SOPHOS

Navigating the AI Hype in Cybersecurity

How to safely, securely take advantage of AI to enhance your organization’s cyber defenses

Contents

Introduction

Cybersecurity is awash with AI hype. Organizations are bombarded with both alluring promises of AI-powered cybersecurity transformation—elevated protection, lower costs, reduced specialist headcount needs—and dire warnings that AI is ushering in a brand-new era of cyberattacks.

This guide is designed to help organizations navigate the hype and misconceptions around AI in cybersecurity. It explains what AI can (and cannot) do to elevate organizational cyber defenses and explores the cybersecurity and operational risks that AI has introduced. The guide also provides advice on how to mitigate these risks in order to safely, securely take advantage of AI to enhance both cyber protection and return on investment.

Along the way, the guide shares insights into the reality of AI usage, expectations, and concerns based on findings from a vendor-agnostic survey of 400 IT and cybersecurity leaders conducted in late 2024. These frontline perspectives provide valuable context and act as a useful comparison point for organizations exploring their AI position. For the full findings, see Beyond the hype: The business reality of AI for cybersecurity

Ultimately, with or without AI, the goal remains the same: to optimally deliver the level of cyber resilience needed for your organization to succeed while minimizing total expenditure. Or, in other words, to best use the (invariably limited) cybersecurity budget to support the business. This guide will help you get there in the era of AI.

The benefits of AI for cybersecurity

AI is a short acronym that covers a range of capabilities that can support and accelerate cybersecurity in many ways. The good news is that AI brings greater incremental advantages to defenders than adversaries. Two common AI approaches used in cybersecurity are deep learning models and generative AI.

Deep learning

Deep learning (DL) models APPLY learnings to perform tasks. They can accelerate the application of knowledge far beyond what is possible by humans. For example, appropriately trained DL models can identify if a file is malicious or benign in a fraction of a second without ever having seen that file before.

DL is ideal for performing repetitive tasks on a vast scale. It creates a *statistical* model that views new items under the distribution of everything it has learned from its very large training data set. For example, DL models can assess millions of file samples without wavering to identify whether they contain malware. As a result, DL is widely used to elevate the protection capabilities in cybersecurity products.

DL models enable defenders to successfully deal with the huge volumes of threats created by adversaries using automation and cybercrime-as-a-service. The DL models can also be updated and adapted as attacks evolve, keeping them up to date with the threat environment.

Generative AI

Generative AI (GenAI) models assimilate inputs and use them to CREATE new content. Example applications include:

- Creating a natural language summary of threat activity to date and recommended next steps for the analyst to take
- Surfacing insights into attacker behavior by analyzing commands that create detections
- Enabling analysts to use natural language search rather than code-based queries to investigate suspicious detections
- Prioritizing application of patches based on propensity for a vulnerability to be exploited

GenAI is a powerful tool for accelerating security operations. By taking care of much of the heavy data lifting, it empowers analysts to make smart decisions fast and enables them to focus their time where it will have the greatest impact. In this way, GenAI can often relieve some of the pressure on analysts, reducing the risk of burnout and employee attrition. GenAI can also help lower the technology barrier to security operations, enabling less experienced analysts to quickly make a positive contribution and accelerate their skill development.

The journey to GenAI

The basis of modern GenAI is the transformer, a deep learning neural network that learns the context and relationship between inputs (for example, the words in a sentence) and uses this learning to create relevant outputs. Transformers are often used in natural language processing (NLP) tasks, such as translating text and providing answers to questions. In fact, the T of ChatGPT stands for Transformer.

While transformers are widely used in GenAI, not all transformers are generative. For example, BERT (Bidirectional Encoder Representations from Transformers) is an open-source machine learning framework for NLP that can read input text bidirectionally, i.e., both left-to-right and right-to-left. This approach enables it to significantly improve contextual understanding of unlabelled text. At Sophos we have been using BERT for many years to identify and defend against business email compromise attacks.

Not one-size-fits-all

AI models vary greatly in size. **Massive models**, such as Microsoft Copilot and Google Gemini, are large language models (LLMs) trained on a very extensive set of data that can perform a wide range of tasks. Conversely, small models are typically designed and trained on a very specific data set to perform a single task, such as to detect malicious URLs or executables. While more limited in scope, **small models** have cost, speed, and performance overhead advantages over larger ones.

AI limitations

AI alone is not the answer, at least not for the foreseeable future. AI complements, but does not fully replace, human expertise. Threats are hugely complex and performing effective security operations requires both technical skill and the ability to apply insights in the context of the organization. AI alone cannot keep organizations ahead of today's skilled, well-funded cybercriminal organizations.

TYPE

Deep Learning AI *Apply*

Uses artificial neural networks to recognize patterns and make decisions in a way that mimics the human brain. It APPLIES learnings to perform tasks.

Example: **Detect malicious URLs**
AI model is trained to identify malicious websites, enabling security products to block access to them

Generative AI *Create*

Uses the structure and pattern of existing data to CREATE (generate) brand new content.

Example: **Threat case summary**
AI model creates a summary of the threat activity and gives analysts recommended next steps

SIZE

Massive AI Models

Multi-purpose tools that are trained on vast quantities of publicly available data and can assist with a broad range of tasks.

Example: **Microsoft Copilot, Google Gemini**

Small AI Models

Outcome-focused models are designed, trained and built for specific use cases.

Example: **Android malware detection model**

AI adoption rates

AI is already widely embedded in the cybersecurity infrastructure of most organizations:

- 73% say their cybersecurity solutions include deep learning models
- 65% say their cybersecurity solutions include generative AI capabilities

Cybersecurity applications of AI are not limited to external vendors and 34% of organizations already use GenAI in-house to elevate their cybersecurity, for example, to help generate phishing test emails.

AI adoption is likely to become near universal within a short time frame, with AI capabilities now on the requirements list of 99% (with rounding) of organizations when selecting a cybersecurity platform:

- 57% say AI capabilities are essential/extremely important
- 41% say AI capabilities are important

With this level of adoption and future usage, understanding the risks and associated mitigations for AI in cybersecurity is a priority for organizations of all sizes and business focus.

73%
Use cybersecurity tools with deep learning models

65%
Use cybersecurity tools with GenAI capabilities

99%
Require AI capabilities when choosing a cybersecurity platform

GenAI: Great expectations

The hype surrounding GenAI has resulted in high expectations for how this technology can enhance cybersecurity outcomes. The survey revealed the top benefit that organizations want GenAI capabilities in cybersecurity tools to deliver, as shown in the table below.

#1 desired benefit from generative AI
Responses ranked first

1=	Improved protection from cyberthreats (20%)
1=	Improved return on cybersecurity spend (ROI) (20%)
3	Increased IT analyst efficiency and impact (17%)
4	Confidence that we are keeping up with cybersecurity innovations (15%)
5=	Greater peace of mind that our organization is well-defended from attacks (14%)
5=	Reduced employee burnout (i.e., automating tasks to free up cybersecurity employee time (14%)

What benefits, if any, do you want generative AI capabilities in cybersecurity tools to deliver? Responses ranked first (n=400)

The broad spread of responses reveals that there is no single, standout desired benefit from GenAI in cybersecurity. At the same time, the most common desired gains relate to improved cyber protection or business performance (both financial and operational). The data also suggests that the inclusion of GenAI capabilities in cybersecurity solutions delivers peace of mind and confidence that an organization is keeping up with the latest protection capabilities.

The positioning of reduced employee burnout at the bottom of the ranking suggests that organizations are less aware of or less concerned about the potential for GenAI to support users. With cybersecurity staff in short supply, reducing attrition is an important area for focus and one where AI can help.

Improved **protection** and increased **ROI** are the top benefits organizations want from GenAI

The risks of AI for cybersecurity

The use of AI in cybersecurity is a two-sided coin. While AI offers tremendous benefits to defenders in the battle against adversaries, it also introduces a number of risks:

1. **Threat risk:** The use of AI in cyberattacks
2. **Defense risk:** Poor quality and poorly implemented AI
3. **Operational risk:** Over-reliance on AI
4. **Financial risk:** Poor return on AI investment
5. **Hijack risk:** The compromise of public AI models by adversaries

1. Threat risk: The use of AI in cyberattacks

While there has been considerable hyperbole about how AI is creating a whole new threat landscape, the reality is less dramatic. Discussions about AI on cybercrime forums are limited in number and many threat actors remain skeptical about AI. Where observed, attempts to develop malware, attack tools, and exploits using AI are typically primitive and low-quality.

Like legitimate organizations, adversaries are primarily taking advantage of AI to improve the quality of their content and the efficiency of their operations—albeit with very different goals. For further details of the latest threat landscape and AI-led attacks, see the Sophos blog.

Improving content quality

One of the quickest, easiest and most accessible applications of AI in cyberattacks is to elevate the quality and credibility of phishing emails and scams, making victims more likely to fall for attacks.

Classic phishing "tells" such as poor grammar, bad spelling, and weak formatting are easily eliminated with AI tools. A well-written email for use in phishing campaigns can be created by public LLMs in less than a minute. Similarly, convincing and well-written texts and social media messages that aim to trick recipients into clicking links or sharing personal information are now easily accessible in any language. LLMs also make it easy for attackers to incorporate timely information into their attacks, further increasing the victim's propensity to fall for the scam.

Generative AI tools have also opened the door to a new era of scams that impersonate senior staff to trick unsuspecting victims into making financial transfers. Voice cloning technology has advanced to the point that, with enough training, adversaries can trick someone into believing they're speaking to the real person. In these voice phishing, or "vishing," attacks, a bad actor will often impersonate a senior leader and call a member of staff to "ask" them to make an illicit gift card purchase, bank payment, or file transfer.

Adversaries are also using AI-powered deepfake technology to visually impersonate people in their attacks. Deepfake videos have been used to trick unsuspecting employees into making sizeable payments and dupe facial recognition programs for loan applications and bank account registrations.

Improving operational efficiency

Just as many businesses use AI-powered chatbots to improve their users' experience, so do attackers. Some threat actors are using LLMs to enhance the forums they frequent by creating chatbots and auto-responses. In an example shared by Sophos X-Ops, the XSS forum created a dedicated forum chatbot to respond to users' questions. The administrator announced (translated from Russian):

> *"In this section, you can chat with AI (Artificial Intelligence). Ask a question – our AI bot answers you …. This section and the AI-bot are designed to solve simple technical problems, for the technical entertainment of our users [and] to familiarize users with the possibilities of AI."*

Building and training custom models requires extensive AI expertise – which is costly and in short supply. While some cybercrime gangs do have in-house AI expertise, threat actors will typically leverage existing LLMs in their attacks rather than build their own.

Attacker level-set

It's important to put the use of AI by adversaries in context. AI is just one of the many tools in the attacker toolkit. Threat actors have been using automation and cybercrime-as-a-service models to increase the scale and frequency of their attacks for several years. For many organizations, these capabilities will have a greater impact on risk exposure than AI.

2. Defense risk: Poor quality and poorly implemented AI

As we've seen, AI models are already widely embedded in organizational cyber defenses. While their intentions are invariably good, poor quality and poorly implemented AI models can inadvertently introduce considerable cybersecurity risk of their own. The propensity of AI models to introduce risk is dependent on several factors, including:

- **Quality of data on which the models are trained**. The adage "garbage in, garbage out" is particularly relevant to AI. Using low-quality data to train models risks introducing errors, while the use of unbalanced datasets has the potential to distort outputs due to over- or under-representation of certain variables. The greater the quantity of high-quality training data, the better the resulting output.

- **Expertise of the teams that create the models**. Building effective AI models for cybersecurity requires extensive understanding of two separate but complementary areas:

  - **Threats:** To identify what you need the AI model to do, you first need to understand how malware and adversaries operate.

  - **AI:** Once you know what you need the AI to do, you then need to identify and build the right model to achieve the goal.

  To build effective AI models that have a material impact on cybersecurity, it's essential that these two skillsets work closely together, leveraging their mutual expertise.

- **Quality of the product development and rollout process**. In mid-2024, the rollout of a faulty content update in a cybersecurity product brought immediate disruption to businesses around the globe. Poorly tested, quality assessed, and rolled-out AI capabilities have the potential to do even greater harm—with the added risk that the issue may not be easily identified or rectified.

A false sense of (cyber)security

Organizations are largely alert to the risk of poorly developed and deployed AI in cybersecurity solutions. The vast majority (89%) of IT/cybersecurity professionals surveyed say they are concerned about the potential for flaws in cybersecurity tools' generative AI capabilities to harm their organization, with 43% saying they are extremely concerned and 46% somewhat concerned.

It is therefore unsurprising that 99% (with rounding) of organizations say that when evaluating the GenAI capabilities in cybersecurity solutions, they assess the caliber of the cybersecurity processes and controls used in the development of the GenAI:

- 73% say they fully assess the caliber of the cybersecurity processes and controls

- 27% say they partially assess the caliber of the cybersecurity processes and controls

While the high percentage that report conducting a full assessment may initially appear encouraging, in reality it suggests that many organizations have a major blind spot in this area.

Assessing the processes and controls used to develop GenAI capabilities requires transparency from the vendor and a reasonable degree of AI knowledge by the assessor. Unfortunately, both are in short supply. Solution providers rarely make their full GenAI development rollout processes easily available, and IT teams often have limited insights into AI development best practices. For many organizations, this finding suggests that they "don't know what they don't know".

3. Operational risk: Over reliance on AI

AI touches almost every aspect of our day-to-day lives, from finding the best route to the supermarket to recommending shows on TV. Its pervasive nature makes it easy to default too readily to AI and take for granted that AI can do certain tasks better than people. Fortunately, most organizations are aware of and concerned about the cybersecurity consequences of over-reliance on AI:

- 84% are concerned about resulting pressure to reduce cybersecurity professional headcount
- 87% are concerned about a resulting lack of cybersecurity accountability

Being alert to these risks is the first step to mitigating them. It's important to remember that AI is just one tool in an organization's cyber defenses; while it is a valuable part of the security stack, it is not always the right approach and is rarely the entire solution. Each organization is different, and the use of AI should be contextual to the wider business setup and needs.

4. Financial risk: Poor return on AI investment

High caliber GenAI capabilities in cybersecurity solutions are expensive to develop and maintain. IT and cybersecurity leaders are alert to the consequences of this spend, with 80% saying that they think GenAI will significantly increase the cost of their cybersecurity products.

Despite these expectations of price increases, most organizations see GenAI as a path to lowering their overall cybersecurity expenditure, with 87% of respondents saying they are confident that the costs of GenAI in cybersecurity tools will be fully offset by the savings it delivers.

At the same time, organizations recognize that quantifying these costs is a challenge. GenAI expenses are typically built into the overall price of cybersecurity products and services, making it hard to identify how much organizations are spending on GenAI for cybersecurity. Reflecting this lack of visibility, 75% agree that these costs are hard to measure (39% strongly agree, 36% somewhat agree).

Without effective reporting, organizations risk not seeing the desired return on their investments in AI for cybersecurity or, worse, directing investments into AI that could have been more effectively spent elsewhere.

5. Hijack risk: Compromised large language models (LLMs)

The cybersecurity risks of AI extend beyond cybersecurity tools and applications. The rapid global expansion of public LLM usage has opened the door for sophisticated, well-funded actors to compromise the models themselves to help them achieve their goals. This has the potential to play out in several ways, including:

- **Data poisoning**. In their 2023 paper Poisoning Web-Scale Training Datasets is Practical, Carlini et. al. showed that data poisoning (i.e., manipulating the data on which the model is trained to influence its outputs) is a viable threat risk.
- **State actor backdoors**. Many nation states have the resources to create powerful LLMs. By adding secret backdoors and then making the models freely available for public usage, state actors can manipulate the LLM to their advantage if needed.
- **LLM spoofing**. Malicious actors can compromise legitimate LLMs (for example, by adding backdoors) and then advertise the changes as "improvements". To trick people into using their compromised tool, they spoof the reputable provider's name—such as omitting a letter or switching the letter O for the number 0.

For a deep dive into LLM compromise, see the latest research from the Sophos AI team.

Practical steps to navigate the AI hype

While AI brings risks, with a thoughtful approach, organizations can navigate them and safely, securely take advantage of AI to enhance their cyber defenses. Many of these recommendations can also be used to assist with the successful implementation of AI in other areas.

Threat risk: Up-level cyber defenses for the age of AI

A key focus should be on improving resilience to AI-powered threats. Given that adversaries are primarily leveraging AI to elevate the quality and credibility of phishing emails and scams, it makes sense to focus on these areas. Suggestions include:

- **Up-level email protection**. Look for solutions that can detect AI-generated phishing emails and scams, preventing them from reaching your users' inboxes.
- **Deploy protection against Business Email Compromise/VIP protection**. Choose email security solutions that include BEC and VIP protection, for example, scanning the content for tone and style to detect scams.
- **Be particularly suspicious of social media** – often users are not fully engaged when scrolling on social channels, so they are more likely to fall for a scam
- **Put in place processes to mitigate the risk of voice cloning**, such as procedures to follow if an unexpected payment or data sharing request is received. Options include:
  - Calling the requester back to verify the ask
  - Putting in place pass codes or phrases

Defense risk: Assess the quality of the AI used in cybersecurity products

Be alert to the risks and impact of poor-quality AI in your security investments. Ask vendors about their:

- **Training data**. What is the quality, quantity, and source of data on which the models are trained? Better inputs lead to better outputs.
- **Development team**. Find out about the people behind the models. What level of AI expertise do they have? How well do they know threats, adversary behaviors, and security operations?
- **Product engineering and rollout process**. What steps does the vendor go through when developing and deploying AI capabilities in their solutions? What checks and controls are in place?

Ultimately, ask yourself: How much do I trust this organization to be doing AI well, and to put in place the rigorous quality and deployment controls needed?

Operational risk: View AI through a human-first lens

AI won't care if you experience a breach; your people will. And if the worst happens and you get compromised, you'll need experienced team members who can understand and remediate the situation in the context of your business.

- **Maintain perspective**. AI is just one item in the defender's toolkit. Use it, but make clear that cybersecurity accountability is ultimately a human responsibility.
- **Don't replace, accelerate**. The ongoing global shortage of skilled cybersecurity professionals is widely known. Major burnout issues further exacerbate the challenge. Rather than looking at AI to reduce headcount, focus first on how AI can support your staff. By taking care of many low-level, repetitive security operations tasks and providing guided insights, AI can:
  - Free up time for more valuable, business-impacting work
  - Reduce alert overload, helping reduce fatigue
  - Accelerate professional development of skilled analysts
  - Enable less experienced analysts to perform security operations, building a resource pipeline

Financial risk: Apply business rigor to AI investment decisions

This is one of the areas that is easiest for organizations to mitigate, with many factors fully within their control.

- **Set goals**. Be clear, specific and granular about what outcomes you want AI to deliver.
    - Identify what you need. What gaps do you have? Where can AI help?
    - Consider financial, time, and protection gains.
- **Quantify benefits**. Understand how much of a difference AI investments will make.
    - If a goal is to lower the overall costs/TCO of cybersecurity, quantify what savings you will make as a result.
    - If you want AI to reduce IT/cybersecurity attrition, be clear on how exactly the AI tool will impact the team. What tasks will it take out of their queues? How many hours will it free up?
- **Prioritize investments**. AI can help in many ways; some will have a greater impact than others. Identify the important metrics for your organization – financial savings, staff attrition impact, exposure reduction, etc. – and compare how the different options rank.
- **Measure impact**. Investment decisions are made with good intentions. Be sure to see how actual performance relates to initial expectations. Are you seeing the advantages you were expecting? Are there unanticipated gains? And are there areas where you are not seeing the results you hoped for? Use these insights to make any adjustments that are needed.

Ask yourself whether AI is the best way to help you achieve your goal, or if another technology or approach would have a greater impact.

Hijack risk: Remain alert to the danger

This risk is the most difficult for organizations to mitigate. Simply being alert to this risk helps reduce the impact. That said, when choosing public LLMs, look for:

- **Models from well-known, reputable providers**. While not immune to data poisoning attacks, issues with data outputs are more likely to be publicized and shared.
- **Correct provider names**. Attackers spoof the names of reputable providers to trick people into thinking their compromised models are legitimate.

Specialists in AI for cybersecurity are actively working on approaches to neutralize this risk.

Conclusion

AI offers tremendous benefits for cybersecurity. By avoiding the AI hype and adopting a thoughtful outcome-led approach to AI, organizations can take advantage of this technology to enhance their cyber defenses and empower their valued IT and cybersecurity professionals.

About the survey

Source: Beyond the hype: The business reality of AI for cybersecurity

Sophos commissioned independent research specialist Vanson Bourne to survey 400 IT and cybersecurity leaders in organizations with between 50 and 3,000 employees. The survey was conducted during November 2024 and respondents came from 13 industry sectors. To ensure broad industry representation, the survey was vendor agnostic with respondents' organizations using endpoint security solutions from 19 separate vendors.

About Sophos

Sophos is a global cybersecurity leader offering a full portfolio of award-winning cybersecurity products and services, from firewalls, endpoint protection and EDR/XDR tools to managed detection and response (MDR) and incident response (IR) services.

Sophos has been elevating cybersecurity with AI since 2017, bringing together AI and human expertise to stop the broadest range of threats, wherever they run. Deep learning and generative AI capabilities that solve the most critical problems for customers are embedded across our products and services and delivered through the largest AI-native security platform in the industry. Trained on data from attacks in more than 600,000 diverse customer environments, our adaptive AI platform delivers unrivalled protection from advanced threats and enhances the power of defenders.

To learn more and explore Sophos solutions visit www.sophos.com

2025-01-15 (WP-MP)

SOPHOS