is risk involved when utilizing third-party generative AI tools. A key point to comprehend is the difference between public and private AI engines. OpenAI's ChatGPT is a public engine, but it is subject to more safeguards than many other public tools. Nevertheless, there are many third-party add-ons to OpenAI which add features or simplify the user interface. Inputs are being funneled through that third party, posing a risk. Further, it is easier for cybercriminals to attack that third party than to attack OpenAI. Private AI engines eliminate that danger. However, the database of information they can access is far more limited as a result. Organizations need to lay out which engines can be used for what purposes.

Another scenario: It is far from inconceivable that the bad guys could infiltrate a gen AI engine or third-party tool and quietly use it to deliver malware to users. Imagine if a company had gone to great lengths to secure its users and data only for an employee to introduce ransomware into the enterprise courtesy of a compromised gen AI application? Additionally, third-party tools that enable multiple people to be present in the same GPT session make it possible for someone to be present for malicious purposes.

The bottom line is that there is risk of exposure and compromise whenever employees using ChatGPT and other gen AI engines send data out of the organization.

Whether due to inputs, outputs or third-party tools, there are significant concerns to address in the usage of generative AI. As its pace of adoption is far in advance of awareness of the dangers, there is a need to institute measures to prevent misuse or abuse. But what exactly are those measures?

How Service Providers Can Ensure Their Customers are Protected

MSPs, MSSPs and vCISOs are the SMB's trusted partner when it comes to security. Customers hold you accountable. As such, service providers are expected to be proactive when new risks emerge, inform their customers about emerging threats and help protect them.

According to the same McKinsey report, only 21% of those using gen AI in the enterprise have established policies governing employee usage. Of those, 38% have taken steps to mitigate cybersecurity risks but only 32% are addressing the potential for gen AI inaccuracy. The rise of gen AI and its associated risks, therefore, represents a great opportunity to demonstrate superior customer service – while unearthing new possibilities.

By contacting existing customers and helping them to understand the risks presented by gen AI in their organization, it shows them that you are on top of the latest threats, understand the cybersecurity landscape, have their interests at heart and that you are proactive in your approach to the cybersecurity dangers surrounding generative AI.

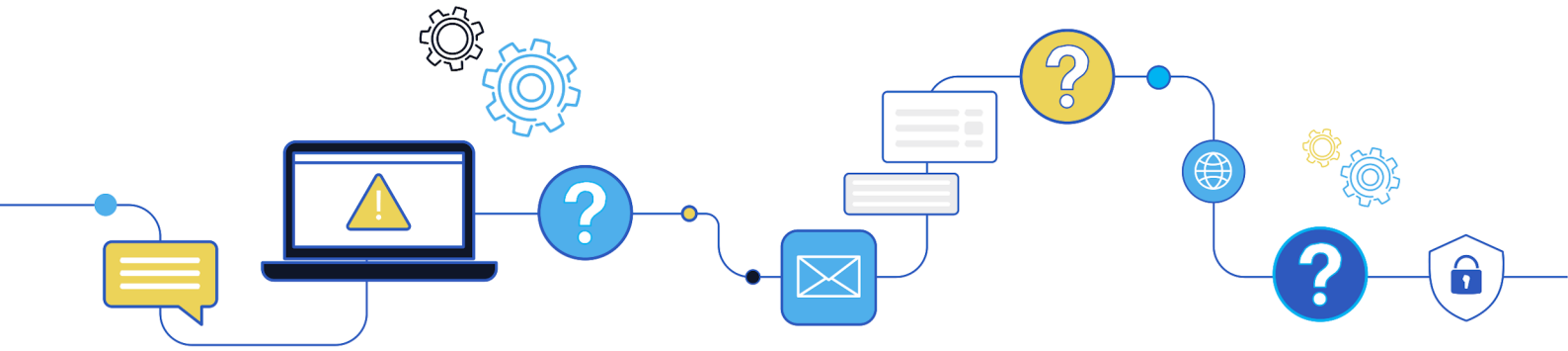
It may be wise, therefore, for service providers to address the threat posed by gen AI by a) make customers aware of the risks by raising the subject for discussion and asking questions to raise awareness of the cybersecurity challenges inherent in gen AI b) instruct them on immediate actions they can take and processes they should implement and c) recommend technologies, tools they can adopt to minimize the danger.

Raise Awareness by Questioning

Here are a few example questions to ask to raise customer awareness of gen AI risk:

- Are employees allowed to use consumer AI content generation services such as ChatGPT, Bard or Bing Chat?
- Does the company educate employees on the safe use of generative AI web services?
- Does the company prevent employees from inputting sensitive or private data into consumer generative AI products? Example: Are existing policies or controls already in place such as data loss protection (DLP), browser control, or SASE that can be used to limit the use of ChatGPT or other generative AI web service?
- Is the output of AI-generated content proofed for mistakes, fact checked and double-checked for bias or hallucinations?
- Does the organization use only certified open-source or secure foundation models?

By asking a few such questions and completing a thorough assessment of gen AI in your customer environments, you can quickly make it clear that there is work to do to ensure gen AI usage is done securely.



Immediate Actions Your Customers Should Take

Every organization using gen AI needs to carefully assess risk and adopt security policies and processes to handle data inputs and outputs appropriately. When implementing generative AI, therefore, MSPs, MSSPs and vCISOs should advise their customers to take immediate actions to enhance their security posture. Here's a list of immediate actions that can be taken:

1. Educate and Train Employees

Ensure that all users are educated about the risks associated with gen AI and trained on how to use the technology safely and responsibly.

2. Implement Robust Authentication Protocols

Enforce strong, multi-factor authentication to access any system associated with gen AI to prevent unauthorized access.

3. Use Secure and Trusted Tools

Select well-vetted, secure, and reputable gen AI tools; avoid untrusted or third-party tools that may pose security risks.

4. Regularly Update Software

Keep all software, including gen AI tools, up to date to protect against vulnerabilities.

5. Secure Sensitive Data

Protect the data used in gen AI scripts by encrypting sensitive information and implementing stringent data handling policies.

6. Ensure Safe Usage of GenAI Outputs

Establish guidelines, protocols and policies for the safe utilization of the content and information generated and monitor compliance to prevent misuse or harmful consequences.

These are some immediate actions that organizations can take to bolster their defenses against gen AI insecurity. As an MSP, MSSP or vCISO, you should instruct your customers to take these actions ASAP.

Security Technologies and Tools to Implement

In addition, there are tools and technologies that service providers can recommend to customers to minimize gen AI concerns. These include larger AI and security vendors as well as startups. Here are some examples:

Cadea:

Data security and role-based access control for AI.

CalypsoAI:

Independent testing and validation of large language models (LLMs) and protection from LLM threats.

Lasso Security:

LLMs protection suite, safeguarding every GenAI and LLM touchpoint with an easy-to-deploy solution.

LLMShield:

Prevents leaks of company secrets via public LLMs and AI chatbots.

ProtectAI:

Performs security scans on ML and AI applications.

TrojAI:

Assesses, measures, and tracks AI/ML/LLM model risks and vulnerabilities to effectively manage risk exposure.

By raising awareness of the potential dangers generative AI brings to the organization and educating your customers on the steps they can take to mitigate them, you can help them navigate the difficult waters of gen AI implementation. Security and privacy concerns may cause some to wish to put gen AI usage on pause. But there is no stopping it. The benefits far outweigh the risks. Nevertheless, the risks are substantial. You are in an excellent position to offer them real help on how to implement gen AI security measures, how to treat bias, watch for hallucination and avoid data compromise by recommending the steps covered above.

How CyberSherpas Can Help

The CyberSherpas vCISO platform incorporates dozens of security policies including the newest controls specific to gen AI use in the organization. The CyberSherpas vCISO platform automatically generates a gen AI policy customized for each organization based on its needs. Once implemented, this helps businesses ensure that the data prompts used for generative AI are appropriate, fully masked and secure – and that outputs are appropriate and avoid risk.

CyberSherpas uses generative AI protection and usage policy defines controls and best practices for protecting data and privacy, managing change, and developing software. This is critical in mitigating potential risks and ensuring alignment with relevant laws, regulations, and industry standards. Such policy typically applies to all individuals and entities involved in the usage, development, management, and oversight of generative AI within the organization. It lays out steps to maintain safe usage of consumer AI products, how generated content can and can't be used, and how to integrate AI into company products or processes.

Contact CyberSherpas today for a demonstration of how to raise awareness among your clients about the risks associated with gen AI, quickly assess their risk level, and initiate a discussion about how to protect from these threats.

The CyberSherpas platform gives you an easy way to realize your customers' risk, assist them in creating tailored generative AI security policies and recommending processes and technological tools that can help them protect from these risks.

These actions provide tangible steps your customers can take to safeguard themselves. In addition, this positions the MSP/MSSP/vCISO as a proactive partner who cares. It also provides an avenue to recommending more products and services to existing customers.

Generative AI technology is a fertile ground for emerging threats. CyberSherpas is always up-to-date and equips its users with solutions for emerging threats.

Book a demo to hear more about how CyberSherpas helps vCISOs provide comprehensive security to their customers and how it helps protect against emerging gen AI threats.

Book a demo