LEARNING MADE EASY

Palo Alto Networks Special Edition

AI Security

for dummies®

A Wiley Brand

Understand the promises and risks of AI

Build secure AI apps and agents by design

Embrace unified, platform-based secure AI

Brought to you by

paloalto® NETWORKS

Steve Kaelble

About Palo Alto Networks

Palo Alto Networks is the global cybersecurity leader, committed to making each day safer than the one before with industry-leading, AI-powered solutions in network security, cloud security, and security operations. Powered by Precision AI, our technologies deliver precise threat detection and swift response, minimizing false positives and enhancing security effectiveness. Our platformization approach integrates diverse security solutions into a unified, scalable platform, streamlining management and providing operational efficiencies with comprehensive protection. From defending network perimeters to safeguarding cloud environments and ensuring rapid incident response, Palo Alto Networks empowers business to achieve Zero Trust security and confidently embrace digital transformation in an ever-evolving threat landscape. This unwavering commitment to security and innovation makes us the cybersecurity partner of choice. For more information, visit www.paloaltonetworks.com.

AI Security

Palo Alto Networks Special Edition

by Steve Kaelble

for dummies®
A Wiley Brand

AI Security For Dummies®, Palo Alto Networks Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-394-35737-6 (pbk); ISBN 978-1-394-35738-3 (ebk);
ISBN 978-1-394-35739-0 (ebk)

Publisher's Acknowledgments

Editor: Elizabeth Kuball
Acquisitions Editor: Traci Martin
Senior Managing Editor: Rev Mengle
Client Account Manager: Cynthia Tweed
Content Refinement Specialist: Umeshkumar Rajasekhar
Technical Contributors: Hitendar Sethi, Munjal Munshi, Monique Lance, Paul Carlstrom, Ranae Sandholm

Table of Contents

Introduction

The momentum has been building for a while now, but the world has most definitely blown past the tipping point when it comes to artificial intelligence (AI). AI plays an increasing role in wide-ranging tasks, boosting productivity, delivering a competitive edge, and taking on tasks that only science-fiction authors could have envisioned not that long ago.

Also blooming in this changing landscape are lots of new risks, many with the ability to totally evade traditional information technology (IT) security measures. Some risks arrive as employees innovate on their own through third-party AI apps, some are inherent even in enterprise-built AI applications. Some risks grow as employees tap into AI in ways they don't even realize are unsafe. There are risks to data, to operations, to compliance, to reputations, and to corporate bottom lines.

The risks are multiplied exponentially by the very attributes that make AI amazing. AI does things stunningly well and efficiently, scaling up without complaint — but it can do the wrong things quite efficiently and at massive scale, too. Add in AI agents with permissions to take concrete actions, and the potential problems get even more dire.

There are security solutions out there, but in the world of AI, point products and built-in safeguards serve only to create a patchwork security blanket with lots of holes. Your organization needs a unified, platform approach that governs human and AI interactions, that builds security into AI instead of just bolting it on, and that safely controls autonomous workflows. Your solution must protect the whole lifecycle, including a constant watch at runtime, when a program is executing.

About the Book

AI Security For Dummies, Palo Alto Networks Special Edition, is your guide and road map to this brave and sometimes frightening new world that isn't just around the corner, but already on every corner. Brought to you by, and packed with insights from, industry

leaders in IT security, this book outlines where things stand now, where they're headed, what the risks are, and what your enterprise can do about them.

Read on for more on the threat landscape, the risks to sensitive data and intellectual property, the promises and dangers of third-party AI apps, and the possibilities of internal AI application development. Learn concrete actions you can take to discover AI activity, build governance and guardrails, protect data and control access, test thoroughly and monitor continuously, and when necessary, respond effectively.

Foolish Assumptions

In writing this book, I made a few assumptions about you:

- Your work is in IT, maybe as chief information officer (CIO), chief information security officer (CISO), or cloud security architect — or you may be on the business side with both eagerness and worry about AI and your enterprise.
- You know your enterprise needs to fully embrace AI opportunities, but you know it has to happen safely.
- You'd appreciate a high-level overview of the landscape and the challenges, with concrete and actionable tips for succeeding.

Icons Used in This Book

Keep an eye out for iconic guideposts for your reading:

REMEMBER

You've got 64 pages of insights here, but if your time is short, be sure not to miss the paragraphs marked with this icon.

TIP

I mention offering actionable insights, and the Tip icon points to some of these insights.

WARNING

IT security aims to react smartly to warnings, and this icon points out some concerning things to be mindful about.

Beyond the Book

After you get through the book, you'll have lots of answers but likely plenty of new questions and an appetite for more information. Check into the AI security resources and solutions at `www.deploybravely.com`.

IN THIS CHAPTER

- Watching AI grow and grow and grow
- Developing your own enterprise AI
- Understanding the new AI threats
- Exploring AI from a business risk perspective

Chapter 1
Embracing AI: Opportunities and Threats

The fact that you opened this book means you're well aware, the sky's the limit when it comes to artificial intelligence (AI). There's not a growth-minded company in the world that isn't figuring out how AI fits into the future of the enterprise. This chapter explores just how explosive the growth is, both in terms of accessing generative AI (GenAI) apps and building enterprise AI systems. It outlines some of the new threats facing this brave new world and details why those on the business side need to approach AI with a cautious eye.

Measuring the Growth in AI

From the point of view of many people, the age of AI really took off toward the end of 2022, with the public launch of ChatGPT. Measured in the number of headlines and whatever way you might measure the collective imagination, that seems reasonable. Indeed, some metrics have tallied a billion new users of AI apps in the course of a couple of years.

REMEMBER

Truth be told, though, AI has been working its way into our daily lives for quite some time — since well before the current explosion in attention. Recognizing that fact is a key to understanding just how widespread the current usage is and future growth will be. AI does so much more than make travel plans, find recipes, help with homework, build successful résumés, or create clever social media memes.

It's not an overstatement to observe that AI is changing how we live and work, and how organizations evolve. Self-driving vehicles are already out there navigating, and that will one day include long-haul trucks. Smart grids are already balancing national energy flows. AI tutors are helping students excel. Musical artists and brands are creating AI-first digital experiences. Data-infused AI is making machine-speed decisions and taking immediate actions.

Here are a few stats backing up just how far AI is working its way into daily life, not just fun popular culture:

- A study cited by ZDNet has found nearly six in ten employees using public GenAI apps every week.
- Collaboration is enhanced, with 73 percent using AI to improve communications at work, according to Grammarly.
- A McKinsey study found that, compared to what their bosses believe, three times as many employees are using AI for a significant amount of their work.
- Gartner believes that 80 percent of software as a service (SaaS) apps will embed AI in the foreseeable future.
- Beyond that, a Pitchbook study expects that by 2030 there will be 12,000 native AI applications — built specifically for AI purposes instead of just adding in AI functionality.

The numbers make it clear that AI is finding its way into daily work more and more, and that employees are increasingly ready to embrace it.

REMEMBER

That's hardly surprising — the potential benefits are numerous and impressive. GenAI can speed up all kinds of time-consuming business processes, including automating multistep workflows, generating content from emails to marketing to product descriptions, and analyzing large data sets in a snap. GenAI can crank

out impressive reports and summarize lengthy documents. For software developers, it can turn out code, debug automatically, test scripts, and really speed up development cycles.

As far as customer experiences go, chatbots and virtual assistants can provide powerful, polite, and tireless support. AI can come up with impressively personalized product recommendations. And when humans are still in the loop serving customers, AI can make them a whole lot more productive and speedy at problem-solving.

Over on the research and development side, AI can analyze market trends, diving into unstructured data to spot trends and inform strategic decisions. It can assist with new product design and help bring new opportunities to market quickly.

The folks in risk management and compliance may really be enamored with the way AI can handle predictive risk identification, by comparing historical patterns with current behaviors. That's great for spotting everything from supply chain troubles to emerging cyberthreats to new types of fraud. AI can help monitor compliance and generate reports, too.

GenAI can help identify employee skill gaps and then it can help with the training needed to fill those gaps. It's potentially great for predicting staffing needs.

And across the organization, you'll find people intrigued by the possibilities of agentic AI, through which AI agents chain together actions supporting a workflow. It's miles beyond traditional, rules-based automation because the agents act autonomously with adaptability, making decisions independently and aiming for a higher-level goal.

What's not to like? Yeah, you knew there was going to be a *but* to this conversation.

WARNING

As the volume of AI adoption grows, so do the data incidents that are linked to AI use. Just in the first quarter of 2025, data loss prevention (DLP) incidents were up by about two and a half times. As the new year unfolded, GenAI-related data incidents were under 5 percent of the total, and by about the end of the quarter, the share had hit 15 percent.

It isn't all a matter of malicious actors stirring up trouble, either — much of the problem isn't even intentional. Employees

using GenAI apps and undetected GenAI plug-ins may be sharing confidential data without even knowing it. For example, uploaded data can end up training AI models, which can result in an unintentional data leak and may compromise proprietary information or intellectual property.

The potential loss of sensitive data is a frightening enough problem, but that's only the beginning of the peril that's resulting from the adoption and use of GenAI and the growth in third-party GenAI traffic. Every new app that enters the picture is a potential new breach point. The attack surface gets bigger and bigger the more employees tap into Gen AI.

Many tools employees use are integrated through browser extensions or plug-ins, or through application programming interfaces (APIs). Those are all potential new vulnerabilities. Insecure apps could open the door to threats such as credential harvesting or the injection of malicious payloads.

WARNING

One more angle to worry about is compliance. Anytime data is at risk, your enterprise is at risk of compliance or regulatory violations. If there is data leakage involving protected health information (PHI) or personally identifiable information (PII), for example, you could run afoul of a whole alphabet full of governmental mandates, from HIPAA to GDPR to SOX. Even if data stays where it should stay, it may be harder to verify compliance when AI is in the picture.

All of that means that reputations are at risk. Breaches, or even the risk of breaches, can seriously damage organizational trust and brand credibility, cause brand erosion, and ultimately steer business to the competition.

Organizations no doubt put in a lot of effort trying to maintain visibility and control over information technology (IT) assets, but so-called "shadow AI" usage can get around that visibility and control. That makes it much harder to mitigate risks or respond to incidents.

And finally, although GenAI works a lot of pretty amazing magic, most people have experienced or at least read about troubles with model bias or errant outputs. The more employees tap into third-party GenAI, the greater the chances that they'll end up connecting with unvetted models generating content that may not be helpful.

Building AI into the Enterprise

You would think that much of the revolution and the risk relates to GenAI plug-ins and apps that employees may be tapping into in order to rev up their work capabilities. That's only part of the picture. Development of enterprise AI applications is also skyrocketing.

It wasn't long ago that this situation may have been seen as an emerging trend, marked by experiments and proof-of-concepts. Suffice it to say that the trends aren't just emerging — they're fully emerged and, increasingly, business as usual. Experiments are now production deployments, and AI is increasingly built into core workflows.

REMEMBER

For example, nearly three-quarters of European companies are benefiting from AI-assisted customer support already, according to *Digitalisation World.* AI can even make ordering a burger more efficient — a report in Google Cloud Blog states that three-quarters of Wendy's drive-through orders are automated.

When it comes to research and product development, AI is proving its worth quickly. For example, according to McKinsey research, AI can triple the speed of drug molecule design, making new drugs get to market faster.

Clinical decision support and completion of charting are increasingly essential uses in health care. AI scribes are listening in exam rooms, taking notes and cutting down on what doctors unlovingly call their "pajama time" of charting while at home. AI is also spotting abnormalities in imaging scans and helping health systems navigate time-consuming prior authorizations required by insurance.

State governments are building agency action plans based on new legislation. Consultants are creating analyst research reports must faster with AI.

Productivity and automation are, thus, key drivers for enterprise AI adoption. Being able to better benefit from structured and unstructured data at scale is intriguing and valuable. AI is allowing better corporate decision making, and agentic systems are opening all new ways to boost both employee productivity and customer service.

But this shiny coin has a darker flip side. The more AI works its way into enterprise application development, the more it brings along its scary new world of risks.

More than 60 percent of enterprises don't have great visibility into machine learning (ML) data sets. Malicious content and privacy violations are also risks; large language model (LLM) data corruption can lead to ML model poisoning.

WARNING

The Open Worldwide Application Security Project (OWASP) is always on the lookout for IT risks and lists these new attack types as top LLM risks:

- Prompt injection
- Sensitive information disclosure
- Supply chain vulnerabilities
- Data and model poisoning
- Improper output handling
- Excessive agency
- System prompt leakage
- Vector and embedding weaknesses
- Misinformation
- Unbounded consumption

It's important to note that enterprise AI applications connect with an extended technology ecosystem. They're not off in an AI silo — they're well integrated all over the technology landscape. Users may access AI functionality through web or mobile front ends, apps connect through APIs, and AI agents act autonomously, accessing and changing data.

The AI infrastructure spans virtual machines (VMs), containers, and serverless landscapes. It taps into ML libraries and makes use of code and packages, as well as drivers and optimization tools. It all runs on the computing power of central processing units (CPUs) and graphics processing units (GPUs).

WARNING

All of this exposes enterprise AI applications to a variety of development and runtime threats. The infrastructure could face supply chain attacks or misconfiguration issues during development, access and authentication risks, remote code and malware

execution, unpatched vulnerabilities, and lack of segmentation at runtime.

Enterprise AI applications live and breathe data sets. They may be trained on internal data, as well as data from external sources. And as AI engages in inference, it will access vector databases that may be managed or self-hosted.

Development threats to these data sets include poisoned data and the potential leak of sensitive data. At runtime, there could be unrestricted database queries, as well as insecure outputs. And a publicly writable data set is a significant risk.

Plug-ins and agents are key elements of the enterprise AI ecosystem as well, which may need translation or search plug-ins, SQL generation, file ops, and an assortment of custom agents. These elements could suffer from insecure output handling in development as well as unrestricted communications at runtime.

There may be a variety of external AI models integrated into the enterprise AI ecosystem, such as ChatGPT, Claude, Gemini, Llama, or any number of others. Bringing external connections into the picture adds such development threats as model theft and data poisoning, and runtime concerns such as prompt injection and model denial of service (DoS) attacks.

AI agents further expand the app architecture, bringing a variety of actions into the picture and employing short-term and long-term memory. They can connect with such things as payment processors, calendar schedulers, ticket resolvers, and email senders.

Agents create tremendously exciting possibilities, and lots of new risks. Identity spoofing and authentication issues can plague web connections, and apps are vulnerable to malware execution and unsafe URL processing. Unrestricted database queries are possible, which means potentially sensitive data exposure and publicly writable data sets can be imperiled. Goal hijacking is a real threat when it comes to AI agents, as well as cascading hallucinations. Privileges can be compromised and tools misused.

Exploring the Threat Landscape

Those who have been around the business of IT security long enough are well aware that everything always changes. Each new advancement brings a new set of risks and threats and ways that bad actors can do their dirty work. Every time you deal with one new risk, another emerges, like a high-stakes game of Whac-A-Mole. There is no room for rest or complacency when the attack surface keeps growing.

WARNING

The fast-growing AI ecosystem is creating a whole new landscape of threats, some that seem familiar and some that are truly new frontiers specific to AI itself. Here are just some of the threats that are taking root in this landscape (and you can be sure that in a few months or a year, something new will come along to add to this list).

Prompt injection

Prompt injection is an attack on AI model inputs, typically known as *prompts.* The aim is to make the model do things it wasn't intended to do, such as reveal secrets, change its behavior, or perform actions beyond the scope of its creators' intentions.

In a direct injection, a malicious actor puts bad instructions right into a prompt. It may tell the model to ignore its previous instructions and release a set of confidential data.

In an indirect injection, similarly evil instructions are hidden in external data sources, such as database records or PDFs or web pages that the AI opens and reads.

Either way, sensitive data may end up seeing the light of day, or API keys or other credentials. The LLM may end up taking an errant action, such as sending fraudulent emails or calling APIs. Multistep AI agent workflows can be steered in the wrong direction through prompt injection.

Data poisoning

Most people who are in the know about AI already know that outputs can, on some occasions, be unintentionally biased or downright incorrect. Hallucinations happen, even inadvertently. But

data poisoning is like intentionally spiking the model's drink with some sort of hallucinogenic substance.

When bad actors feed malicious, biased, or otherwise corrupted data into the model's training or fine-tuning, the model learns incorrect facts or makes the wrong associations or picks up unsafe behaviors. The same can happen if poisoning is achieved through retrieval-augmented generation (RAG) pipelines feeding data in from live sources. Either way, data poisoning can lead to harmful outputs and undermine trust in AI-driven actions and decisions, just when you're trying to get users to trust your apps.


REMEMBER

The problem may not happen all at once, either. Models can degrade over time as they're retrained, letting biased or manipulated data work its way in gradually. Even a good model is potentially at risk for this kind of drift.

Model inversion

For the average user, it's impossible to know all the information that has gone into an AI model's training, and as long as the output is accurate, most people don't really care or need to know. But there may be some valuable nuggets of information hidden among all the training data, and model inversion is one way to try to mine that gold.


WARNING

Model inversion is when attackers query a model again and again and again, in ways that aim to reconstruct training data or make inferences about it. The result may be digging out proprietary data such as source code, customer information, or perhaps details from medical records that were anonymized for training. It's called *inversion* because it's starting from the end and trying to work back to the beginning.

Even in the absence of an organized attempt at model inversion, it's possible for data leakage to happen. In some cases, snippets of sensitive training data can peek through in responses, and whoever's data was exposed is not going to appreciate the leak.

Threats to agentic AI

Agentic AI systems are built to take actions independently and execute workflows. These are systems that follow multistep processes autonomously, calling APIs and accessing data.

That makes them vulnerable to a variety of threats of their own. Goal hijacking, for example, happens when an attacker manipulates intermediate steps to steer the agentic system toward a malicious outcome. API chaining, where an agent's output becomes the trusted input for another agent, can cause trouble if that output is somehow manipulated.

WARNING

Also, if tools in an agentic system are over-permissioned, agents may end up with access to data or APIs that they shouldn't be allowed to access. And again, even unintentional blips can have consequences, such as if a model hallucination ends up changing the behavior of an automated workflow.

Membership inference

This type of AI-related attack aims to figure out whether a certain record was included when the model was trained. That may seem like a minor breach, not quite as dastardly as some, but make no mistake, the consequences can be major.

Consider, for example, if we're talking about data about a specific person. Just knowing who the customers are can be a breach of confidentiality — in fact, it can be a significant regulatory violation. For example, the Health Insurance Portability and Accountability Act (HIPAA) makes it illegal to simply reveal that a person is a patient, and it levies some very steep fines when protected health information is breached. Running afoul of Europe's General Data Protection Regulation (GDPR) is equally unpleasant.

Adversarial attacks

When you're under attack, it sure helps to know that there's an adversary out there and who it is. But some attackers are able to craft inputs that seem pretty normal to humans while fooling an AI model.

Little purposeful glitches of this kind may cause an LLM to output the wrong information. It could cause a predictive AI model to misclassify some data or misinterpret something. It could even jailbreak the safety mechanisms.

Supply chain risks

On one hand, enterprise AI development is the act of building your own applications from scratch, but on the other hand,

you're not likely building it entirely from scratch. It may depend on open-source models and data sets, perhaps Python libraries.

WARNING

That puts your application at risk for malicious code or model weights among its dependencies. Open data could be tainted. Attackers could deliberately sow dependency confusion by putting out malicious packages with names similar to what your developers were really looking for. And when something is open source, there's often some risk of legal exposure that you don't know about.

And who knows what else?

These are just some of the threats that are increasingly worrisome in an AI-driven ecosystem. There are more that we didn't list — and even more that we can't foresee as of now.

Imagine troubles that could be stirred up by malware that's prompt-aware and knows how to spot LLM interaction so it can dive in with malicious content and intent. Consider how much chaos could be caused if one compromised AI agent could start to wrongly influence others. What if AI-generated personas get really good at misinformation and fraud?

Go ahead and let your imagination go to dark places. Read up on some dystopian sci-fi. There's trouble around the corner, so it's best to be ready for it.

Understanding Business Risks

If you're an IT security specialist, you probably wake up from nightmares on a pretty regular basis. Integrating AI into your landscape no doubt adds some new plotlines to your nightmares, but you already were having trouble sleeping.

If you're on the business side, you may have more mixed feelings. You can't log in to your favorite industry trade publication without being bombarded with daily headlines about the promise of AI, the cool things your competitors are doing, the amazing possibilities. Your big fear? Missing out.

That fear is valid, of course, but you need to balance that fear against the very real business risks that AI mishaps can bring.

Let's examine a few of them, to be sure you take that plunge in a way that's not just starry-eyed, but also levelheaded.

WARNING

Consider insider threats. They're nothing new, but AI systems can turbocharge those risks because they open high-powered access to data as well as automated actions that may not be fully visible.

Let's imagine that confidential data finds its way into shadow AI apps, either intentionally or not. That security risk creates a boatload of business risks, such as regulatory violations, reputational damage, legal problems, and loss of intellectual property. Operational disruptions are quite possible, too.

Or think about weaponized AI applications. Remember that AI not only helps your business work better and faster — malicious actors can get a productivity boost of their own, at your expense. Phishing is a power sport with the help of AI. Synthetic identities facilitate fraud and access risk. AI-generated disinformation can target your reputation. Financial losses are a real business risk, along with competitive disadvantages. And you don't want regulators on your back.

Over-permissioned or unchained AI agents also bring financial and compliance exposures, the possibility of operational downtime, and the risk of stakeholder mistrust in important AI-driven operations.

Beyond that, poorly governed data can be risky to the performance of AI models, through potentially biased outputs, data leaks, loss of traceability, and other concerns. From a business perspective, this can bring regulatory fines and other financial losses. And a big-picture business worry is that users lose trust in AI. You need AI for a successful and competitive future, and you need stakeholders to trust it. The future of AI has to be built on a secure foundation of trust.

IN THIS CHAPTER

- Understanding and controlling employee use of third-party AI
- Gaining visibility into the AI app landscape
- Outlining best practices in AI security
- Developing solid policies and governance
- Seizing the Holy Grail of AI security

Chapter 2
Safeguarding Enterprise Data in the Age of AI

Seemingly everyone uses artificial intelligence (AI), and seemingly every enterprise, too. There are lots of apps that make it simple, with no computer expertise required, so inventive employees are upping their game at work by exploring them. What could go wrong?

This chapter dives into the many things that can go wrong. But more important, it outlines how to protect your business and get the use of third-party apps under prudent control, with careful monitoring and solid governance to achieve data security in the use of AI.

Protecting Use of Third-Party Apps

As shared in Chapter 1, the number of native AI apps is exploding. The conversational chatbots such as ChatGPT and Gemini have gotten a lot of attention, but there are hundreds or thousands more apps out there, with new third-party apps launched all the time.

Beyond conversational chats, code assistants and generators are making life easier for developers. Video and image generators are handling new creative needs. Writing assistants are helping expand the reach and efficiency of marketing teams and other practitioners of the written word. There are AI meeting assistants, productivity assistants, and enterprise search specialists. Plenty more are being built in-house (more on that in Chapter 3).

REMEMBER

Indeed, according to Palo Alto Networks "State of Generative AI" research, generative AI (GenAI) traffic surged by nearly 900 percent in 2024. New AI models and apps are developed and introduced in mere months, taking on all sorts of exciting new tasks. Hugging Face is home to more than two million models, and it records millions of downloads every month. No wonder that, on average, organizations that take a deep dive into what employees are up to find 66 GenAI apps in their environments.

That is a crazy sprawl of apps, many of them on the insecure side, and you can bet those are not all authorized or approved. Salesforce researchers found that more than half of employees have used unapproved GenAI tools at work. Shadow AI apps are creating security blind spots in a big way. Uninspected prompts and responses are leading to the loss of sensitive data. And malicious content is finding its way into some responses, posing all kinds of risks.

Needless to say, this kind of Wild West AI landscape can't continue. There are just far too many downsides accompanying the exciting upsides. It's critically important for organizations to secure GenAI usage, with purpose-built protections aimed at employee use of third-party GenAI apps.

TIP

That's easier said than done, but from a big-picture perspective, here are some of the problems, paired with the solutions every organization needs:

- **Unchecked sprawl of insecure apps:** A good start is having a comprehensive GenAI app catalog that can keep up with the explosion in new tools.
- **AI apps storing, learning, and reiterating data:** From that comprehensive list of apps in use, you need visibility into the ones that are training on data.

- **Ingestion of unstructured inputs:** This problem means any type of data can be provided. The answer is granular data controls that have context-aware detectors employing machine learning (ML). Your controls must be able to protect against threats that are delivered by way of responses and the links that they contain.
- **Rapidly evolving apps are easily accessible:** Powerful apps are all over the various marketplaces, so you need visibility into and control of third-party AI plug-ins. You have to be able to detect plug-ins that claim excessive permissions.

That's a big wish list, but the reassuring news is that it's not unrealistic to have these wishes fulfilled. Let's look into some insights and ideas for how to make it happen.

Keeping an Eye on AI Usage

As we take a deeper dive into the operational aspects of securing GenAI usage, it's all about gaining full visibility into all of the work being done with the help of AI, and then taking action to be sure it's happening safely.

As every security expert will tell you, you can't secure what you don't know about, which makes discovery the logical, necessary beginning.


TIP

That means going on a hunt for all AI tools, browser extensions, and other various kinds of plug-ins. You'll be checking everything from network traffic to browser activity to software as a service (SaaS) usage as you work the discovery step, which will provide a foundation for setting policy and establishing controls.

One multinational consulting firm went through this exercise and found that employees were regularly accessing nearly 400 GenAI apps, three of which were known to be high-risk. A large financial firm went through a manual process of discovery to look for GenAI apps, and then tapped into an AI security platform that found 76 more that the manual search missed.

Having now discovered AI apps, you can move on to the task of blocking malicious or unapproved apps and plug-ins (that multinational consulting firm mentioned earlier blocked those three apps pronto). There are various ways to block bad apps, but the

most efficient and effective approach is deployment of a platform built specifically with AI security in mind. The aim is to essentially have an allowlist of approved AI applications, so you can then keep everything else away from your valuable and sensitive resources.

After you've determined what apps are approved, your system must govern them. It's not just a matter of saying "yes" and calling it a day, because even approved applications can still be misused. Your visibility and control process must keep tabs on who's using GenAI apps, which ones, and what they're doing, so you can protect against sensitive data exposure through that use.

What does that look like? It's monitoring employees' use of external AI tools through continuous collection of telemetry, including a secure browser that provides visibility into how data is used and shared. You're continually auditing data flows and identifying proper use patterns, seeing what prompts ought to look like.

Knowing what's normal is the key to spotting anomalous behavior. Strange queries can be a red flag, or unusually large data transfers, or connections that are unexpected.


TIP

What happens when one of those red flags comes up? Whether it's data exposure or policy violations or something else, it's time for incident response. An effective system must have incident response procedures that are established up front and as clear as possible, so your operation can dive right into containment, investigation, and any necessary remediation. AI incident playbooks, created with the involvement of data privacy and compliance teams, are essential. Any solution must include a record of all activity that has taken place, session recordings, and event timelines for thorough investigation and forensics.


REMEMBER

One point that should go without saying but is absolutely vital to remember: This is not a checklist but an ongoing loop of activity. AI is growing and changing every minute of every day, at a rate that can be downright alarming. What you knew yesterday won't be good enough tomorrow, so this whole process needs to happen continually.

Getting a Handle on Third-Party Apps

REMEMBER

To stay ahead of risks, it's vital to quickly identify emerging GenAI apps and platforms and their attributes. The battle for AI security happens on a lot of fronts, but third-party apps are a key place to focus. To see an example of how this can play out, it's worth having a look at how the process of securing GenAI usage is handled by Palo Alto Networks.

The idea is to pull together into one place comprehensive data protection. Real-time visibility and security of AI use comes from being able to granularly see at the user level all data and apps being used, how they are used, and what actions are being performed on these apps. That allows attack surface protection that blocks unsanctioned and high-risk apps, uniformly and granularly applies information security policies, and keeps threats at bay.

Here's what it looks like:

- **Gaining visibility into AI apps:** This is the discovery phase that uncovers GenAI app usage across all the various possible use cases. The Palo Alto Networks approach that leverages Prisma Browser delivers in-depth visibility into all GenAI use in the organization.
- **Classifying and controlling app access:** Classification of apps can be as simple as deciding whether they're sanctioned or unsanctioned (or at least tolerated). Determine access based on more than 80 granular risk categories that are automatically calculated. Application process controls can be easily set up with the help of out-of-the-box best-practice policies.
- **Establishing data access controls:** It isn't just apps that need to be classified; data does, too. This approach taps into large language model (LLM)–powered, context-aware ML models to classify data, with more than 300 potential categories. Contextual in-line policies can then prevent GenAI apps from getting their hands on sensitive data. According to a study, 85 percent of the work happens in the browser. So, a secure browser makes absolute sense to start gaining visibility and implementing control with encrypted traffic.

- **Building security controls:** This begins with uncovering interconnected GenAI apps within SaaS marketplaces. It's essential to identify, monitor, and remediate any AI bots that are not authorized. Security controls detect threats being delivered in AI responses, including malicious files or URLs. And they monitor and maintain GenAI app posture for compliance.
- **Monitoring continuously:** Old or new risks can arrive anytime, so risk monitoring must be continuous. The Palo Alto Networks approach keeps tabs on app adoption and usage insights across the various app categories, and delivers comprehensive reports on GenAI app usage, risks, security, and compliance. It also makes contextual recommendations for strengthening security controls. Its secure browser provides ongoing visibility into what AI applications are used and monitors and controls how data is shared in these apps.

TIP

Remember that every new app that enters your ecosystem has an impact on your enterprise security posture. That's why you must automatically and promptly vet every one of them. Given the buzz about AI, apps will arrive in a grassroots (dare I say viral) fashion, either on their own or by way of plug-ins and connectors. A close eye is essential.

That automatic detection and control also needs to be focused on detecting and managing permissions. As it happens, excessive permissions are among the biggest problems in the GenAI ecosystem. Apps and plug-ins, even the authorized ones, must be allowed only the permissions and access they need, nothing more.

Establishing Best Practices

Much of the discussion here has to do with technological approaches to securing your AI ecosystem — and that's absolutely essential. Your enterprise needs all the protection it can get from external threats and accidental exposures.

That said, the job of AI security isn't complete until the whole team is onboard. Given that a fair amount of the AI-related danger has to do with risky employee behavior, educating the workforce is an essential part of getting where you need to be.

You can have the most advanced controls available but still end up in a bad place if employees are doing things with AI tools and have no idea how they operate and what risks they can introduce. AI tools don't come with warning labels like cigarettes or beer cans do — education is the organization's job, and it should really be viewed as the first line of defense.

TIP

Effective training begins with raising awareness of the risks associated with GenAI, and part of that lesson is sharing how these tools work. Employees must understand that GenAI tools are always learning, and they're learning from their interactions with users. They're capturing information, including the information that goes into prompts. Every interaction, therefore, is a data-sharing decision.

Be sure all employees know very specifically what a risky prompt may look like and what not to feed an AI tool, at least one that's not approved for handling sensitive information. Protected health information (PHI), personally identifiable information (PII), financial data, intellectual property, trade secrets, source code — a smart employee may know not to post such things on Facebook, without realizing that it's also inappropriate for an AI prompt.

REMEMBER

Education about best practices also must dive into ethics and a broad understanding of the potential downsides of AI-generated outputs. To begin with, employees need to be reminded of the dangers of inaccuracy, bias, or misinformation. They need to understand that users are always responsible for outputs — "the AI said so" is not an excuse for sharing inaccuracies or getting into compliance trouble. That means they must keep the old "trust but verify" adage in mind. They should seek information sources and cross-check AI claims against known and trusted sources.

Transparency must always be the expectation, too. Employees need to be up-front when using materials generated by AI and maintain records of which tasks happened with the help of AI. They also need to always be respectful of intellectual property.

It bears repeating that because AI is always evolving, education about best practices needs to evolve, too. Examine incident data and audit findings for ideas of new warnings and tips to share. Employee feedback can help improve content as well.

Building Policies and Governance

Just as it's vital to get employees in the know and onboard with the task of AI security, policies and governance must reflect the challenges of today and tomorrow to ensure safe adoption of AI. It must be crystal clear how systems are to be used, monitored, and managed by the organization.

TIP

An acceptable use policy is a good start (and when it's adopted, the policy should be an early lesson in the educational effort discussed earlier). Employees must know explicitly how they can use GenAI tools and how they can't. That means defining both the permissible use cases and listing the approved AI tools.

Needless to say, any policy must prohibit entering any confidential information into unvetted or unapproved systems. It should establish rules for how agents are integrated into workflows. It should make it clear who owns the intellectual property of AI inputs and outputs and require that AI-generated content be identified as such.

Governance will need to align with the organization's overall data protection frameworks. From a data handling perspective, that includes defining what categories of data may be used with AI models, whether anonymization is necessary, what data residency rules apply, and what human approvals are required.

The compliance department will be a key player in creating such policies, of course. That's a good thing because they also are the experts in ensuring that AI use adheres to the rules and regulations coming from both governments and industry standards bodies. There is plenty of attention already in such areas as privacy, data minimization, audit trails, data transfers across geographic borders, and business associate agreements with AI vendors.

Expect governments and standards organizations to play an ever-more-influential part in the compliance picture. Their policies are evolving rapidly, which means all enterprises building AI into their operations will need to pay close attention, keep up, and stay in alignment.

For example, the current U.S. administration has mandated that federal agency security risk management processes take into

account management of AI software vulnerabilities and compromises. "America's AI Action Plan," put forth by the White House, states that "promoting resilient and secure AI development and deployment should be a core activity of the U.S. government." It calls for new federal efforts focused on analyzing and sharing information about AI vulnerabilities.

Beyond that, voluntary bodies such as the National Institute of Standards and Technology (NIST) and Open Worldwide Application Security Project (OWASP) are ramping up AI security overlays for already established cybersecurity standards.

Meanwhile, given how much AI activity will happen across a vast, cloud-based infrastructure, there's an urgent need for foundational cloud and network security capabilities. That includes application-level visibility that reveals which applications, data sources, and application programming interfaces (APIs) connect with AI tools.

It's also essential to enforce policies and processes preventing lateral movement of threats that may find their way into a network. The usual menu of protections will be needed, including zero-trust segmentation and identity-based access controls. Your focus on cloud and network security capabilities also must include protection of cloud-native workloads, because GenAI applications often are deployed within containers, in serverless environments, or across multi-cloud infrastructures.

Achieving Data Security

AI couldn't revolutionize the world without close and continuous interactions with data of all kinds. Agents, copilots, and AI plug-ins are able to connect directly with data sources, including corporate data. That's great for productivity, of course, but it also happens to increase the risk significantly.

WARNING

This creates a need to rethink data security. Traditional frameworks for data loss prevention (DLP) were built for a world of structured, predictable files and dataflows and keyword-based rules. They can't keep up with the new flows of unstructured content, including AI-generated text, images, and source code.

What's needed are upgraded approaches to achieve data security in today's landscape and tomorrow's evolution of it. For example, SaaS security posture management (SSPM) offers real-time, continuous visibility into how SaaS-based AI agents, copilots, and plug-ins are behaving.

Secure browsers, another upgraded approach, deliver granular visibility and control at the user level into all app use, including GenAI tools, on any device.


TIP

Also important are efforts to properly prepare and sanitize data, before it has the chance to hit a model. Anonymization and pseudonymization processes can prevent training or inference data from being traced back to real identities. Automated redaction can remove identifiers from input data.

How can you be certain you're protecting all sensitive data appropriately? Classification is essential, and AI-augmented classification can handle this task automatically with incredible precision, far more effectively than traditional regex-based data classification. This protects data while greatly reducing false positives.

The best approaches include pretrained ML classifiers, as well as user-trainable models for documents and images, which helps protect such assets as patents, contracts, and source code. Accurate detection of sensitive data needs to be implemented at the point of access to these tools — in the browser.

It's also vital to carefully evaluate risks common across public cloud environments and AI environments. Best practices include a cloud network risk assessment that spots exposures and analyzes posture. Cloud firewall benchmarking should check for security gaps, and an AI risk assessment should identify exposures across the whole AI ecosystem.

Because of the challenges inherent in hybrid infrastructures, multi-cloud environments, and distributed teams, siloed tools can't cut it for managing security and performance. You'll find the greatest protection by unifying networking, security, and operations into a single, cloud-delivered architecture. That approach keep protection as simple as possible as the enterprise scales.

The only way to win at this game is to have AI on your security team, too. Autonomous AI agents can help optimize the digital experience and adapt to changing information flows, while also maintaining the highest security. Accelerated innovation is essential for corporate success, but security is just as essential. There's no reason they can't coexist happily.

IN THIS CHAPTER

- Getting into the enterprise AI business
- Thinking about security from the start
- Staying safe at runtime
- Building effective guardrails
- Testing to ensure security
- Getting control of access
- Ensuring appropriate anonymity
- Securing data pipelines and storage

Chapter 3
Building Secure AI Applications

The last chapter talked about employees diving into do-it-yourself artificial intelligence (AI) app innovations. There are plenty of great reasons for your company to develop its own enterprise AI, too, and that's the focus of this chapter. It covers embedding security throughout the process, from design through runtime through data handling.

Building Your Own AI Apps

Everybody who is anybody is building AI into their workflows, products, and services. It isn't just the way of the future — it's the way things are *today.* Growth-oriented enterprises know they need to step up or get out of the way of their competitors.

Off-the-shelf AI tools will only get your organization so far. Building your own proprietary AI lets you tailor models to your customer experiences, your workflows, and your data structures. Customizing can create competitive differentiation.

You also may consider building your own AI to better protect your data and security. You likely have a fair amount of sensitive data, particularly if you're in a highly regulated sector such as health care or finance. Internal AI development can allow greater control over such things as privacy and data residency (and truth be told, those are vital considerations even if you're in a less regulated industry).

REMEMBER

Creating custom AI apps also can help you more effectively integrate their functionality into your current enterprise systems. You can achieve greater control over how data moves between systems, which can be both more efficient and more secure, with less reliance on third-party application programming interfaces (APIs).

Ultimately, going with purpose-built, custom enterprise AI systems can strengthen strategic capabilities for the longer term. You can get into AI quickly by outsourcing your technology, but for many companies, making AI a core competency is the way to go for long-term strategic wins.

No wonder, then, that nearly half of enterprises are building AI applications, according to Menlo Ventures research. Hugging Face — an open-source AI hub for machine learning (ML) models, data sets, and various tools — reports more than two million models on its platform. And Palo Alto Networks research expects there will be at least 100,000 enterprise AI apps by the end of 2026.

Most companies don't dive into training their own foundation models from scratch, as that can be incredibly costly and complicated. Plenty of pretrained foundation models are available to be adopted and fine-tuned.

That approach allows enterprises to layer in their own proprietary data and domain-specific tuning. Model orchestration frameworks make it possible to tap into multiple models to handle different tasks, such as reasoning or decision support or summarization. Models can be deployed in private environments to help contain exposure risks.

That said, there are still plenty of risks. For one thing, if you tap into external models, you don't have full visibility into their training data, their potential for bias, and any other vulnerabilities. You also must watch as models evolve and potentially drift,

which means constant reevaluation to maintain compatibility and effective guardrails.

There's always a risk for data leakage, a risk that is greater whenever there are outside connections. Reliance on external models also introduces supply chain risks and dependencies. Putting vital functionality into the hands of outside vendors puts you at someone else's mercy — and their potential problems can become yours.

WARNING

One other risk that isn't all that different from tapping into third-party apps: Your AI development can get out ahead of your policy development if you're not careful. As you race forward for fear of missing out, you can run into compliance and governance gaps, and that's just as possible with internal AI development as it is with tapping into third-party tools.

From an overall security perspective, companies that are venturing into their own AI development are stepping out into an entirely new and complex threat landscape. Traditional cybersecurity solutions weren't designed to address these perils.

Including Security from the Start

No matter what you're developing, it's vital to integrate security into the AI/ML development lifecycle.

Start with design; this work needs to include threat modeling and a secure architecture review, and ensure that there's least-privilege access.

In development, your team must employ secure coding, data validation, and dependency scanning.

When it comes to training, encrypted storage is a good practice, along with model integrity validation and isolation of training data.

The testing and validation phase of AI application development must include adversarial robustness testing and red teaming, and it must take policy enforcement into consideration.

When it comes time for deployment, it's important to keep scanning for backdoors and unsafe serialization, as well as the whole

host of AI-specific threat vectors, with continuous security testing within the continuous integration/continuous delivery (CI/CD) pipeline.

Let's talk about point products. They're out there helping with some of these needs, with niche players taking on specific aspects of AI security such as data leak protection, vulnerability scanning, and detection and response. The problem is, you end up with a fragmented security posture. You may still be vulnerable to such threats as model theft, excessive agency, or model denial of service (DoS).

Similarly, most native GenAI platforms have some built-in guardrails, which is great, but those are limited, too. All kinds of adversarial prompt attacks, not to mention sophisticated jailbreaks, can get past these guardrails. They often won't stand in the way of sensitive data leakage, they can miss malicious URLs embedded in model outputs, and they often lack basic firewalling or multi-cloud support.

What about traditional firewalls? They're generally not ready for new AI threat vectors either. They're not prepared for the full spectrum of AI-specific threats.

REMEMBER

In any case, even without their shortcomings, these options provide a very fragmented security blanket. That means there are going to be blind spots in terms of knowing what AI assets are running, what their dependencies are, and how their permissions are set. You can't operate safely without better visibility, and you can't get that without adopting a broad-platform approach to AI-related security.

A platform approach is eminently possible, and it's the most effective and simplest path forward. Following are some of the elements you need from a platform solution.

Model security

Your solution needs to proactively identify vulnerabilities before deployment. It's better to uncover issues before they're out in the wild.

Model scanning involves both static and dynamic analyses that are checking for risks hidden in training data, issues with architecture, and behavioral red flags. You're looking for things such as

backdoors and unsafe serialization, model tampering, and malicious scripts hidden within models themselves.

Agent security

Your AI security solution must secure AI agents, an up-and-coming opportunity and risk. It should watch over agents on no-code and low-code platforms, and monitor for such threats as identity impersonation, tool misuse, and memory poisoning or manipulation.

In the same way you protect your landscape from humans, you must be sure agents are operating with least privilege. And in the AI world, you also must ensure agents are not acting on hallucinations. (Look for more on agentic AI in Chapter 4.)

Posture management

AI/ML security posture management creates continuous visibility into the security configurations of your data pipelines, your deployed models, and your training jobs. This is proactive monitoring for misconfigurations, sensitive data exposures, and excessive permissions.

Think about cloud security posture management and how it automates visibility into your cloud environments and infrastructure. You want the same kind of peace of mind in your AI ecosystem.

Red teaming

Your security must think like a villain to contain the real villains. Systematic red teaming and adversarial testing of your AI systems means spinning up real-world attack simulations to spot vulnerabilities and predict emerging bad behaviors.


TIP

What you need are automated penetration tests on AI apps and models. You must stress-test your AI deployments, learning and adapting the way a bad guy would. Your solution should provide detailed reports offering ideas for mitigation.

Runtime security

Though you want to build as many security practices into the left side of the development model, runtime is where the rubber meets the road. Now that you're in production, you're on the lookout for

real-time threats such as prompt injection, toxic content generation, data leaks, novel agentic attacks, and model DoS.

Runtime security can happen at various enforcement points, including the network layer and APIs. (More on this vital element later in the chapter.)

Protecting Runtime Operations

No matter how many safeguards you've implemented during development and testing, there's no telling what's going to happen when the curtain goes up and the show gets underway.

This is true, of course, with everything you develop and put into production. It just happens to be especially complicated in the world of AI. AI systems are built as compound systems, requiring implementation of an entire AI stack that integrates multiple models and plug-ins and vector databases and perhaps even internet search capabilities.

REMEMBER

Every element introduces new potential vulnerabilities that can inconveniently appear at runtime. You need an enterprise AI security approach that boils down to three key functions: discovery, protection, and monitoring.

Discovering the AI ecosystem

This must be an automatic function that puts AI applications, models, users, and data sets under the microscope. You must gain a sophisticated understanding of how your cloud environments are using AI applications.

That begins with analyzing dataflows, visualizing the asset inventory, and understanding how all parts communicate. Discovery includes mapping apps with their relevant models, users, plug-ins, data sources, and external sites.

You're looking for connections, and you're bound to find ones you didn't know existed. These insights help you know where to focus your runtime security and how to ensure compliance across cloud-based implementations.

Protecting against AI-specific threats

At runtime, your AI-focused security has a number of key focus areas, including the following:

- **Application protection:** This includes robust defense against malicious URLs, achieved by means of scanning URLs going back and forth between AI apps and models. Block applications from displaying or fetching malicious URLs, which helps keep users safe from receiving bad URLs in model output. Data exfiltration attacks also may try to trick a model to build a URL that gets an end user to unwittingly send sensitive data to the attacker's server.
- **Model protection:** Direct and indirect prompt injections are key culprits to watch for here. Impersonation is one issue, but prompt injection can also lead to goal hijacking and do-anything-now attacks.
- **Data protection:** Your AI security must be able to ensure that training data doesn't get leaked in application outputs. It should watch data patterns in prompts and responses to look for troubling activity. For models that generate database queries, it should prevent unauthorized changes to databases.
- **Agentic threat protection:** It's more and more critical to secure AI agents from such trouble as identity impersonation and memory manipulation. Tools and APIs connected to agents are in danger of abuse, so your security must protect against tool misuse.

Monitoring runtime risks

REMEMBER

AI is always evolving, and runtime is always running. That's why continuous monitoring is essential. A runtime security program must constantly analyze AI runtime risk posture and provide insights into vulnerabilities that are right there in operational systems.

Your security must be able to find and identify unprotected AI applications, spotlight risky communication pathways connected to AI applications, and watch for potential entry points where malicious activity could find its way in. Proactive monitoring yields greater security and reduced risk exposure.

Including Governance and Guardrails

Data is at the absolute heart of AI systems, the key to the intelligence that happens to be artificial, and also the subject of many of the actions that AI systems take. We've talked a lot about protecting that data, which is essential, but it's also important to consider the governance and guardrails around the data.

REMEMBER

AI is, after all, only as trustworthy as the data on which it runs. A robust governance framework must ensure that all data used in developing, fine-tuning, and deploying AI systems is accurate, compliant, ethically sourced, and secure. Here are some key aspects of governance:

- **Data classification:** Think of this as the foundation of governance — you can't govern what you don't understand. Data should be categorized in terms of its sensitivity, its business value, and any regulatory obligations. Is it public and safe to share, or internal and not for public release? Is it confidential, business-critical information, or perhaps restricted in terms of being personally identifiable information (PII) or protected health information (PHI)?
- **Ownership:** All data sets should have designated data owners who are responsible for overseeing usage and ensuring compliance. Data stewards are in charge of quality, integrity, and documentation, which can be vital for traceability and potential audits.
- **Lifecycle management:** From data ingestion and storage to deletion or archiving, data governance requires end-to-end lifecycle management. This is how you keep on top of whether training data is current and approved, and prevent shadow data from finding its way into the mix.
- **Quality assurance:** Reliable outputs depend upon good data. Governance practices here include cleansing to remove corrupted or duplicate entries, bias detection, and ongoing monitoring to watch for data drift.

Within applications, *guardrails* are procedural controls that ensure AI models stay aligned with the standards set forth by policies and governance. They may include context filters and controls, input validations, access controls, model lineage tracking, and version control.

AI systems that generate content may require moderation guardrails to help the organization steer clear of reputational hazards. Toxicity classifiers and other context-aware filters should be watching for biases, misinformation, or even hate speech. Model outputs should also align with brand tone and ethical standards.

Running the Tests

Before your organization's new enterprise AI application ever sees the light of day, the application must be subjected to rigorous vulnerability management, penetration testing, and red teaming. These are key approaches for ensuring that you find weaknesses before attackers have a chance to do so.

REMEMBER

Red teaming, for example, simulates real-world attacks, checking to see how AI systems respond to such threats as prompt injection, data poisoning, and model manipulation. This isn't a new concept, of course, but it needs to be different from static red teaming approaches that rely on predefined test cases. AI is dynamic and so are AI threat actors, so this exercise must be dynamic as well.

That means understanding the context of the application, such as health care or financial services, and intelligently targeting the kinds of data and vulnerabilities that attackers would try to exploit. This red teaming shouldn't be a pass-or-fail test. If the testing engine fails, it needs to be able to learn and figure out new strategies and keep on testing until it finds an exploitable path. In other words, it needs to behave the way an AI-powered threat actor would behave.

Results should filter back to runtime parameters, and be used to create comprehensive reports showing where the attacks succeeded and what sensitive data would have been extracted. These kinds of real-time recommendations are essential for improving AI security posture. And notice that I said "real-time" — this testing is continuous and ongoing, part of a runtime security effort.

Controlling Access

Let's start by making the long story short. Access controls are essential in AI system security, just as they are in every other aspect of IT security. In fact, the access concepts pertaining to AI and agents will sound quite familiar, because they're very similar to the ways you secure your infrastructure from human threats.

That means implementing stringent access controls to ensure that only authorized AI models, applications, and personnel are allowed to access specific data sets. Notice that I list *personnel* third on that list, behind nonhuman actors such as AI models and applications. Given that these players are out there running autonomously, or at least semi-autonomously, your access controls are right there on the front lines guarding the henhouse from the foxes.

REMEMBER

Tight access controls are how you keep unauthorized personnel or nonhumans from extracting sensitive data or retraining models with the wrong data. This is the way to be sure rogue AI agents aren't calling APIs beyond their intended scope. It's how you ensure compromised apps aren't accessing restricted data sets. Here are some key access concepts:

- **Role-based access controls (RBACs):** These controls base permissions on job functions instead of individual identities. Among humans, data scientists may have different access from operations engineers or end users on the business side. RBAC should restrict permissions for applications and agents significantly, greatly limiting or preventing access to specific data.
- **Attribute-based access controls (ABACs):** One problem with RBACs is that they don't always operate on enough context to make the right decisions, especially when it comes to dynamic AI workloads. ABACs enforce policies based on user attributes, as well as attributes of the resource, what kind of action is being taken, and other contextual elements. This is more fine-grained policy enforcement, and it can help ensure that AI agents and tools are sticking to their approved parameters.
- **Principles of least privilege and least control:** As is the case with humans, AI models and agents should only be

allowed whatever access and controls are necessary for their specific functions. With regard to data, that could mean AI agents can only read, not write or modify. They should be limited in their API usage and which end points or services they can call. They should only interact with permissible environments and query allowed models.

Proper posture management means enforcing these access restrictions, along with least privilege and least control. The right tools keep a constant watch over agent configurations and permissions, look for over-permissioned roles, spot unsafe combinations of APIs, and automatically revoke improper access and privileges. This has to be a continuous part of runtime security, because threats never sleep.

Creating Anonymity

The more an AI system knows, the more powerful and intelligent it can be. Imagine what AI could accomplish if its model could be trained on all available attributes of every cancer patient in America — not just disease details, but age, gender, race, diet, exercise, environmental factors, and more. Then toss in treatments and outcomes.

That kind of a system could be unspeakably powerful in its ability to predict diagnoses and recommend the right treatments and lifestyle changes. It also would run the risk of violating the Health Insurance Portability and Accountability Act (HIPAA), if it were possible to identify any of the patients whose details were used to train the system.

REMEMBER

Creating anonymity is the way to balance privacy concerns and regulations with the importance of useful data. Your organization may need an AI data minimization strategy to de-identify sensitive elements before model training and ensure compliance, while preserving the statistical integrity required to get what you need from the data.

Anonymization is the process of irreversibly removing all personal identifiers. Done properly, individuals will never be able to be identified, even if auxiliary data sources are tossed into the mix. Techniques include stripping out names, patient ID numbers,

Social Security numbers, that kind of thing. You can also employ generalization to replace specific values with broader categories, or perturbation that uses statistical noise to block exact values.

Pseudonymization, on the other hand, takes those identifiers and replaces them with tokens. Data can be linked across systems while not exposing original values. Someone's name, for example, is replaced by a random number, or cryptographic hashes are used to prevent reverse engineering. There may be maps retained connecting the original with the pseudonym, but those are kept out of the training data.

WARNING

With anonymized data, there's a risk of reidentification if other data sets are combined into the AI system without proper controls. Reverse mapping is a danger with pseudonymization, especially if hashing is weak or tokens are reused.

Another strategy is synthetic data, which is artificially created through probability distributions. There is no link to real people, so privacy risks aren't an issue if done properly, but there is potential for bias.

Securing Pipelines and Storage

Here's one more consideration when it comes to data security in the world of AI, and this, too, should sound familiar. Data pipelines and storage must be secure so data is always protected as it journeys to and from AI systems.

As with every other way you use data, you must use secure ingestion pipelines that authenticate data sources and sanitize inputs. Data should be encrypted while at rest and in transit. It needs to live in a secure data lake or data warehouse, with robust at-rest encryption and systems enforcing least-privilege access.

IN THIS CHAPTER

- Getting workflows accomplished without humans
- Controlling agent identities and permissions
- Keeping agent communications secure
- Monitoring agentic AI continuously

Chapter 4
Securing Agentic AI

Agentic artificial intelligence (AI) promises both the most tantalizing technological opportunities in ages and some of the most troubling risks. It's AI that is truly in action, working through autonomous workflows efficiently and effectively.

This chapter explores how agentic AI works, what it can accomplish, and what risks it poses. It then spells out how to mitigate those risks through agentic authentication and authorization, tightening security on agent communications, and keeping a close eye on their activities.

Tapping Into Autonomous AI Workflows

When ChatGPT burst onto the scene in 2022 and invited regular folks to give generative AI (GenAI) a try, many people were wowed by the possibilities. They were interacting with a large language model (LLM) that seemed to know practically anything and everything and could put it all into words quickly and skillfully. Pretty amazing stuff, for sure, but agentic AI stands ready to *really* blow people's minds. It's not just thinking, not just writing, but going out and doing things on its own.

At its core, agentic AI is automation, but that description really doesn't do it justice. Automation is when a stoplight gives you a green arrow after you pull into the left-turn lane. Or the ATM that hands you two $20 bills. Or that contraption at the fancy doughnut shop that spits out rings of dough and floats them in hot oil to cook on one side, flips them over to cook the other side, and then drizzles them with glaze. Tasty, yes, but pretty limited capabilities.

REMEMBER

Agentic AI, on the other hand, involves fully autonomous workflows guided by artificially intelligent orchestration layers controlling teams of specialized agents. For example:

- You call the cable company when your TV signal goes out. The conversational agent asks what the problem is, taps into other agents to consider possible causes, checks for a broader outage, and then sends a signal to reboot your cable modem. You get back to the football game broadcast before the next first down, without waiting for a human.
- A hospital system's agentic AI includes a bed management agent that tracks patient occupancy, a discharge prediction agent that makes educated guesses about when patients are likely to be sent home, a staffing agent that recommends nursing schedules based on patient census and dispatches crews to clean rooms after discharge, and a communication agent that keeps the humans informed.
- A pharmaceutical company wants to accelerate drug discovery, so its agentic AI platform has one agent scanning and summarizing medical journals in search of new ideas to try, another running simulations of molecular interactions to predict the effectiveness of potential new chemical compounds, and one that pulls together lab results and clinical trial data.
- This book is about IT security, so imagine an agentic workflow that determines what normal information technology behavior looks like, monitors system logs looking for anomalies, analyzes alerts to determine severity, prioritizes problems, fixes issues by applying patches or revoking credentials, and then creates compliance reports. That requires multiple coordinated agents.

All of these things can now happen with little to no interaction with humans. Agents, under the umbrella of agentic AI orchestrate workflows, autonomously analyze situations, decide what needs to be done, and then take care of it.

REMEMBER

It's worth making a distinction between AI agents and agentic AI. The agents are out there handling whatever tasks they're meant to handle — think of agentic AI as the orchestration layer for the workflow the agents are handling.

There are a number of ways this orchestration can happen. A hub-and-spoke pattern will feature a central orchestration point that assigns subtasks to task- and domain-specific agents. A peer-to-peer collaboration has agents communicating directly by way of application programming interfaces (APIs), dividing up the duties. A hierarchical arrangement sets high-level objectives that are then split into subgoals for agents to execute.

Any of these agentic AI approaches may be accomplished by software as a service (SaaS) agents, as well as internally developed enterprise AI agents, or some combination. Essentially, agents are working together to achieve a common goal and work their way through some sort of predetermined workflow.

But again, it's not just automation, because the agentic AI system is making its own decisions along the way, based on circumstances and contexts. In certain cases, there may be human approvals somewhere in the loop, but for the most part, we're talking about autonomous decision-making and execution.

The idea is powerful and the possibilities are endless. But the risks are real. Every agent's tool set is a potential ingress point for malicious action. There's a risk of implied trust chains, where agents mistakenly assume other agents are legitimate.

If agents have too much autonomy or too broad a set of permissions, they can end up doing things they're not supposed to do. That can also happen if their actions are based on LLM reasoning coming from a model that has been manipulated through prompt injection or data poisoning. And working SaaS agents into the workflow increases the risk of shadow AI — agents that escape the oversight of IT.

Knowing the Agents

Let's consider for a bit one of the potential risks. It's a risk that has been around for years when it comes to humans, and it's increasingly a risk involving nonhuman agents. The challenge is trust: How do you and your agentic AI system know that the various agents are who they say they are? How do you know they're trustworthy? How do you know they're doing what they're supposed to be doing (and nothing else)?

Humans have been fooled and conned by other humans for as long as they've been interacting with each other. As it happens, agentic AI systems are just as susceptible to trickery. That means all the kinds of safeguards needed to protect against malicious humans need to also be in place in fully agentic orchestrations. We're talking about strict identity and access management.

REMEMBER

Humans need credentials and permissions, and agents also must be able to authenticate who they are and then act only within whatever boundaries are authorized. Without strong controls, you can end up with agents being spoofed or hijacked or over-permissioned. That opens the door for data being leaked, fraudulent actions, and operational chaos, among other perils.

Each and every AI agent that exists to execute a task must be set up with an identity that is unique and traceable. It'll be like any other service account or workload identity, and it will facilitate accountability for all actions taken by the agent, audit trails, and permissions that are carefully limited to the functions of that particular agent.

Similarly, the supporting infrastructure needs to be set up with verifiable machine identities. Containers, microservices, orchestration layers — the whole environment in which the agent is working needs to be trusted before any credentials are exchanged.

A smart approach is token-based authentication, as opposed to static keys and passwords. Tokens such as OAuth 2.0 indicate delegated access rather than the agent's permanent identity. They can include contextual attributes that establish such things as type of task and sensitivity level. And they can be set to automatically expire.

After authentication, authorization models can then spell out what the agent is allowed to do. Role-based access controls (RBACs) can, for example, be tied to predefined permissions. You may have HR workflow agents or data retrieval agents or other similar descriptions. It's a decent approach for agent functions that are fairly narrow and predictable.

Attribute-based access controls (ABACs) are a bit more fine-grained and especially well-suited for environments with adaptive and autonomous agents. ABACs consult a list of contextual attributes to dynamically grant or deny access. Time of day, location, agent purpose, and data sensitivity are among the possibilities. They can be defined to allow access only if certain combinations of contexts are confirmed, and deny access in all other cases.

You can get even more fine-grained by adding in policy-based access controls (PBACs), which may start with ABAC logic and then combine that with specific policies for careful runtime enforcement.

The time-honored principles of least privilege and least control are essential here. Agents shouldn't be allowed any more access than their function needs, nor any more permission.

Because agents often need to access sensitive systems, their credentials must be managed carefully and securely. That can mean centralized secrets management for issuing and revoking tokens, and perhaps include short-lived credentials that are around only long enough to cover a session or workflow. Agent identities should be provisioned and deprovisioned in ways similar to human identity lifecycle management. Needless to say, agent identity and authorization must always be validated at runtime before anything sensitive is allowed to take place.

REMEMBER

Herein lies a big challenge with agentic AI. There's a delicate balance between usability and security. It would be ideal to provide agents with complex and flexible authorizations. For example, a powerful scenario would be for an agent to be allowed access to a bank account, but then provide it with only the authorizations needed to do whatever transactions pertain to the specific task at hand.

In that scenario, the agent's permissions would also depend on its state, its architecture, and its behavior patterns. If that agent is

exhibiting unexpected behaviors, authorizations would be immediately revoked.

Securing Communications and Workflows

As explained earlier in the chapter, agentic AI isn't just a matter of an AI agent taking some sort of action — it's taking part in an orchestrated, collaborative series of actions involving a whole ecosystem of agents. That can only happen when all the players talk to each other, and it can only happen safely if those conversations are fully secure.

Secure communication channels between AI agents and the broader enterprise ecosystem are essential. All components and parties must be able to communicate with confidentiality and integrity, and they must be able to ensure the authenticity of messages exchanged by and between agents.

Agentic AI systems must be built with a security-first philosophy. That's no small feat, given that the whole landscape is constantly changing. Communication protocols and agent architectures are redesigned and prototyped continuously, and new and never-seen-before vulnerabilities are commonplace.

Agents interacting with tools and other systems do so through a number of different communication mechanisms. APIs, function calling, message queues, and standard networking protocols are among the ways communications are initiated. These methods get the job done, but not always in the best way from a security perspective. They may not convey whether an agent is supposed to have specific access, and they don't really offer a way to keep rogue agents at bay.

TIP

For now, best practices include the following:

- **Designing secure APIs for agent interactions:** APIs are the primary ways that agents talk to applications and get things done. Strong identity checks are essential, with each agent having a digital ID to help prove legitimacy. Least privilege ensures agents aren't allowed to go overboard with their actions, and it's essential to validate data as it arrives and

departs to be sure it's as expected and doesn't include malicious instructions.

- **Encrypting communications:** This is always a good idea for keeping outsiders out, whether data is in motion or at rest. Consider mutual Transport Layer Security (TLS), so that both parties verify each other's identity. And rotate encryption keys for added long-term security.
- **Securing the messaging queues:** Agents may find it handy to drop messages into queues or event streams, but doing this safely requires authentication and authorization for both the sender and the recipient. Again, encrypt messages, validate message formats to deter malicious data, and keep messages only as long as necessary.
- **Building robust orchestration mechanisms:** Agents must be limited to executing tasks within defined boundaries and with proper validation. The orchestration layer should clearly define those boundaries and double-check that actions are legitimate and aligned with policy. Sandboxing agents helps contain agent misbehavior, and constant monitoring is vital. For certain high-risk or sensitive operations, consider including a human approval step.

Watching the Agents

Agentic AI is not your average software, that's for sure. It doesn't look like it or behave like it. Agents are dynamic and evolving, and they're able to execute decisions with little to no oversight. That is, of course, the whole point. Agentic AI is analyzing situations and taking care of business.

REMEMBER

That said, just because agents can get on with their day with little oversight doesn't mean they don't *need* oversight. Perhaps it's better to think in terms of agentic AI not needing micromanagement of each and every little decision. But in the big picture, agentic AI absolutely needs oversight — continuous oversight. It's out there taking real-world actions, so you really can't afford to just trust that agents are behaving.

The problem is, the usual frameworks for providing that oversight and securing traditional applications aren't enough in this

world. Agentic AI brings risks and potential vulnerabilities that are fundamentally new. Here are some of the kinds of oversight that you'll need:

- **Comprehensive logging:** Everything an agent does should leave a digital trail. Logging is always important for trouble-shooting, but in this case it's also about accountability and forensics. Every action and decision should be logged, including what action was taken, involving what systems, and what the outcome was. Include the context of what triggered the action, the model versions used, and the prompts provided. Centralize the logs and keep them as long as needed for investigation — and be sure to protect confidential information.
- **Real-time monitoring:** You've seen the words *runtime* and *real-time* all over this book, because I can't stress enough the importance of watching everything closely as it happens. You want dashboards and alerts monitoring agent activities, analytics that show when agents are starting to stray from expected behavior, automated safeguards that can pause or disable a potentially errant agent, and correlations with other systems to cross-check logs with other tools and control systems.
- **Behavioral baselines:** Earlier I mention being able to track when agents are straying from normal behavior. You can only tell that if you know what normal looks like. That means knowing what systems an agent normally uses, what data it accesses, how often it acts, and what kinds of actions it takes. Frequency, timing, and scope of actions are all important things to know, so your monitoring systems can tell when things are out of whack.
- **Incident response playbooks:** When something is amiss, it's important to act incredibly quickly. Agentic AI is good and fast at what it does, which means if it's going down the wrong path, it can do a lot of harm in no time. That's why you need your responses planned out in advance, for whatever situation may arise. Spell out steps to achieve immediate containment, verify what the problem is, and analyze the root cause so you can fix today's issue and prevent tomorrow's. Establish clear roles for mitigation and set up communication channels. The playbook should set out how to restore service once it's safe, and how to learn from the event.

IN THIS CHAPTER

- Realizing that security must change
- Convincing the C-suite
- Building the pillars of AI security
- Moving past fear toward opportunity

Chapter 5
Safeguarding the Future of Secure AI

This chapter envisions the future of AI — secure AI — and underscores that security protocols must move just as quickly into the future. It documents the importance of helping C-level executives recognize the business threats, summarizes the pillars of robust AI security, and offers a hopeful vision for stepping boldly and bravely toward a secure AI future.

Recognizing the Security Paradigm Shift

The risks are so new and so different that traditional cybersecurity has been left behind. What were modern tools not all that long ago are hobbled by the fact that they were made for more predictable systems, more static perimeters, and human-controlled activities.

AI, on the other hand, is continuously learning and dynamically growing. It's driven by vast data sets that are often filled with sensitive information, enabled by application programming interface (API) integrations and in many cases supercharged by third-party services. That expands the attack surface exponentially.

REMEMBER

The first imperative of this new paradigm is to secure human usage. It would have been hard to imagine just a few short years ago how easy it has become to experiment with AI, without any information technology experience — and that's exactly what employees are doing. They're trying out phenomenal new ideas with third-party generative tools, and in the process, they're sometimes exposing regulated data and trade secrets.

You can't fault their enthusiasm for innovation, but you must balance that enthusiasm with governance. Employees must learn how AI works, what safe practices are, what use is acceptable and what is not, and what content should never be allowed near a prompt. They should be trained when AI is acceptable, when it's not, when they should be skeptical, and why they must be transparent.

Knowledge is important, but guardrails are essential, too. Your security must prevent employees from using unvetted or dangerous tools, flag and filter risky prompts, and watch for malicious outputs before employees (and the organization) get burned by them.

REMEMBER

The second imperative is building secure AI applications. Safe and secure practices are required regardless of whether your employees are creating workflows with the help of third-party generative AI (GenAI) apps or whether your IT teams are developing enterprise AI applications in-house.

Security must be front of mind in every phase of the AI development lifecycle, from design to model training to integration to deployment to runtime monitoring. That includes model scanning in search of embedded threats, secure data handling processes, careful preparation and sanitization of model training data, ongoing red teaming both in development and during runtime, and AI posture management.

Security can't be added in or bolted on or piecemeal. Point solutions will leave risky gaps, as will whatever security features are built into the components of the AI system and the infrastructure on which it runs.

What you need instead is a unified platform approach that discovers all corners of your AI ecosystem to provide full visibility; that knows the latest threats and works continually to protect you from them; that focuses like a laser on safeguarding data;

and that takes care of model scanning, red teaming, and runtime monitoring. You need a platform that can live on-premises through a next-generation firewall or in a Secure Access Service Edge (SASE)–native cloud architecture (or both).


REMEMBER

The third imperative of this new paradigm is ensuring that agentic AI workflows are secure. The stakes are as high as they can possibly be with agentic AI, because AI isn't just creating content or making inferences — it's actually taking actions and making workflows happen without humans at the wheel.

Agentic AI magic often happens through APIs, each of which is a new potential point of compromise. Securing agentic AI means recognizing agents like they're people, subjecting them to identity and authentication and authorization frameworks. It means safeguarding agents' communications and learning what their behavior should look like, so it'll be clear when they're not behaving.

It's daunting, it's critical, and thankfully, it's entirely possible with the right tools. The new paradigm of AI security involves transforming mindsets into a new style of continuous, adaptive protection and governance. The best time to start thinking this new way is right now.

Bringing the C-Suite Onboard

Who isn't excited about the promise of AI? It's a safe bet that the people in the C-suite are as enthusiastic as anyone else — and as concerned about the prospect of falling behind the competition. They may be among the biggest cheerleaders for embracing the innovations that AI can bring the organization. These days, AI pretty much sells itself.


REMEMBER

What doesn't sell itself quite as well is the argument about the concerns and caveats. It's kind of a bummer really. And although C-level executives may not need all the details about APIs, prompt injection, and data poisoning, what they do need is a solid understanding of the business risks related to AI. They get the business risk of missing out, but they need to fully grasp all the other business risks associated with AI gone wrong.

Consider what happens when a model misuses or leaks data or produces biased outputs, or an agentic system inadvertently

opens the door to malicious actors. The end result can be a major hit to brand trust, to revenues and expenses, to regulatory standing, and to shareholder value.

Failures in the AI world can lead to operational disruptions and all the associated costs. They can invite litigation. They can set you back in an area where competitive advantage is essential. Chief risk officers have plenty of worries on their plate, and now they need to think about such things as autonomous AI agents going astray in ways that aren't immediately visible.

C-suite leaders should gain some familiarity with the major threat categories, such as data leakage, model manipulation, regulatory noncompliance, and agents going beyond their boundaries. Any of these threats can have an impact on fiduciary duties.

They need to understand how AI must be part of business governance. They have to know who is responsible for policies and enforcement, how human resources fits in with regard to education and employee expectations, and how marketing leaders fit in.

Ultimately, they need to understand why it's so important to invest sufficiently in secure AI foundations. They need to understand why the security systems they approved in the past aren't sufficient for the future. They must be persuaded that the powerful business benefits of AI can arrive safely only if security is taken seriously.

Building a Robust Security Framework

This book is an attempt to explain these risks — not to create fear, but to build familiarity and confidence. C-suite leaders especially should understand how to secure both the use and development of AI, the risks associated with it, and the potential business impact of these risks.

The first part of this chapter begins a conversation about three key imperatives for fully grasping the new paradigm of AI security. I cannot stress those concepts enough — they are so important that they can be seen as the three pillars of AI security. The following sections take a bit of a deeper look at how these three pillars together hold up the robust security framework required for AI security.

Governing human and AI interactions

REMEMBER

This pillar recognizes the fact that the first defense is behavioral in nature. AI may be about taking humans out of the loop to boost productivity and free people to use their brains elsewhere, but ultimately, it's human engagement that is at the start and finish of AI systems.

Humans are prompting GenAI tools, querying models, interacting with AI-driven customer interfaces, and setting up workflows for AI to assist with. Humans are making the decisions about what AI should be trying to achieve, and they're making a lot of the mistakes that lead to AI-related trouble.

That's why organizations must govern human and AI interactions. They must set the rules for who gets to use AI and in what ways, as well as what data does or doesn't connect with AI, and how AI data gets classified. Organizations must establish acceptable use policies and educate their people about those policies. Users must be trained to recognize AI risks, just as the organization likely already trains them to recognize phishing attempts.

This pillar sets the guardrails for people, and it also includes determining when people should set guardrails for AI by being in the loop for oversight of certain critical decisions. But looked at another way, this pillar establishes how AI-enabled innovation can happen, and how it can happen safely.

Building security into AI

The words of that headline — which is the second pillar of the robust security framework — were chosen carefully. It doesn't say that you're applying security protocols to AI. It says you're building security *into* it, which suggests it happens from the start.

REMEMBER

In other words, you can't retrofit security after AI models are developed and deployed. It must be engineered into every phase of the lifecycle, starting on the left side, all the way from design to model training and data ingestion to validation to runtime.

Building security into AI includes scanning models for various embedded threats and back doors before they're put into use. It ensures that data sets are properly anonymized and sanitized, verified, and well protected.

Posture management approaches should always be assessing exposures that may impact models and APIs. Red teaming should be continuously looking for weaknesses and keeping ahead of attackers. Data should be protected in transit and at rest through encryption.

It's really no different from the security mindset that has been right in the middle of DevSecOps for who knows how long. This is just a reminder that it's equally essential in AI systems development, something that can be easy to forget when everyone is working quickly to keep from falling behind the fierce AI competition.

Controlling autonomous workflows

The things AI is doing right now are amazing, but there's no doubt that the technology is only scratching the surface of what's possible. And what will really wow people in the future is likely to involve agentic AI and the autonomous workflows that it enables.

We're really near the beginning of the story about AI not just thinking and reasoning, but deciding and doing things. Agentic AI is orchestrating the behavior of agents whose tasks join together to complete powerful workflows. The potential is fantastic, and the potential risk is challenging to foresee. That's why controlling autonomous workflows is the third pillar of building a robust security framework.

Talk about a need for rules and boundaries. Agents whose jobs are executing multistep tasks and interacting with external systems absolutely must operate within clearly defined boundaries. These agents must have unique identities and adhere to token-based authentication. Just like people, they must be made to adhere to access and authorization models such as role-based or attribute-based access control.

Communications across the agentic ecosystem must be protected through encryption and should be logged and validated. No actions should happen until all parties to a communication have validated that their counterparts are legit.

REMEMBER

Proper agent behavior must be fully understood. That understanding drives the ability of continuous runtime monitoring to spot out-of-the-ordinary behavior, which must be evaluated and potentially stopped. Abnormal agent behavior could be a sign of prompt injection, goal hijacking, or API access that is out of scope.

Embracing the Future with Confidence

You're nearing the end of this book. If you've read it cover to cover, your view of AI may have evolved — from initial excitement to a more grounded understanding of the risks that come with it. That shift is intentional.

AI is powerful. And power, when understood clearly, is something to manage — not something to fear.

REMEMBER

There really is no need to fear AI or even to restrict it. You absolutely should embrace AI — just do so with your eyes fully open to the security needs. What's important is to understand AI's capabilities and the way it works, anticipate the risks that go along with its innovations, prepare to mitigate those risks, and ultimately gain the confidence to use AI bravely.

Clarity is the first step toward that confidence — it's achieved through understanding how AI works and what the vulnerabilities are. That understanding brings into focus the steps needed to move securely into the AI-centric future. Security really doesn't need to get in the way of innovation, and in fact, it makes successful innovation possible.

Yes, the threat landscape will always be there, and yes, it will keep on changing. The ongoing evolution of AI will lead to a continued evolution of threats. You've already heard of many threats that didn't even exist a few years ago, and you can bet you'll be adding more to your vocabulary in the next few years.

That sounds discouraging, but here's a twist to consider: AI itself is coming to its own rescue. AI-powered detection systems will increasingly be the key to analyzing behavioral patterns in real time. AI will be on the front lines, spotting anomalies that humans would never notice, the same way it can spot tumors on medical imaging that radiologists can't yet see. AI will be able to accelerate incident response, too, offering greater speed and unparalleled precision.

As agentic AI makes more and more workflows autonomous, AI will increasingly make security itself an autonomous operation. Protective agents will be there with the human analysts, learning and adapting as the threats adapt.

The future will belong to organizations that engage AI with intention — not by rushing ahead blindly, and not by standing still, but by putting the right controls in place to allow innovation to happen safely. Secure AI is not a constraint — it's the foundation that makes progress possible. Get that foundation right, and AI becomes a durable advantage rather than a risk waiting to surface.

REMEMBER

The goal is not to slow AI down, but to make its progress sustainable. With clear policies, continuous security, and a commitment to learning, AI becomes not a liability to manage, but a capability you can confidently build on.

IN THIS CHAPTER

- Seeing how AI is used
- Changing mindsets about AI security
- Focusing on security throughout
- Watching over supply chain elements
- Keeping runtime safe
- Taking a platform approach

Chapter 6
Ten Questions to Ask Your AI Security Vendor

As shared in the rest of this book, traditional cybersecurity defenses are not up to the task of fully protecting your artificial intelligence (AI) environment as AI systems and generative AI (GenAI) tools proliferate across the enterprise. Security must be embedded throughout the whole AI adoption and development lifecycle.

What kind of approach will effectively secure your organization's AI ecosystem? As you ponder that, here are ten questions that you should ask your security vendor:

- **Do we get a complete, real-time inventory of all AI and GenAI assets being used in the organization, including external AI tools, custom AI apps, and AI agents?** This is essential because effective security starts with visibility. This must extend to every model, agent, data set, and dependency, as well as the unvetted "shadow AI" systems proliferating in the enterprise.
- **Do your tools extend to protect user-facing AI adoption, such as controlling employee access and usage of external GenAI applications?** Security must protect "how

AI is used" as much as "how AI is built." This means the browser has become a new security front door. You must have visibility and granular control over all AI applications used by employees — sanctioned or not — to prevent data exposure and manage "shadow AI" risks.

REMEMBER

- **How does your solution enable a culture of radical transparency and accountability, from the board level down to the developer?** AI security is an executive mandate, not just a technical task for information technology (IT) security folks. Your vendor must support the creation of a culture of radical transparency — for example, providing detailed bills of materials for AI components. The vendor must also facilitate clear accountability for AI-driven decisions, including board-level oversight and auditable logging.
- **How do your solutions enforce the confidentiality, integrity, and availability (CIA) triad across the full AI lifecycle?** The CIA triad is a foundational cybersecurity model for what organizations should protect across the entire machine learning security operations (MLSecOps) pipeline, and it's a key aspect of the Palo Alto Networks unified platform approach. AI security fundamentals require seeing the CIA triad through a modern interpretation. Confidentiality demands robust access control for training data and model code. Integrity requires traceability and defense against data poisoning. Availability means protection against prompt manipulation and resource exhaustion such as distributed denial of service (DDoS).
- **Can your data loss prevention (DLP) solution identify and block sensitive data such as proprietary source code or intellectual property from being inadvertently exposed to AI models?** The model's intelligence is defined by its data, and sensitive data exposure is a top risk. Traditional DLP is often inadequate. You'll need AI-specific DLP that can recognize pattern-based leakage, enforce context-aware policies, and protect intellectual property from both users and model outputs.
- **What specialized tools do you use for continuous, real-time protection against AI-specific runtime threats such as prompt injection and model denial of service (DoS)?** AI is uniquely vulnerable to nondeterministic attacks such as sophisticated prompt injection, which can trick the

model's core logic or override safety measures. Your security vendor needs specialized tools, such as AI firewalls, and AI-aware monitoring to defend inference-time operations.

REMEMBER

- **Does your platform integrate security controls into every stage of the AI/MLSecOps pipeline, from scoping to monitoring?** Security can't be an afterthought; it must be a core principle integrated throughout the MLSecOps pipeline. A platform must ensure that security controls are woven into every phase — scoping, data preparation, model training, testing, and deployment — to eliminate critical vulnerabilities.
- **How do you protect against supply chain risks introduced by pretrained models, open-source dependencies, and external components?** The security posture of your final AI system is determined by its weakest link, and that often involves external elements. Your vendor must offer specialized security measures such as model code signing and scanning for vulnerabilities in pretrained components to establish an auditable, trustworthy supply chain.
- **How does your solution continuously perform adversarial testing and red teaming on AI systems, including autonomous agents?** AI models are susceptible to deception, so continuous, specialized red teaming is required to validate that ethical and business guardrails can't be bypassed. For autonomous AI (agentic AI), testing must cover interactions between components to ensure safe failure under direct adversarial attack.

REMEMBER

- **Is your security solution a unified platform capable of defending the entire AI lifecycle, or is it a collection of point products stitched together with no unified management?** The expanding, complex AI attack surface demands a holistic, platform-centric security architecture. That's the way to cover it all, to protect both "how AI is used" and "how AI is built." Traditional security tools designed for deterministic systems can't keep pace with dynamic, probabilistic AI. That makes a specialized, integrated platform essential for resilience.


PRISMA AIRS™
Let your AI innovations energize the world.
Prisma AIRS™ enables today's leaders to bravely reimagine their industry.
paloalto NETWORKS
deploybravely.com

Unify AI security to confidently win the race

Artificial intelligence (AI) is the bus your business can't afford to miss, but AI threats are risks you can't afford to take. The risks have new names, but more important, they're beyond the reach of traditional cybersecurity. *AI Security For Dummies* outlines why you must build secure AI from the start with a platform approach. Learn how to gain visibility, control access, optimize red teaming, monitor runtime, protect data, govern humans and AI agents — and embrace the future with confidence!

Inside...

- Recognizing the AI threat landscape
- Educating leaders of the business risk
- Governing AI and human interactions
- Building in security (not bolting on)
- Controlling autonomous, agentic AI
- Ensuring full AI security at runtime
- Trading point products for a platform

paloalto NETWORKS

Steve Kaelble is the author of many books in the *For Dummies* series, and his writing has also been published in magazines, newspapers, corporate annual reports, and coffee table books. When not immersed in the *For Dummies* world or writing articles, he engages in healthcare communications.

ISBN: 978-1-394-35737-6
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.