LEARNING MADE EASY

Cyera Special Edition

AI Security

for dummies®

A Wiley Brand

Understand how AI puts data at risk

Discover, govern, and protect your AI data

Measure and boost AI data security maturity

Brought to you by

cyera

Sol Rashidi
Zain Malik
Kelsey Pierce

About Cyera

Cyera is the industry-leading AI Security Platform, empowering global enterprises to adopt AI at scale by securing the data that fuels it. The platform provides precise visibility and intelligent controls to protect data at rest, in motion, and in use, whether accessed by humans or AI agents. Cyera is trusted by a growing number of Fortune 1000 companies. Learn more at cyera.com.

AI Security

Cyera Special Edition

by Sol Rashidi
Zain Malik
Kelsey Pierce

for dummies®
A Wiley Brand

AI Security For Dummies®, Cyera Special Edition

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

ISBN 978-1-394-44127-3 (pbk); ISBN 978-1-394-44128-0 (ebk); ISBN 978-1-394-44129-7 (ebk)

Publisher's Acknowledgments

Writer: Steve Kaelble
Project Manager: Rebecca Senninger
Acquisitions Editor: Traci Martin
Senior Managing Editor: Rev Mengle
Client Account Manager: Jeremith Coward
Production Editor: Athiyappan Lalith Kumar

Introduction

The whole world is in agreement that artificial intelligence (AI) is a game-changer. It's having an astonishing impact on the way organizations of all kinds operate, and has settled into everyday life for most people, whether they realize it or not.

Despite AI's revolutionary potential, many enterprises mistakenly treat AI security as no different from standard IT security. The fact is, AI brings a host of security challenges that traditional security tools simply weren't designed to address.

AI security risks can be serious, and they often surface in places you're not actively monitoring. Many organizations are using AI tools in ways that are not fully visible across the enterprise, which makes those risks harder to anticipate. The good news is that with the right visibility and controls in place, you can adopt AI thoughtfully and reduce exposure before it becomes a problem.

The key to securing AI is visibility and control of your data. Data, after all, is the fuel for AI. Without trusted data, with the right access control and governance, AI is not possible.

Thrive in this new world by adopting a careful approach to discovering, governing, and protecting the data powering your AI, monitoring it faithfully and promptly addressing any issues, and proactively building trust with all stakeholders. Secure your data first, use tools and concepts created with AI in mind, and build a foundation to enable the miracles of AI without the mayhem you fear.

About the Book

This book provides a practical approach to AI data security. It's packed with details about how we've gotten to this place in the AI story, the different ways AI may be deployed across your organization, what the risks are, and how you can avoid them.

Read on for details on the AI data security path, including five key steps along the road to security: AI discovery and classification,

governance, technical protection and control, detection of and recovery from AI troubles, and assurance that your AI systems are secure and in compliance. There's advice on assessing your company's AI data security maturity and setting maturity goals, plus a list of ten AI data security problems and proactive ways to prevent them.

Icons Used in This Book

Check the margins of the book for these guideposts:

REMEMBER

We've kept the book short, but don't hurry through it so fast that you miss the key points marked by this icon.

TIP

The point is to give you helpful takeaways for AI data security, and this icon identifies one of those bits of advice.

WARNING

You don't need to be warned about the potential dangers ahead, but this icon points to a specific caveat to remember.

Beyond the Book

We hope you finish the book feeling less nervous and more inspired, eager to get out ahead of the potential AI data security risks. Check these resources to learn more:

- `Cyera.com`: The insights in this book are from the experts at Cyera, creator of a data security platform that's all about AI. Learn about AI-enabled Cyera Data Security Posture Management, AI Guardian, and Data Loss Protection, and access many other helpful resources.
- `AI security self-assessment`: A quick way to review your organization's AI data security maturity.
- `Certified Security for AI Fundamentals`: An engaging and enlightening online certification course from Cyera.

IN THIS CHAPTER

- Reminiscing about AI history
- Reviewing AI categories
- Seeing AI data security as essential

Chapter 1
Fueling AI with Data

Without good data, and lots of it, we would not have artificial intelligence (AI). AI lives and breathes data, greatly increasing data's value while also putting it at significant risk. Meticulous, intelligent data security is essential because you must simultaneously lock data down while setting it free to work its magic. This chapter sets the stage with some AI history, level-sets with AI background, and spells out the vital nature of AI data security.

Riding the AI Wave

For many people, AI has been a hot topic for a few years now, pushed prominently into the headlines and popular culture when ChatGPT became publicly available, free for the asking. Suddenly, and perhaps for the first time, regular folks were able to get a sense of how game-changing AI technology could be. Game-changing in a good way, that is — science fiction had for years been filling imaginations with a much more dystopian take on the potential impact of AI.

REMEMBER

Of course, those more attuned to technological details know AI has been around for decades, at least as far back as the 1950s, when Alan Turing famously pondered whether machines can think and proposed his Turing Test for judging the behavior of machines.

In the 1960s and ’70s, researchers pushed forward with rule-based systems that worked only if inputs and context were tightly controlled and predictable. Expert systems and neural networks broke new ground in the 1980s and ’90s, moving computers from being programmed with knowledge to being able to learn it through spotting patterns. The internet brought an explosion of data in the 2000s, and as computing power also accelerated, it became possible to train computer models with massive volumes of information. Machine learning (ML) was breaking new ground all the time.

And in 2022, ChatGPT arrived on the public scene, joined before long by other monikers vying to also become household names, such as Copilot and Gemini. As regular folks dove in, so did organizations and enterprises, creating and tapping into large language models (LLMs) to boost practically every industry.

REMEMBER

Generative AI — which listens and answers in natural language, and can pretty miraculously create verbal and visual content on its own — is astonishing enough. But the generative AI explosion has been rapidly followed by the crazy-fast adoption of autonomous AI agents that don’t just create, don’t just respond, but go out and take actions. AI agents aren’t just tools, but virtual teammates.

Indeed, AI has very quickly morphed from an abstract concept to behaving like a colleague. Enterprises have rapidly embraced its capabilities as productivity accelerators that fit right in with everyday work. It’s operational across the enterprise, influencing decisions, acting faster than humans, and touching sensitive data.

Here are some stats from early 2026 that are astounding, and keep in mind that exponential AI growth will likely cause the numbers to become ever more jaw-dropping.

- AI usage has multiplied 61 times over 24 months.
- Nearly 72 percent of enterprise AI tools are classified as high or critical risk.
- Nearly 84 percent of enterprise data is already flowing through these risky tools.

Notice that we said “tools” in the plural form. AI isn’t just a single project or platform, but a widely distributed way of life. Assistants are built into software-as-a-service (SaaS) platforms, AI bots are chatting with customers, copilots are helping out all over

the organization, and analytics systems are analyzing at lightning speeds. AI capabilities are maturing and growing rapidly — the challenge is ensuring that the structures for securing them are able to keep pace.

REMEMBER

Suffice it to say, the miracles are happening all over, and so are the increasing risks. We'll talk more about risks in Chapter 2, but it's important to underscore a key point. The aim here isn't to spread fear, or persuade you to batten down the hatches and toss your data into an impenetrable vault and throw away the key. The future involves opening up data to exciting possibilities, while also keeping it safe.

Getting to Know AI

Let's outline some of the technologies and terms we'll be discussing in depth. It can, in fact, be a bit of a challenge to clearly define what "AI" really means in practice, because it can be deployed in many different forms that carry very different risk profiles. For starters, here's some detail on the broad categories describing ways AI might be implemented:

- **Public AI:** This generally refers to AI tools that can be accessed over the internet directly. That, of course, includes the generative AI apps that have taken the public by storm, such as ChatGPT, Gemini, Claude, Perplexity, and Canva. These tools are great because they are easy and fast to adopt, but they sit outside of your organization's direct control. As such, if you're not careful, the data entered into these tools could end up being retained or processed in ways you don't know about. Public AI is the big reason you may have shadow AI, which is unapproved and quite possibly unknown AI usage that brings real data risk.
- **Embedded AI:** This is AI you'll find inside other software platforms. Copilot is a good example, because it adds AI power to a lot of Microsoft applications. You'll find AI inside productivity suites, analytics tools, help-desk platforms, customer-relationship management systems, and other platforms. Embedded AI can have stronger administrative controls, which is good. Still, many organizations don't fully understand or review the configuration choices that can impact data access, output handling, and memory behavior.

- **Homegrown AI:** As the name suggests, these are AI systems built or assembled by organizations themselves. Could be internal chatbots, copilots, retrieval-augmented generation (RAG) systems, or autonomous agents living on cloud platforms or at home on self-hosted models. Enterprises own the data pipelines, integrations, memory, prompt logic, outcomes — and most definitely, the security.

Secure AI Starts with Secure Data

If you wanted to boil AI security down to its most basic principles, one of the biggest ideas is that data is key. AI is built on data, trained on data, predicts and acts because of what it has learned from data, and in the work it does, it continually interacts with data and often impacts data. Data is everywhere along the AI security journey.

That's why AI security starts with data, and why securing data is nonnegotiable. In traditional applications, data has followed generally predictable paths, but AI tends to blur the boundaries. One prompt can bring together data from all over, from documents and application programming interfaces to memory. It can generate an output that is new data, and that new data might even be used as a future input down the line.

Security must understand where data enters the AI lifecycle, how it's classified, and how it's governed. Security isn't just bolted onto a chatbot, but woven into the lifecycle and embedded into AI design. It grows along with AI adoption, adapts as systems change and add capabilities, and keeps watching as AI becomes more autonomous.

WARNING

Data security in AI is nonnegotiable because the risks are real and vast. Poor governance can lead to leaks of regulated or proprietary data, and once sensitive data ends up erroneously in a model, it can be problematic to fix. Over-permissioned systems can open the door to improper users and uses. Analyses can be flawed if data is mismanaged.

In the end, data problems devolve into trust issues. And stakeholders will only embrace AI that they can trust.

IN THIS CHAPTER

- Counting up the potential troubles
- Becoming aware of what lurks in the shadows
- Learning the new ways data is at risk
- Understanding risks at different points of the lifecycle
- Seeing risks in the AI supply chain

Chapter 2
Understanding AI Data Security Risks

Back in simpler times, it wasn't hard to tell when things were heading south. Cars simply wouldn't start. Component shipments were delayed. Computers crashed. If you had a problem, you knew it. AI systems have problems, too, but they aren't always as obvious. They can experience and create risk, and go about it very quietly.

This chapter focuses on the many kinds of risks that can threaten AI systems, beginning with the wisdom of one prominent top ten list of risks. Keep reading to learn more about shadow AI, explore novel new threats to data, find out the different concerns that exist at different points in the lifecycle, and wrap your head around supply chain risks.

Imagining What Could Go Wrong

What could go wrong as AI taps into vast stores of data to learn, references that data in its outputs, and in the case of agentic AI, takes actions that might impact data? Let us count the ways.

Or maybe better yet, let OWASP count the ways. The Open Worldwide Application Security Project has been creating Top Ten lists for more than two decades.

WARNING

Here are some of the vulnerabilities OWASP has put into the spotlight related to large language models (LLMs).

- **Prompt injection:** This happens when a user prompt intentionally messes up the LLM's behavior or outputs. Prompt injection can cause a model to create harmful or biased content, violate guidelines, or even open the door to unauthorized access.
- **Sensitive information disclosure:** Data makes AI happen, but serious problems can arise if an LLM mishandles sensitive data. We're talking about personal identifiable information, health records, financial data, business secrets, security credentials, and other things that are supposed to be confidential.
- **Supply chain risks:** Creating LLMs often involves tapping into third-party models. They are part of the supply chain that allows AI to function, but they can also create vulnerabilities that can lead to system failures, breaches, or bad outputs.
- **Data and model poisoning:** This happens when vulnerabilities, biases, or backdoors are introduced into pretraining, fine-tuning, or embedded data. Capabilities can be impaired, outputs can be messed up, and ethical behavior can be thrown out the window.
- **Excessive agency:** LLM-based systems are often granted some level of agency, letting them call functions or otherwise interface with other systems. Excessive agency is when some sort of unexpected or manipulated output is able to cause damaging action.

Stirring Trouble in the Shadows

WARNING

One of the best things about today's AI technologies is also one of the most challenging. It can be so easy to adopt, even for non-IT folks, that eager trailblazers often cut a path into dangerous territory without even realizing it. The risk we are talking about is

shadow AI, which refers to AI systems that are in use in the enterprise but without any formal approval, no governance, and little to no visibility.

Shadow AI could be employees using public generative AI tools through browsers, teams turning out their own experimental models and agents, or business units that are using AI features built into a software-as-a-service (SaaS) platform, without doing any security review.

It's often hard for your people to know they're doing anything wrong. They're probably boosting productivity, and likely feeling pretty good about it. But as you'll read in various parts of this book, proper governance is essential, and you can't govern (much less secure) something you can't see. Shadow AI is a sign that usage is growing faster than governance, and it likely puts sensitive data at risk.

The risks vary from one shadow to the next in this world of uncontrolled deployment scenarios. You can imagine how much trouble might be caused by an autonomous agent triggering workflows or updating records. But using a public chatbot to brainstorm new ideas is risky, too, because sensitive data can easily find its way into prompts. Or, plug a RAG system into an internal knowledge base, and you could mix up a batch of misinformation.

Messing with the Data

Just a few years ago, not too many people outside of tech gurus had ever heard such terms as data poisoning, prompt injection, or model inversion. These are all AI-related concerns in which security problems threaten not just the infrastructure but the data itself.

Data poisoning can target systems during training or retrieval. The title is really pretty self-explanatory — it involves injecting training sets or knowledge sources with data that's misleading, biased, or downright malicious.

The result may not be as dramatic as when a movie villain sneaks a few drops of something poisonous into a martini and the victim immediately keels over. This poisoning might not cause any noticeable outward symptoms at all, which, when you think about

it is even worse, because data poisoning may leave the model up and running while causing unsafe actions or incorrect conclusions.

WARNING

As long as we're talking movie comparisons, imagine *prompt injection* being like when a character is secretly hypnotized and is caused to do all the wrong things. The result is often comedic in the movies, but real-life prompt injection isn't the least bit funny. Attackers craft malicious inputs, sometimes in prompts themselves, sometimes hidden in documents or web pages, and those inputs override system instructions or safeguards. The result could be inappropriate responses, disclosure of sensitive information, or misuse of tools.

We'll finish this cinematic journey by thinking of *model inversion* like a brilliant Sherlock Holmes or Benoit Blanc who has gone to the dark side. This kind of attack meticulously analyzes model outputs, using careful and often repeated queries to reconstruct sensitive attributes or infer whether certain private records or information were in the training data. It's reverse engineering a functioning model to turn it into a crime scene.

REMEMBER

These kinds of threats are often enabled in part by trust in data pipelines. AI systems generally trust that prompts and training data and retrieval sources are benign, and they usually are. When they aren't, a model can have trouble with data integrity and confidentiality. Unlike the victim who consumes a poisoned martini and drops dead, there might not be any obvious signs of compromise.

Spotting Supply Chain Risks

AI systems rarely operate alone, but work their magic through a supply chain that features such things as data sources, foundation models, plugins, orchestration frameworks, application programming interfaces (API) that let agents take action, and various third-party services. It's a sophisticated and integrated chain of players and information sources. And every link in the chain brings with it a variety of risks.

Let's talk about foundation models for a moment. They can be essential for AI systems, but because they are often developed and hosted by external organizations, you may not have

full transparency into training data or such operational details as update schedules or retention practices. Tapping into open-source models and libraries may bring in information without a lot of vetting. SaaS platforms may send data through third-party models.

Agents and plugins also enrich the system and complicate the supply chain while expanding the attack surface. Plug in a plugin, and it could grant an AI system access various things such as file storage, external APIs, ticketing systems, and more. What if that component is a bit too permissive?

Supply chain risk is tricky for a number of reasons. As described above, it can expand the blast radius. But there's also a risk that comes from the unknowns of depending on an outside supply chain over which the organization has little control. A compromised data source, a dependency update, or some random change made by a vendor — all of these things can impact the AI systems to which they're linked.

IN THIS CHAPTER

- Discovering and classifying your AI assets
- Governing your AI activities
- Establishing protections and controls
- Detecting and mitigating trouble
- Validating and assuring your security

Chapter 3
Traveling the AI Data Security Path

The first couple of chapters set the stage for where we've come with AI, along with the many new risks that accompany the exciting rewards. You need more than a simple checklist to move forward safely. AI data security isn't a task to complete, but a journey to travel — a journey that guides you through the discovery, classification, and governance of data, on your road to data intelligence.

This chapter spells out how that AI data security journey should look. Read on to learn five distinct stages of this journey: discovering and classifying your AI assets; governing your AI systems; protecting and controlling them constantly; detecting issues and recovering from them; and gaining assurance by validating that your systems are secure, policy-aligned, and ready for any necessary audits.

Knowing What You Have

It's often said that you can't secure what you can't see. No wonder so many people surround their homes with security cameras. But cameras only give you extra eyes. You still need to determine

whether that thing you see on the screen is the purring cat from next door or a stinky skunk or a mischievous fox, and whether that human image is a teenager sneaking out or an intruder sneaking in.

Same goes for your AI assets. There's plenty of potentially malicious activity to be concerned about, but one of the big initial risks is simply being unaware of where your people are using AI and what data that AI touches. The first step on the journey is to discover, classify, and inventory your assets. Shine a light into the shadows to really see what's there.


TIP

Asset discovery means finding and identifying every AI system across the enterprise. The list should include all the officially approved AI tools, the AI functionality embedded into lots of apps, and all the shadow AI tools that employees are tapping into on their own.

You will likely be amazed when you learn just how much AI is happening outside of your traditional security oversight. Oh, it's out there, interacting with sensitive enterprise data, but your traditional security tools aren't giving you visibility. That means you can't effectively enforce policies, uncover misuse, or keep data from leaking.

Once you've identified AI systems, you must figure out what data they're accessing. You need to classify structured data, such as customer databases, financial records, and operational metrics that you're gathering. You also need to classify unstructured data, from emails to chats to documents to source code.

Unstructured data is likely to include sensitive info, and because it's unstructured, it's often ungoverned. It's also ripe for the picking by retrieval-augmented generation (RAG) tools. Embedded AI tools often peek into poorly governed data sources.

Mapping AI systems to the sensitive data they touch helps you fully understand AI's potential to dig into intellectual property, personal data, confidential business files, or regulated information. Prioritize your risk by correlating AI assets with data sensitivity levels.

Discovery and classification are made all the more complicated by the fact that enterprises operate all over the place — on-premises, in the cloud, across software-as-a-service (SaaS) platforms.

AI-native discovery and classification tools are what's needed to make sense of it all, continuously and in real-time.

Setting the Rules

That first step helps you know where things stand and offers hints at what kinds of rules you need to protect your data. The second step is governance, which is a fancy way to say making the rules.

REMEMBER

By establishing solid governance, you're turning the visibility you've gained into the accountability you need. Governance spells out who owns what AI system, what the rules and acceptable uses are, how deployment decisions are made, and who gets access (and note that "who gets access" could be people or other systems).

Clear usage policies are the start of effective AI governance. They spell out the who, what, when, why, and how of AI usage. What kinds of tools can employees use? What kinds of data can be run through those systems? What business processes may rely on AI-created outputs? In the case of embedded or homegrown AI, your policies may include approval workflows and risk reviews, and a full documentation of acceptable use cases.

Your governance process should include creating AI risk registers and automated assessment workflows. Once each system has been cataloged with its purpose, its owner, its risk tier, and its acceptable data types, you can more easily track which models are dealing with sensitive info. That helps you determine and prioritize the controls that you'll build in the next step.

TIP

Another benefit of good governance is that it causes people to stop and take a breath before diving into AI. Face it, AI is fun and exciting, and people often dive in without even really thinking through whether there's a valid business purpose (the "why" we mentioned earlier). Just because something is new and popular doesn't mean it makes sense to adopt. Line it all up with strategic goals and measurable outcomes before giving the green light.

Governance includes identity and access management. Your controls in the next step will likely include such things as role-based access controls (RBAC), but you need governance to establish how those controls are set. Use the governance step to align AI access with job function and data sensitivity.

The policies you put in place need to be adaptive, adjusting based on data attributes such as sensitivity, criticality, and usage patterns. No more creating static rules and posting them on SharePoint — your governance needs to be informed by real-time data monitoring and be able to expand policies based on what it sees.

As you figure out the rules, you must consider what rules and standards the outside world applies to your enterprise. All kinds of different compliance frameworks might apply, depending on what business you're in or who your potential customers are.

You may need to be thinking about the National Institute of Standards and Technology Artificial Intelligence Risk Management Framework. Or the Health Insurance Portability and Accountability Act. Or the European Union AI Act. Or ISO 42001. Or something completely different.

And as you consider those external powers-that-be, you also must establish the internal governance structures. Who exactly is in charge of making and enforcing these rules? Lots of organizations these days have AI governance committees that bring together people from legal, security, information technology, compliance, data management, and the various business units that are using the AI.

Protecting Your AI Systems

By now, you may be seeing how this resembles a journey, with growth steps along the way. You discover and classify your AI, then figure out what the rules should be. Next comes the actual work of establishing the proper controls and protecting your data and AI systems.

You're turning the thought exercise of policymaking into action, your protection strategy into reality. The big idea is to put appropriate limits on what data AI systems can see, what they can do, how their outputs are handled, and that kind of thing.

When it comes to controls, identity and access control is the most obvious place to begin. You need RBAC and strong identity management to be sure that your AI models, capabilities, and datasets are open only to authorized users, agents, or services.

TIP

You have undoubtedly already had plenty of conversations about the principle of *least privilege*, through which users and systems get only the permissions they need for their roles and tasks, and nothing beyond that. It's a vital concept for AI, too. It's one way that you keep a model or agent from digging into broad datasets of sensitive info, or executing high-risk actions without proper oversight.

You'll also be tapping into encryption and application programming interface (API) security. The more the world moves into agentic AI, the more AI systems are interacting with data through APIs, external services, and retrieval pipelines. You need careful management of API keys and tokens, along with encryption of data at rest and in transit.

Data loss prevention (DLP) is another important concept with which you are likely already familiar, and it's just as important with AI. Especially important, in fact, because AI systems are not just ingesting sensitive information but also generating it, so you need protections on both the input and output sides. DLP solutions must be able to inspect training data, prompts, and generated responses. They need to be on the lookout for any kind of sensitive information, ready to flag any violations or concerns.

You'll need multiple layers of defense to protect against prompt injection. That includes filtering, input validation, context separation, and well-defined controls over the tools and data sources accessible to the model.

Plan to build AI firewalls and runtime protection mechanisms to help enforce your safeguards. These controls keep an eye on prompts and responses, aiming to enforce safety policies before bad data reaches the model, and before unsafe outputs are delivered to users.

TIP

Strong data validation and output controls are essential. You must validate data fed into training pipelines and retrieval systems, while filtering improper outputs and maintaining audit trails. Your controls should build a closed loop of protection, covering inputs, actions, and outputs.

Staying Alert and Responding

REMEMBER

You can sleep better knowing that you have mapped all your AI assets, created thoughtful and complete governance, and built the strongest possible controls. That doesn't mean your protection gets to sleep, though. The next part of the journey is where all the action takes place. Your systems must be constantly detecting and on the lookout for potential problems. And they must promptly respond to any alarm bells and start the vital work of mitigation or recovery, as needed.

Your AI security must be monitoring AI activity in real time, staying alert to how AI systems are behaving. This "detect and recover" phase is all about identifying unsafe, unexpected, or degraded AI behavior and then containing it before the risk spreads any further. Your aim must be to catch issues such as drift, policy violations, or misuse early on, before they have the chance to lead to an operational failure, data leak, or compliance issue.

This stage is a combination of observability, threat detection, and carefully written response playbooks. Comprehensive logging is a foundation for AI observability, giving security teams visibility into user activities, prompts, tool calls, retrieved content, and model responses.

Audit logs help build a picture of day-to-day AI operations. They give insights into how models decide to take a certain action or deliver a particular output. With evidence in hand of which data sources are involved, investigations are more successful. And strong observability comes in handy at compliance time, when you're called upon to demonstrate faithful operations and rule-following.

TIP

Build on this observability with real-time threat detection. This means keeping watch over prompt patterns, unusual usage patterns, or tool calls that are out of the ordinary. Constant watch helps set a baseline for what is normal, so that it's easier to spot any abnormal behaviors that could indicate prompt injection attempts, data exfiltration, or other concerns.

Model usage analytics are an important part of defining the normal and spotting the anomalies. Analytics also help your governance teams refine policies and help leaders optimize performance.

Another key focus of the "detect" phase is identifying model drift. Model outputs can, as they age, become less accurate and less aligned with policy, and in some cases, more prone to hallucination. Drift doesn't necessarily mean someone's intentionally messing with the model. It can happen as vendor models are updated or upstream data evolves. Whatever the cause, it's vital to continually evaluate error rates, policy compliance, and overall output quality.

As we noted earlier, this step in the journey has two parts. Once you've detected something, you must respond and recover as needed. You simply have to expect that things are going to happen, and be ready when they do.

That's where a well-defined incident response plan comes in. Again, this is not exactly a new idea, as you likely already have detailed cybersecurity incident response plans for traditional threats. You just have to add some new ideas and categories when you bring AI into the mix. You're dealing with things your earlier plans never imagined, such as prompt injection attacks, AI responses that inadvertently expose data, hallucinations, or the misuse of agents or tools.

TIP

As is required by other cybersecurity situations, clear playbooks must define what constitutes an AI security event, who handles it, and what must happen to contain and remediate the issue. If inaccurate or policy-violating outputs start emerging, for example, your playbook might call for temporarily disabling the offending feature, or cutting off user access, or reverting to some fallback.

Your recovery plan may include maintaining versioned models and configurations, in case rollbacks are part of the recovery. Your response teams must be able to revert to a known-safe version if a model starts to act strangely, hallucinate, or drift. You might need to pause integrations, disable certain plugins, or reset agent states.

Incident response also needs to include breach readiness. AI incidents can trigger legal or regulatory obligations, which means that any AI-related data exposure must be treated just like any other breach. Notify stakeholders, preserve logs, investigate root causes, and report whatever you are required to report. Indeed, some of today's emerging AI regulations are counting on you to

not only maintain preventive controls, but also plan for how you will respond to incidents and maintain accountability.

Ultimately, as much as you work to prevent troubles, you're still bound to experience incidents of some kind. It's an IT fact of life. Preparing to respond is all about resilience. Complex systems are going to fail at some point or another, and the important thing is how you react, how quickly you recover, and how well you maintain trust going forward.

Assuring That All Is Well

Speaking of maintaining trust, that's the whole point of the fifth and final stop on our AI data security journey. Yes, you must build strong security controls, monitor continuously, and respond effectively. But ultimately, your ongoing success depends not just on how well you do those things, but how well your users, customers, executives, partners, and any regulators *believe* you are doing those things.

REMEMBER

The name of the game for this final stop is assurance. You must prove that your AI systems are trustworthy, compliant, resilient, and behaving the way they're supposed to behave. Assurance is turning your security and governance efforts into verifiable evidence of your good work.

Assurance begins at the beginning of the journey. You must plan a structured evaluation before you ever deploy, with formal conformance exercises and red-teaming to check out your accuracy and safety. And when we say "structured," we mean these evaluations need clear signoff criteria.

That's the beginning, but far from the end. Assurance isn't a one-time thing. Because your AI systems evolve continuously, your evaluation and documentation must be ongoing. Assurance includes logging model behavior, tracking all the metrics for performance and safety, refreshing risk assessments from time to time, and maintaining model cards. Stale evidence becomes less and less assuring as time passes.

Assurance also puts a spotlight on data integrity and provenance. Be able to trace where training data and retrieval content come from and how they influence model outputs. Lineage helps

explain outcomes, get to the bottom of incidents, and demonstrate compliance.

Ownership is also a key element of assurance. You need defined roles for approving AI deployments, responding to audit findings or regulatory inquiries, and maintaining the necessary evidence. Clear accountability helps make assurance a reality.

Taking an End-to-End Approach

Your attack surface keeps expanding as you scale AI across the enterprise, covering everything from your data sourcing to the model outputs and the agent actions. You need an end-to-end approach to be sure each part of the AI lifecycle is covered and secured. The journey we've outlined in this chapter helps ensure that happens effectively, and to put a fine point on it, it's worth reviewing a complementary framework that adds some helpful specifics to what we have outlined so far.

This is what Cyera refers to as the 7-Phase AI Security Framework, outlining seven pillars required for securing your ecosystem from one end to the other. Each pillar identifies focus areas specifically and is equally specific in spelling out key controls required for security.

- **Data sourcing:** The focus here is discovery, classification, provenance tracking, legal rights, and vendor risks. Controls to adopt include a data registry, an inventory, and a compliance checker.
- **Infrastructure:** As you focus on segment networks, encrypt flows, and container security, adopt such controls as terraform configs, key management system, transport layer security, and virtual private cloud isolation.
- **Data-in-transit:** In this phase, focus on masking, anonymization, and differential privacy. Key controls include detection of personally identifiable information, as well as synthetic data checks.
- **API:** Focus on authenticating access to models and enforcing rate limits. Helpful controls here include RBAC, token validation, and input sanitization.

- **Model provenance:** This pillar focuses on detecting prompt injection, other adversarial behaviors, and unsanctioned uses. Output filters are helpful here, as well as other guardrails and controls on RAG.
- **Incident response:** Real-time alerts, a focus on forensic lineage, and rollback plans are the points here. Controls include automated breach response and alert triage.
- **Continuous monitoring:** Here's where you keep a long-term lookout for drift detection, conduct usage audits, and complete compliance logging. KPI dashboards and retraining triggers are vital controls here.


TIP

If these pointers have given you a greater sense of purpose related to AI security, you can find even more in *Scaling AI: The AI Governance and Security Playbook for Executives*, written by Cyera experts Sol Rashidi and Steve Klementowski, available for reading on Kindle.

IN THIS CHAPTER

- Seeing why maturity models are important
- Assessing your AI security maturity
- Setting goals for AI security

Chapter 4
Assessing Your AI Security Maturity

It's safe to assume that because you opened this book, you're interested in beefing up your organization's AI security. As we've said before, this is a journey, not a one-time thing, in part because there will always be constant changes in models, data, and use cases — and partly because there's always room to get better.

That's where maturity modeling comes in. It's a way to assess where you stand today and where you need to go. This chapter offers more details on maturity models, assessing your enterprise, and setting goals.

Understanding Maturity Levels

AI security maturity models are frameworks that tell the story of this journey in terms of structure, automation, and effectiveness. This work isn't about achieving perfection, but enabling continuous progress. Figure 4-1 shows a framework to consider for measuring AI security maturity.

Security for AI Maturity Model

Component	Level 1 Initial/Ad-hoc	Level 2 Defined	Level 3 Managed	Level 4 Quantitatively Managed	Level 5 Optimizing
Discover & Classify	No inventory of AI tools or data use. Shadow AI and untracked models prevalent.	Inventory started. AI tools and data flows partially documented. Classification ad-hoc.	Centralized inventory and standardized classification. Sensitive data and tools tagged. Shadow AI reduced.	Discovery automated. Data and model usage metrics tracked. KPIs assess inventory completeness and drift.	Inventory self-updating. AI detects usage anomalies and adjusts classification dynamically.
Govern	No AI usage policy. Ownership undefined. AI use unregulated.	Acceptable Use Policy drafted. Risk register begun. Governance roles identified.	Approval workflows include third-party AI. Owners assigned to internal/external models. Risk register maintained.	Risk posture metrics inform policy updates. Governance systems enforce role-based reviews and thresholds.	Governance decisions dynamically updated via AI. Risk tiering, policy routing, and approvals adapt in real time.
Protect & Control	Minimal RBAC or technical enforcement. No model boundaries.	RBAC defined. Some prompt controls and sandboxing applied.	IAM and plugin access controls implemented. Security audits scheduled.	Threat metrics drive controls. RBAC and boundaries updated based on behavior insights and incident KPIs.	Protections adapt in real time. AI-driven enforcement adjusts tokens, tools, and scopes at runtime.
Detect & Recover	No logging or rollback. Misuse undetected.	Logging and rollback paths partially defined. Alerts reactive.	Logging and QA baselines established. Escalations and anomaly detection tied to misuse.	Drift, bias, and hallucination metrics integrated into SIEM. KPIs monitor agent behavior and response speed.	AI-assisted forensics and rollback. Systems auto-freeze unsafe sessions and retrain or redeploy models automatically.
Assure	No QA or documentation. Black-box risk unchecked.	Safety testing defined. Some baselines and evaluations applied.	Standardized evaluation protocols. Risk tiers aligned to go-live requirements.	Conformance KPIs scored across all deployments. Metrics enforce consistent baselines before launch.	Continuous assurance loops. AI tracks, evaluates, and remediates assurance gaps in real time across models.

FIGURE 4-1: Assess your organization's AI security maturity with this framework.

Assessing Yourself

REMEMBER

It doesn't matter what you're trying to improve — fitness, pickleball skills, cooking capabilities, parenting — the best first step is an honest self-assessment. For this important exercise, you should refer to Chapter 3, where we establish five distinct parts of the AI security journey.

One important thing to understand as you set out on that journey: For each stage we outlined in the previous chapter, your organization has likely achieved some of the important elements but has work to do on others. This assessment is all about figuring out what work has been done for each stage and what work still needs to happen.

In other words, for each of those five stages, you should be able to determine where on the maturity scale you are right now — somewhere between level 1 and 5. And don't worry, there's no shame in not scoring a 5; this book is about improvement.

Setting Maturity Goals

REMEMBER

Once you have assessed where you are in each stage, you must set realistic targets for where you need to be, so you can see the gaps that must be filled. While 5 is the top maturity level, that doesn't mean your initial goals must be set at level 5 for every stage. Your priorities should be influenced by real-world factors such as your organization's risk tolerance, budget, and compliance requirements.

Here's an example of how this works in practice. A healthcare organization has some serious compliance requirements with regard to patient data. So when it comes to both controls and detection, that reality necessitates a maturity goal of 5 when considering industry and regulatory environment. It's also going to get a 5 under risk appetite, because data exposure can be very costly, and strong controls are vital. A less regulated startup might be fine with goals in the midrange.

IN THIS CHAPTER

- Gaining visibility, classifying data, enforcing policies
- Adopting AI-specific controls, tools, and practices
- Planning security and response from the start

Chapter 5
Ten AI Security Woes and Wisdoms

We learn from mistakes, but the lessons can be painful. Read on for ten common AI woes, and the wisdom to prevent them.

Flying Blind with Shadow AI

The mistake: Organizations often discover they have ten times more AI applications in use than they knew about — sometimes even more than that. Employees adopt ChatGPT, Microsoft Copilot, and hundreds of other AI services without information technology knowledge, creating the risk of invisible data exfiltration paths. Cyera's 2025 State of AI Data Security Report found 40 percent of organizations have shadow AI operating outside approval and oversight.

The wisdom: Implement continuous AI discovery and inventory management (you can't protect what you can't see). Tools such as Cyera AI Security Posture Management (AI-SPM) identify which AI tools are in use (both approved and shadow AI) and map where they interact with sensitive data.

Treating AI Like Any Other User

The mistake: AI isn't like other users, but organizations often fail to manage AI systems as distinct identity classes with specific access requirements. That can give AI models access to more data than they need, violating least-privilege principles. Only 16 percent of enterprises treat AI as a first-class identity today, even as two-thirds have caught AI over-accessing sensitive data.

The wisdom: Create AI-specific identity and access management policies. Give each AI system a defined scope of access tied to data classification and business context.

Skipping Data Classification Before AI Adoption

The mistake: Enterprises often rush to deploy AI without first discovering and classifying the sensitive data that will fuel models and tools. Unclassified data is the top driver of AI exposure.


TIP

The wisdom: Always discover and classify sensitive data before it enters AI systems. Implement comprehensive data discovery across cloud, software-as-a-service, and AI environments to understand where regulated information resides, how it's used, and who can access it. Data classification makes all other AI security controls possible.

Ignoring the Prompt-Output Interface

The mistake: Sensitive data can flow through the interface between humans and AI — where prompts and outputs are traded. Many organizations overlook this critical security control point, and prompt injection attacks go undetected.

The wisdom: Secure the interface with prompt filtering, output validation, and real-time monitoring. Implement AI firewalls and runtime protection to detect prompt injection attempts, identify sensitive data in model outputs, and block risky interactions before data leaves your control.

Setting AI Policies without Enforcement

The mistake: Some enterprises have AI usage guidelines and acceptable use policies that exist only on paper. Policies are worthless without controls and continuous monitoring. Only 9 percent of organizations currently monitor AI activity in real time.

The wisdom: Implement policy-as-code with automated enforcement. Deploy data loss prevention (DLP) specifically designed for AI, such as `Omni DLP`, with prompt filtering, access control, and data retention rules that execute automatically.

Treating Autonomous Agents Like Simple Chatbots

The mistake: AI agents can remember context, call tools, and take autonomous actions. Many organizations fail to recognize that they require fundamentally different security controls than passive LLMs.


TIP

The wisdom: Implement agent-specific controls, including tool scope restrictions, and communication path validation. Secure Model Context Protocol and other agent communication frameworks. Treat agents as digital coworkers that need identity management, least-privilege access, and continuous oversight.

Assuming Traditional Security Tools Work for AI

The mistake: It's dangerous to rely on existing firewalls, DLP, and cloud access security broker solutions not designed for AI. They miss AI-specific risks such as training data poisoning, model theft, and vector database vulnerabilities.

The wisdom: Deploy AI-native security controls designed for the unique threats, as outlined in the OWASP Top 10 for LLMs, NIST AI RMF, and MITRE ATLAS frameworks. Traditional tools must be supplemented with AI-specific threat detection, model security, and data lineage tracking.

Missing Visibility into AI Data Flows

The mistake: Too many organizations can't trace where training data came from, how it was used, or how it influenced outputs. They lose track of data lineage once information enters AI systems, which creates compliance nightmares and makes incident response nearly impossible.

The wisdom: Implement comprehensive data lineage and access trail capabilities. Track what data AI systems access, when, and how it's used. Facilitate compliance by maintaining audit logs of all AI interactions with sensitive data.

Waiting for an Incident to Build Response Plans

The mistake: Having no AI-specific incident response playbooks is a recipe for disaster. When AI systems hallucinate, leak data, or get compromised, security teams scramble without clear procedures for containment, investigation, or recovery.

The wisdom: Create AI-specific incident response plans before you need them. Define procedures for handling data exfiltration through AI tools, model poisoning, prompt injection attacks, and hallucination-caused impact. Include model rollback procedures and clear escalation paths.

Prioritizing Speed Over Security in AI Adoption

The mistake: It's dangerous to deploy AI without establishing proper governance, believing security will slow innovation. This creates technical debt and risk that becomes exponentially harder to remediate later.


TIP

The wisdom: Build security into AI adoption from day one. Embed security in the design phase using frameworks such as NIST AI RMF's "Govern" function. Start every AI pilot with visibility, classification, and access controls in place.

The New Standard for AI Security. Start Your Certification

Certified Security for AI Fundamentals is a tool-agnostic AI security certification that helps you unlock AI for your enterprise, securely.

Sign up today to gain the expertise and recognition needed to lead AI security and governance programs with confidence. This certification distills lessons from leading enterprise AI initiatives so you can identify AI risks, implement effective controls, and clearly communicate your AI security posture to stakeholders.

cyera

Embrace the AI-driven future by protecting your data

Artificial intelligence is transforming business and life at a stunning pace, enabling possibilities you might never have dreamed about just a few years ago. It's also creating new risks you've never heard of before. There's no need to fear the risks — embrace AI with confidence by securing the data that makes it possible. *AI Security For Dummies* outlines your game plan: discover your AI assets, govern AI activities, protect and control your AI, detect trouble and respond, and assure that your AI is secure.

Inside...

- Seeing why data is key to AI security
- Discovering and classifying AI assets
- Creating wise and effective governance
- Building controls to protect AI activity
- Detecting issues and responding smartly
- Validating the reach of your security
- Assessing and boosting AI security maturity

Sol Rashidi is Chief Strategy Officer for Data & AI at Cyera. A Forbes "AI Maverick," Sol Rashidi is a 9-patent holder and author of *Your AI Survival Guide*. From launching IBM Watson to holding 4x C-Suite roles, Sol bridges the gap between massive data transformations and human ingenuity. She's a Top 100 AI leader focused on technology as an enabler.

Go to Dummies.com™
for videos, step-by-step photos, how-to articles, or to shop!

ISBN: 978-1-394-44127-3
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.