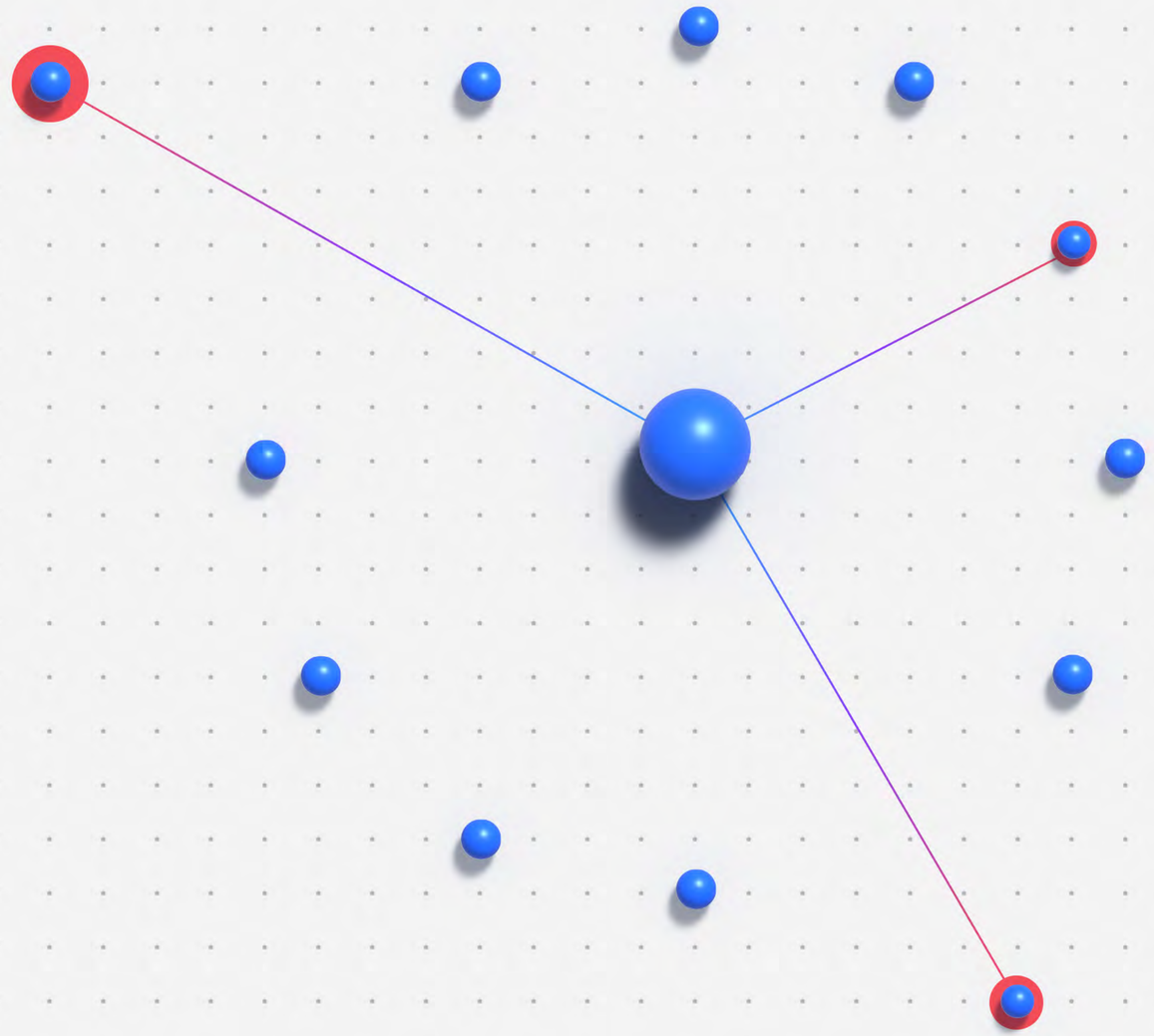
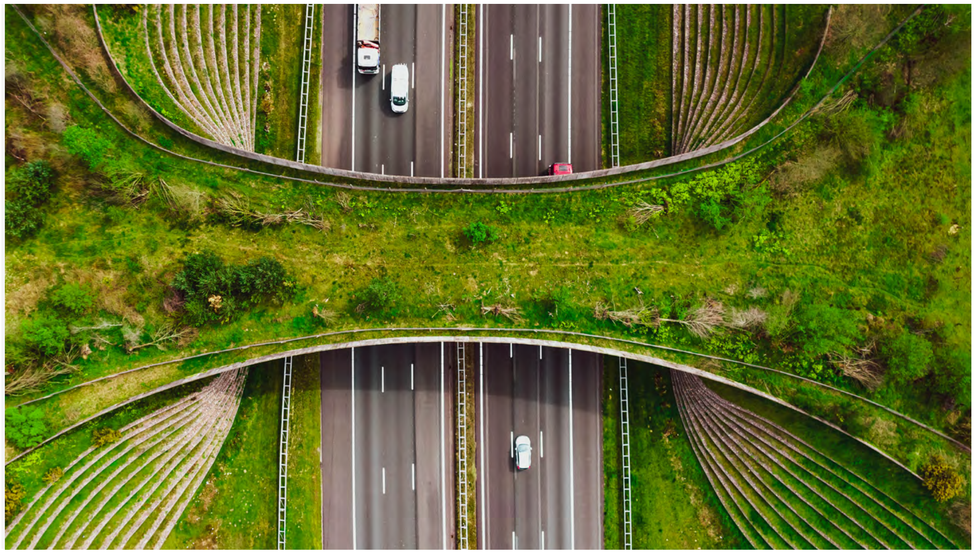
Cybersecurity in the era of generative AI

Learn how today’s security landscape is changing



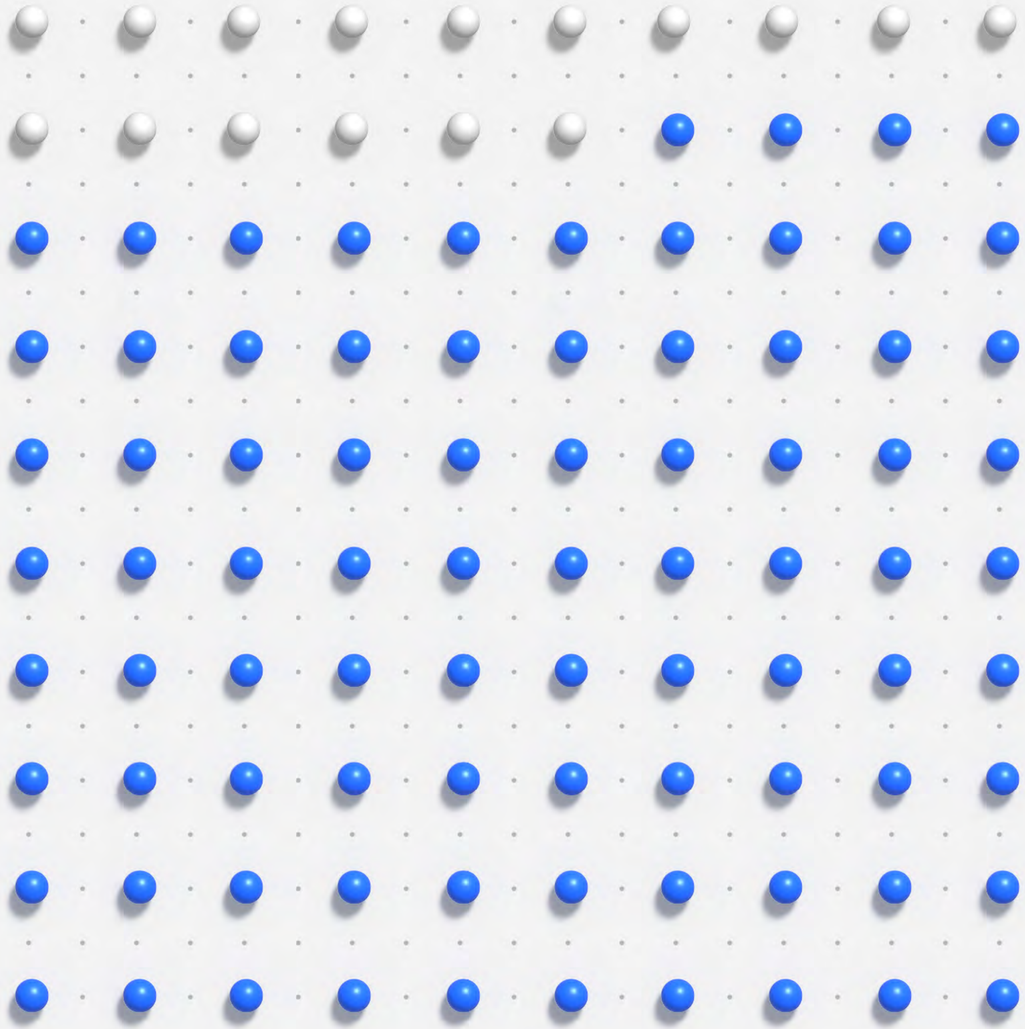
IBM

Contents



84%

of executives plan
to prioritize generative
AI cybersecurity solutions.[1]



01

Introduction

Cybersecurity leaders are facing a generative AI challenge. As their organizations experiment with this transformative technology—which can drive massive productivity gains across the enterprise—they must contain the potential risks and threats that generative AI can bring with it. These risks and threats include everything from accidental data leakage to hackers manipulating the AI to perform malicious tasks.

Some 48% of executives expect nearly half of their staff to use generative AI to augment their daily tasks in the next year.[2] Yet nearly all business leaders (96%) say that adopting this technology makes a security breach in their organization likely within the next 3 years.[1]

With the average cost of a data breach reaching USD 4.45 million globally last year—USD 9.48 million in the US—companies need to reduce risks, not increase them.[3]

Compounding the challenge, hackers are expected to adopt generative AI for the same speed, scale and sophistication it offers enterprises. With it, hackers can create better targeted phishing emails, mimic trusted users' voices, create malware and steal data.

Fortunately, cybersecurity leaders who've already invested in traditional AI solutions such as machine learning (ML) to maximize their time and talent can use AI to fight back. They can use generative AI tools to secure data and users and detect and thwart potential attacks.

There's a lot at stake. We created this guide to help you navigate the challenges and tap into the resilience of generative AI. We explore the ways attackers might use generative AI against you and how you can better protect yourself by using these technologies. Finally, we provide a framework to help you secure AI training data, models and applications across your enterprise.

02

How cybercriminals will use generative AI

Generative AI will likely benefit attackers in the same way it benefits enterprises by providing them with speed, scale, precision and sophistication. It will also upskill newcomers who may lack technical expertise, lowering the bar so that even novice hackers could launch malicious phishing and malware campaigns on a global scale.

In preparing to meet these threats, you should consider the two primary avenues that cybersecurity researchers see for generative AI-related attacks.

Primary avenues for generative AI-related attacks

- Attacking your organization
- Attacking your AI

Attacking your organization

Using large language models (LLMs), cybercriminals can conduct more attacks faster, from email phishing campaigns to the creation of malware code. Though these threats aren't new, the speed and scale hackers can gain from offloading manual tasks to LLMs could overwhelm cybersecurity teams that are already facing an ongoing labor and skills shortage.[4]

AI-engineered phishing
Researchers have shown that generative AI can be prompted to create realistic phishing emails within minutes.[5] They also found these emails to be nearly as effective as those created by someone who is experienced at social engineering phishing attacks. The emails are so convincing that they can challenge even the most well-prepared organizations.

- **More phishing, more clicks:** Though an AI-engineered phish aims to achieve the same goal as one crafted by a human, it becomes a tool that lets attackers speed up and multiply their phishing campaigns. This scenario creates a higher chance that users will mistakenly click a malicious email.
- **Targeted phishing:** Attackers can use generative AI chatbots to study their victims' online profiles, gaining valuable insights into their targets' lives. These chatbots can also generate highly persuasive phishing emails that mimic the targets' own style of language.

Deepfake audio
Cybersecurity leaders are concerned over the threat of generative AI audio deepfakes. In these attacks, criminals could feed recordings of a speaker's voice obtained online into an LLM that could generate whatever audio they want. For example, they could use a company CEO's voice to leave the CFO a message instructing them to pay a bogus invoice to an attacker-controlled bank account.

IBM has shown that attackers can "audio jack" calls in real time. An IBM researcher showed that attackers can do it by intercepting a live phone call, cloning the speaker's voice in real time and generating spoken sentences that could trick listeners into divulging sensitive information. Examples include bank account numbers and online passwords.

The term for this new method of attack is Phishing 3.0, in which what you hear seems legitimate but isn't. Cybersecurity professionals will likely need to combat this form of cyberattack with an array of security codes spoken between individuals to confirm identities. This method will add more layers of cybersecurity prompts to workers' lives.

Attacking your AI

Criminal hackers can try to attack enterprise AI models, in effect using your AI against you. They could “poison” your AI by injecting malicious training data into the AI models and forcing the AI to do what they want. For instance, they could cause your AI to make incorrect predictions about supply chains, or spew hatred over chatbots. They could “jailbreak” your LLMs by using language-based prompts to make them leak confidential financial information, create vulnerable software code and offer faulty cybersecurity response recommendations to your security analysts.

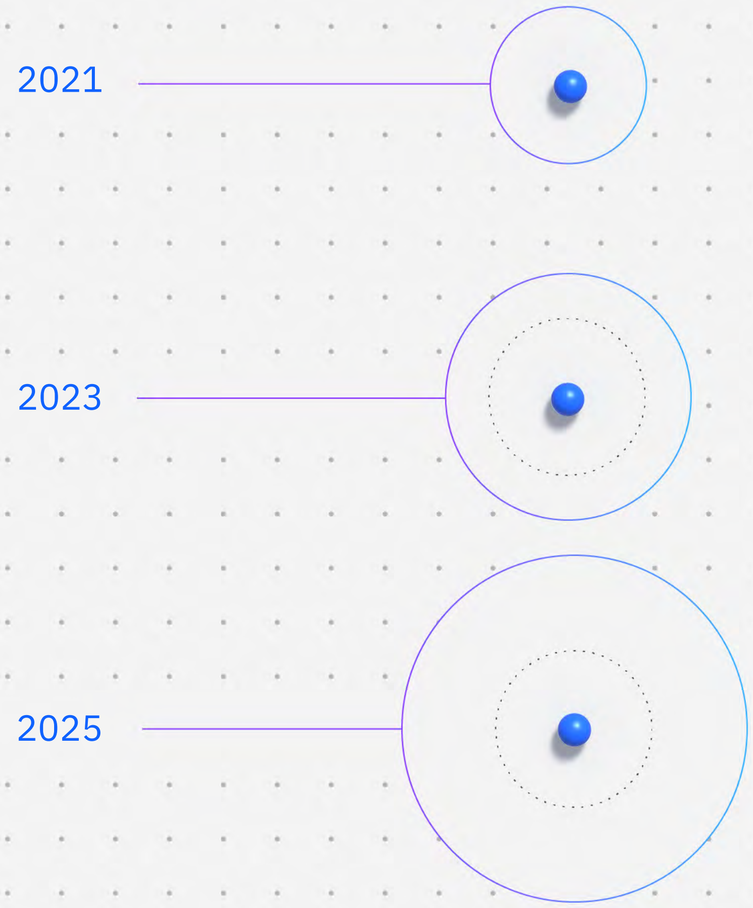
Poisoning your AI
Turning your AI against you may be the highest aspiration of cybercriminals. It can be tricky to pull off but not impossible. By poisoning the data used to train an LLM, an attacker can make it malfunction or behave maliciously—without being detected. The impact of a successful attack could range from creating disinformation to launching a cyberattack on critical infrastructure. These attacks, however, require a hacker to have access to training data. If that data is closed, trusted and secured, these attacks can be difficult to execute. But if your AI models are trained on open-source data sets, then the bar to poisoning your AI is much lower.

Jailbreaking your LLMs
With LLMs, English has essentially become a programming language. Rather than mastering traditional programming languages, such as Python and Java, to create malware to damage computing systems, attackers can use natural language prompts to command an LLM to do what they want. Even with guardrails in place, attackers can use these prompts to bypass or jailbreak safety and moderation features on your AI model. To better understand how attackers can manipulate LLMs, IBM® X-Force® researchers[6] successfully prompted several LLMs to:

- Leak confidential financial information about other users
- Create vulnerable and malicious code
- Offer poor security recommendations

51%

Executives say their 2023 AI cybersecurity budgets are 51% greater than they were in 2021. And they expect those budgets to climb an additional 43% by 2025.[1]



03

How AI maximizes your cybersecurity time and talent

Along with external challenges, cybersecurity leaders face many internal challenges. There are too few people to fill the vacant roles across their profession—700,000 by one government estimate.[4] Add to this situation an explosion of sensitive data, growing infrastructure complexity and an expanding attack surface. All these issues make it harder for cybersecurity leaders and their teams to safeguard data, manage user access and respond to the thousands of threats they face each day.

Every day, nearly a third (32%) of security operations center (SOC) analysts' time is spent investigating and validating incidents that turn out to be false threats.[7] In fact, there are so many alerts that SOC team members are only getting to half (49%) of the ones that they're supposed to review in a typical day.[7] That effort is not only demotivating to workers, but it can lead to critical cybersecurity bottlenecks. In fact, most analysts (81%) say they're slowed down by the daily effort of manual investigation.[7]

Most recently, cybersecurity leaders have been addressing these challenges with traditional AI solutions that use ML to help SOC analysts assess risks and recommend responses based on their organization's own security needs. The benefits of this approach are well established:

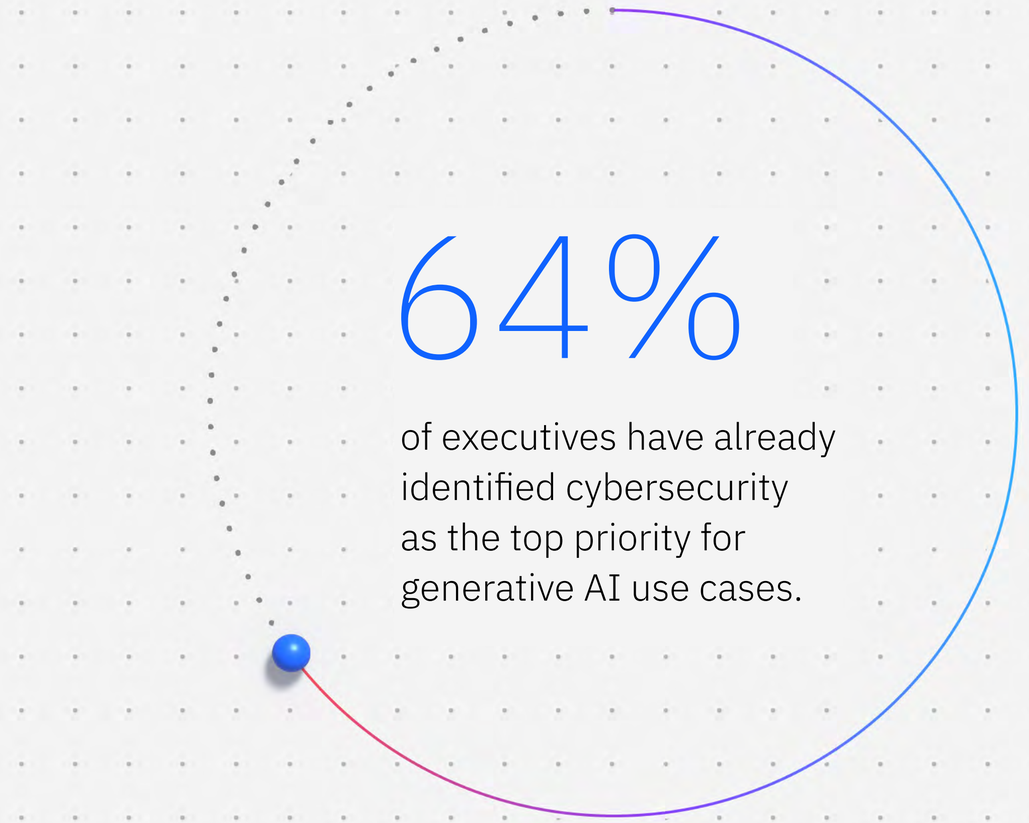
- Organizations that use AI and automation extensively experienced, on average, a 108-day shorter time to identify and contain a breach.[3]
- Leading AI adopters are monitoring 95% of network communications and reducing detection times by one-third.[8]
- Organizations that use AI and automation extensively saved nearly USD 1.8 million.[3]

Generative AI as a force multiplier in cybersecurity

In 2023, business leaders woke up to the potential of generative AI across their organizations. It began to transform nearly every corner of the enterprise, from supply chain management—by way of predictive analytics—to customer and employee experiences ... hello chatbots. But there's one corner it hasn't touched: cybersecurity.

The next 12 months will change all that.

According to an IBM Institute for Business Value study on generative AI in cybersecurity, 64% of executives have already identified cybersecurity as the top priority for generative AI use cases.[9] Moreover, a majority understand the business value of these technologies, with 84% stating they plan to prioritize generative AI cybersecurity solutions over conventional cybersecurity solutions.[1] For their part, cybersecurity leaders should educate and create buy-in for these solutions among business partners and C-suite executives.



Generative AI, when applied to cybersecurity, will be as much of a business accelerator as it has been for the rest of the enterprise—maximizing your security team's time and talent on multiple fronts. Generative AI can manage and automate repetitive, time-consuming tasks on behalf of analysts, freeing them to focus on more strategic aspects of cybersecurity.

It can also level up security professionals' abilities by simplifying complex and technical content, making it faster and easier for them to take on more challenging tasks.

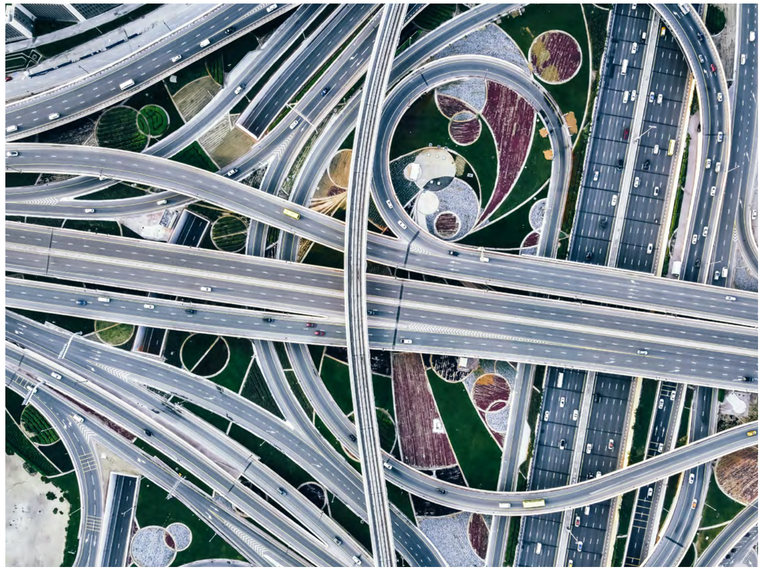
Some specific use cases for you to consider when looking for generative AI solutions to enhance your security operations in the near term include those that:

- **Automate reporting:** These tools can create simple summaries of security cases and incidents that can then be shared with various security and business leaders in the organization. They can be tailored to their level of technical expertise and areas of interest.
- **Accelerate threat hunting:** These tools automatically generate searches to detect threats—all based on natural language descriptions of attack behaviour and patterns—helping speed response to new threats.
- **Interpret machine-generated data:** To help analysts quickly understand security log data, these solutions provide simple explanations of events that have taken place on your system, expediting their investigations and lowering the technical barrier for new workers.
- **Curate threat intelligence:** Generative AI tools can interpret and summarize the most relevant threat intelligence, honing in on campaigns that are most likely to affect clients based on their unique risk profile.

Looking forward, generative AI will also be able to learn and create active responses that optimize over time, for instance, helping security teams find all similar security incidents, update all affected systems and patch all vulnerable software code.

04

Securing AI: Risks and recommendations



The speed and ease that LLMs offer enterprise users are major cybersecurity risks when it comes to wide-scale adoption. As engineers and analysts use generative AI to help create code and develop software scripts, they increase the risk that the AI may not be trained on secure code and practices.

Potential impacts to your organization

- Hasty data security practices
- Software supply chain risk
- Hallucinations
- Data leakage
- The black box effect

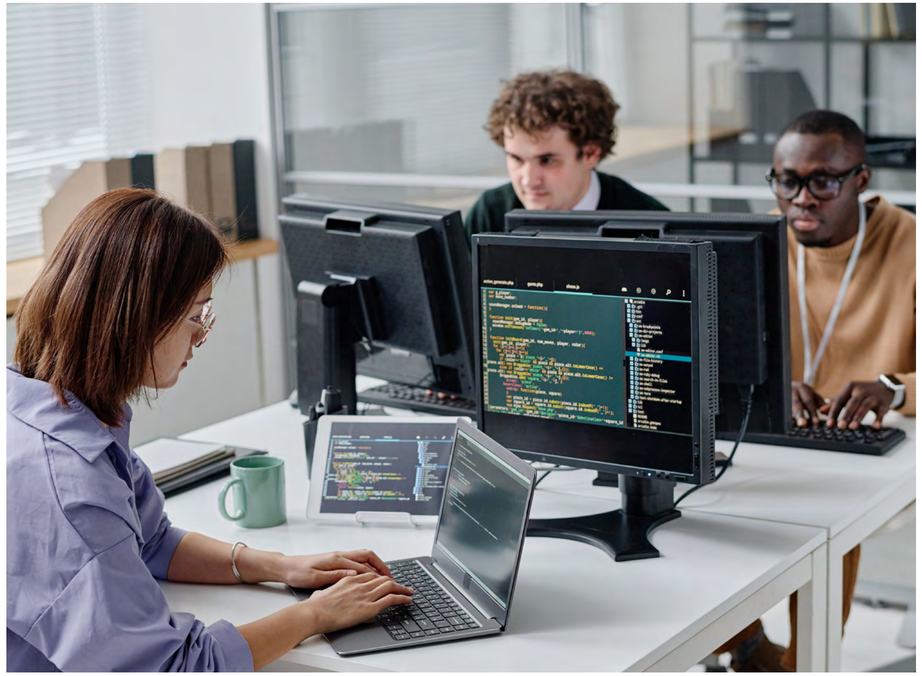
Hasty data security practices
Businesses are rushing to use LLMs, sometimes ignoring data security best practices and standards, including encryption, residency and privileged access controls.

Software supply chain risk
Businesses might be building and innovating on erroneous code assembled through open-source and commercial software components bought from vendors. Common failures and potentially exploitable software flaws across industries could create a new level of exposure.

Hallucinations
Software that's built using LLM-generated code carries the risk of including errors or hallucinations, potentially compromising the integrity and security of the source code. To prevent this issue, your organization should use AI models trained on diverse, balanced and well-structured data. Your organization's AI engineers should be as specific as possible in their prompts to avoid the AI model having to make assumptions or create detail where it's lacking. And they should be trained to rigorously assess the code generated by an LLM to ensure the quality of the output.

Data leakage
Without proper oversight and governance of third-party AI engines, organizations risk exposing confidential data to unauthorized users, regardless of the AI solution's policy on data segmentation or retention. New prompt injection and prompt leakage attacks can expose sensitive information and erode model performance.

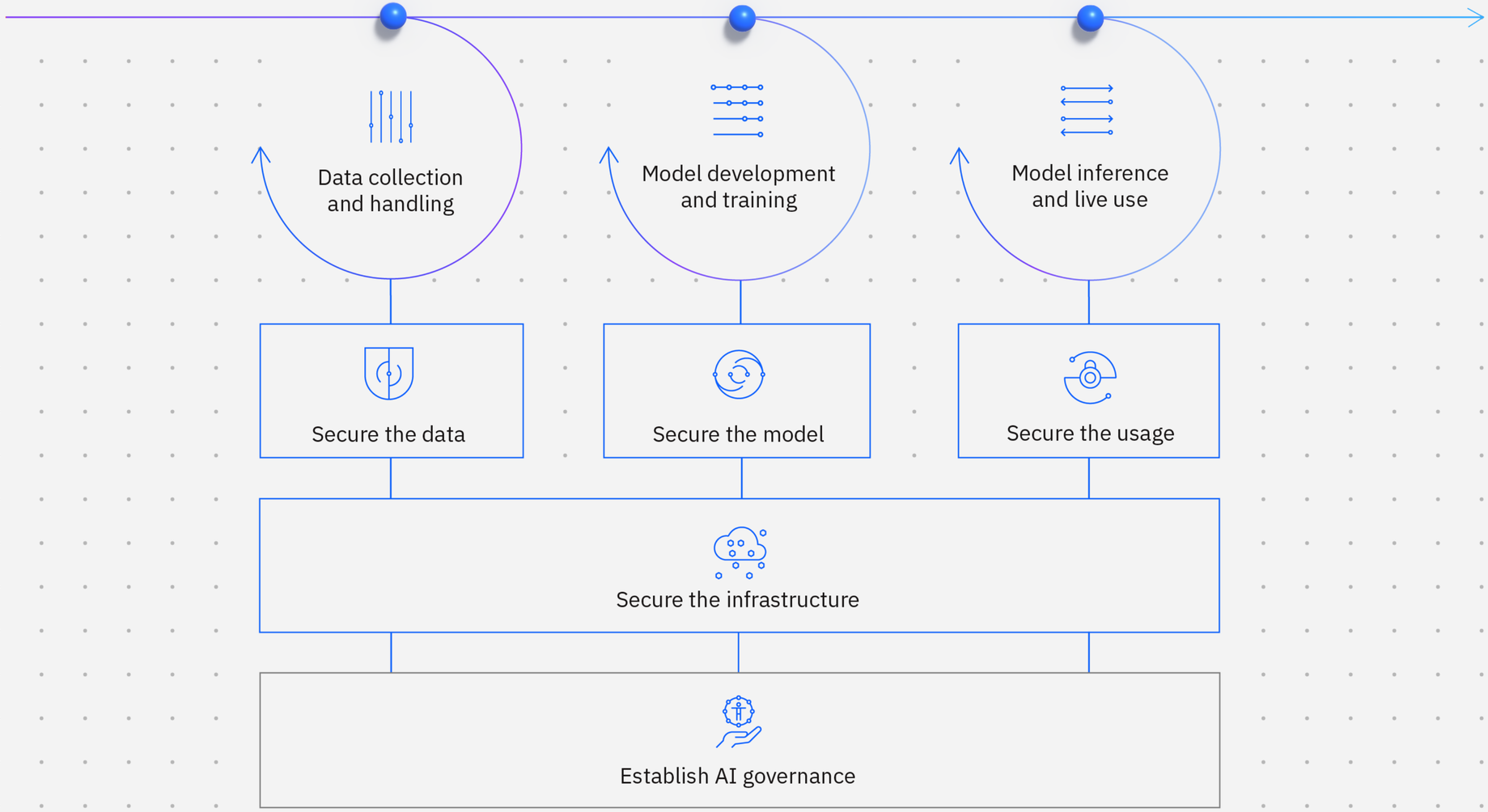
The black box effect
Part of the processing of AI occurs in a black box, meaning that security leaders will lack the visibility, transparency and explainability required of them and their teams. This issue poses a particular risk when proper software engineering practices aren't followed. It's critical that your organization have an AI governance approach that helps it track performance over a model's lifecycle and explain how and why a model produces the results it does.



Security for AI framework

Build trustworthy AI

As AI adoption scales and innovations evolve, cybersecurity guidance will mature. A framework for protecting trusted foundation models, generative AI and the data sets on which it's built will be essential for enterprise-ready AI. Here are some best governance and technical practices that you should share with your security teams.



Secure the data
Protect AI training data from theft, manipulation and compliance violations.

- Use data discovery and classification to detect sensitive data used in training or fine-tuning.
- Implement data security controls across encryption, access management and compliance monitoring.
- Use data loss prevention techniques to prevent sensitive personal information (SPI), personally identifiable information (PII) and regulated data leakage through prompts and application programming interfaces (APIs).

Secure the model
Secure AI model development by scanning for vulnerabilities in the pipeline, hardening integrations, and enforcing policies and access.

- Continuously scan for vulnerabilities, malware and corruption across the AI and ML pipeline.
- Discover and harden API and plug-in integrations to third-party models.
- Configure and enforce policies, controls and role-based access control (RBAC) around ML models, artifacts and data sets.

Secure the usage
Secure the usage of AI models by detecting data or prompt leakage and alerting on evasion, poisoning, extraction or inference attacks.

- Monitor for malicious inputs, such as prompt injections and outputs containing sensitive data or inappropriate content.
- Implement AI security solutions that can detect and respond to AI-specific attacks, such as data poisoning, model evasion and model extraction.
- Develop response playbooks to deny access, and quarantine and disconnect compromised models.

Secure the infrastructure
Extend your existing cybersecurity policies and solutions—including threat detection and response, data security, and identity fraud and device management—across your underlying AI infrastructure.

- Deploy infrastructure security controls as a first line of defense against adversarial access to AI.
- Use existing expertise to optimize security, privacy and compliance standards across distributed environments.
- Harden network security, access control, data encryption, and intrusion detection and prevention around AI environments.
- Invest in new security defenses specifically designed to protect AI.

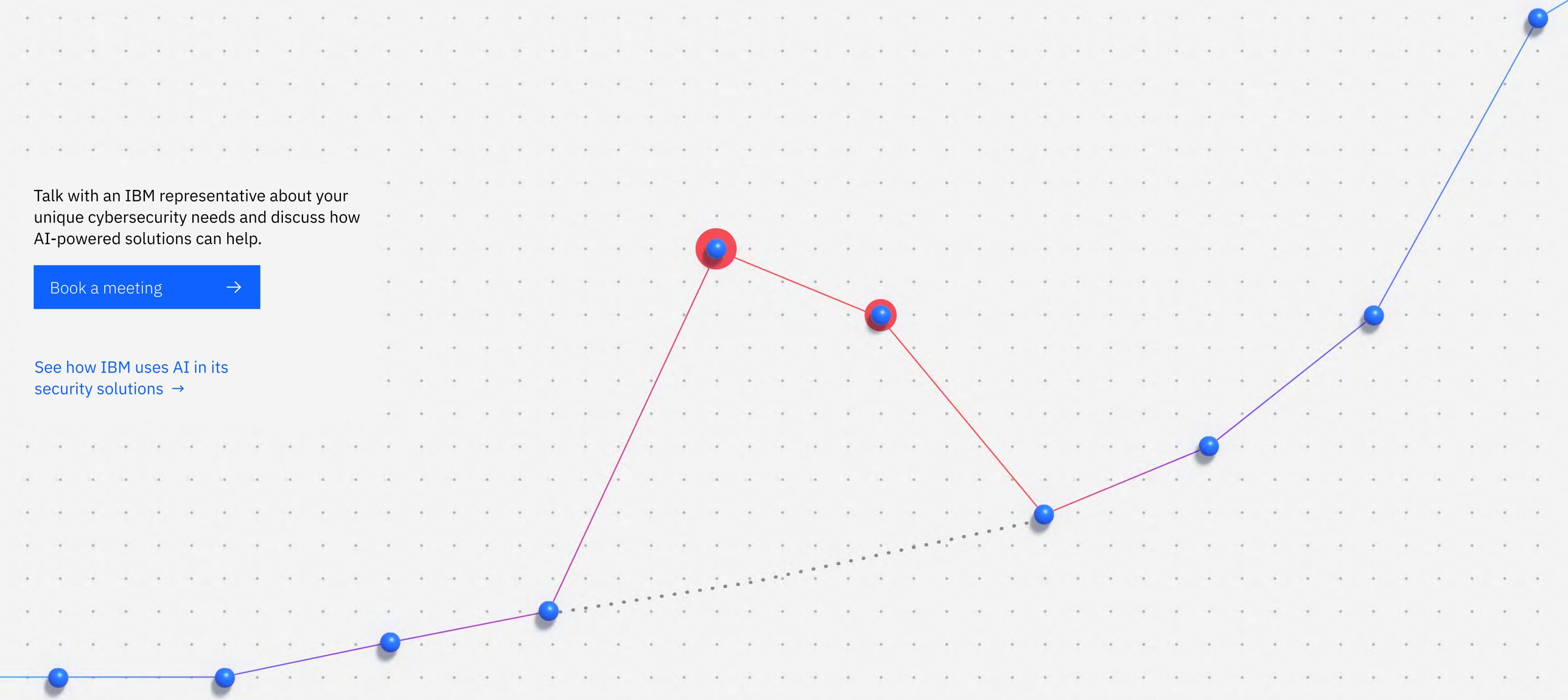
Establish AI governance
Building or buying trustworthy AI requires an AI governance framework that helps you direct, manage and monitor your organization's AI activities. The framework will strengthen your ability to mitigate risk, manage regulatory requirements and address ethical concerns, regardless of your existing data science platform.

- Enable responsible, explainable, high-quality and trustworthy AI models, and automatically document model lineage and metadata.
- Monitor for fairness, bias and drift to detect the need for model retraining.
- Use protections and validation to help enable models that are fair, transparent and compliant.
- Document model facts automatically in support of audits.
- Automate and consolidate multiple tools, applications and platforms while documenting the origin of data sets, models, associated metadata and pipelines.

Talk with an IBM representative about your unique cybersecurity needs and discuss how AI-powered solutions can help.

Book a meeting →

See how IBM uses AI in its security solutions →



1. When it comes to cybersecurity, fight fire with fire, IBM Institute for Business Value, 2023.

2. IBM Consulting unveils Center of Excellence for generative AI, IBM blog, 25 May 2023.

3. Cost of a Data Breach Report 2023, IBM Security, July 2023.

4. Media Advisory: Garbarino Announces Hearing on Growing the National Cybersecurity Talent Pipeline, Homeland Security Committee | Republican, 16 June 2023.

5. AI vs. human deceit: Unravelling the new age of phishing tactics, IBM Security Intelligence, 24 October 2023.

6. Unmasking hypnotized AI: The hidden risks of large language models, IBM Security Intelligence, 8 August 2023.

7. Global Security Operations Center Study Results, a Morning Consult study commissioned by IBM, March 2023.

8. AI and automation for cybersecurity, IBM Institute for Business Value, 3 June 2022.

9. Enterprise generative AI: State of the market, 2023 IBM Institute of Business Value Study.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
December 2023

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

All client examples cited or described are presented as illustrations of the manner in which some clients have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

Generally expected results cannot be provided as each client's results will depend entirely on the client's systems and services ordered. It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs. THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

Statement of Good Security Practices: No IT system or product should be considered completely secure, and no single product, service or security measure can be completely effective in preventing improper use or access. IBM does not warrant that any systems, products or services are immune from, or will make your enterprise immune from, the malicious or illegal conduct of any party.

The client is responsible for ensuring compliance with all applicable laws and regulations. IBM does not provide legal advice nor represent or warrant that its services or products will ensure that the client is compliant with any law or regulation.