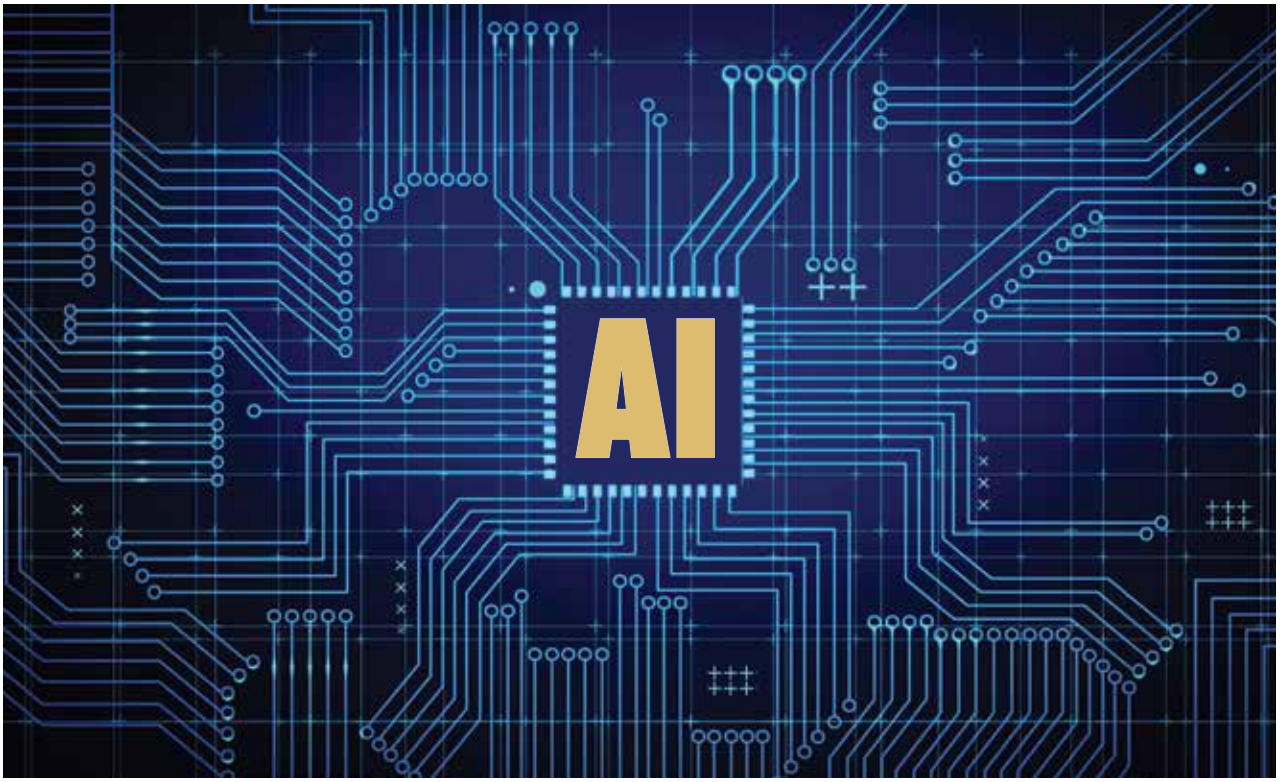
ENTERPRISE AI SECURITY

HANDBOOK



THE TAG ANALYSTS

LEAD ANALYST DR. EDWARD AMOROSO
CEO, TAG INFOSPHERE | RESEARCH PROFESSOR, NYU

TAG

ENTERPRISE AI SECURITY

HANDBOOK

C O N T E N T S

TAG

DR. EDWARD AMOROSO, FOUNDER & CEO

LESTER GOODMAN, DIRECTOR OF CONTENT • DAVID HECHLER, EDITOR

Charlie Ciso

"On the AI-Net, no one knows you're a human."

CHAPTER 1

OVERVIEW

WELCOME! Why in the world would anyone bother spending time writing yet *another* handbook on cybersecurity?

It's a fair question. Just Google the terms—or ask ChatGPT—and you will find enough enterprise security handbooks from vendors, investors, writers, and consortia to keep you busy reading for years. It's a crowded field indeed, this security-handbook-writing racket. And that's before we weave in the critical impact of artificial intelligence (AI), which is our focus in this book.

And yet, in our day-to-day research and advisory work at TAG, we perceive a significant gap in trusted AI security guidance. Worse, we find that the information available on this topic is highly biased. There is a lot of money riding on AI security these days, with investors funding startups to the tune of tens of millions of dollars, creating an environment ripe for misinformation.

The effect is that when tasked to protect the AI that is being used in organizations to cut costs (and let's be honest, AI is being used almost exclusively to cut costs), the CISO has surprisingly few options to choose from for unbiased and competent advice. Vendors, investors, and analysts have inherent biases, and most academics have no clue.

This is why we've developed the book you have in front of you. It was developed in a Zoom laboratory, where our team at TAG met with several dozen AI committees, AI security teams, and other groups to try to understand their plans. And sadly, we found that most weren't sure how to proceed.

After the fiftieth or so discussion, the idea set in that maybe we could help. As we met with each team, we would update what we were learning in an evolving PowerPoint deck that become the repository of the best ideas we'd heard. That slide deck, developed mostly during 2024 and 2025, is the source of this book. It is how it came to be.

PRESSURE TO 'DO SOMETHING' ABOUT AI RISK

One observation that emerged from our work is that security teams are being pushed by leadership, including boards, to *do something* about the risk of AI. Unfortunately, the "something" being requested is not clear. For example, connecting the business outcomes described by boards to the technical AI risks described by CISOs has often been tricky.

For example, if we asked you right now to set this book down for a moment and think through how AI might damage your own organization, you might find this a tougher exercise than one might expect. Go ahead and give it a try. You'll see what we mean. It can be a stretch to connect AI risk to serious business consequence.

Actually, what we found is that security is more often than not put in place to make damn well sure that the cost promises being made regarding AI are not subverted by a bunch of hackers. If the AI is going to fail, we heard repeatedly, then the preference is that it fail on its own, not as a result of a cyberattack.

DECISIONS TO SPEND ON AI SECURITY

All of this is not to say that AI-related cyber threats do not exist for enterprise teams, government agencies, and even citizens. Prompt injection, model bias, jailbreak attacks, and the like are concerns. But the decision to spend money on AI security versus, say, improved identity security, might lean differently if risk quantification were the driver.

And so, during 2024 and 2025, we watched as security teams, mostly in larger businesses, deployed AI security in a haphazard manner. Specifically, we watched as AI committees were created to enable prototypes and pilots which were then showcased to senior executives and boards. This resulted in good demos, but lukewarm risk reduction.

Where we are today, in our estimation, is that the overall mood is beginning to shift. Senior executives are now thankfully beginning to ask their security leaders how AI security can and should become an operational discipline with real return on investment versus just a series of proof of concept (POC) experiments.

The healthy result is the emergence of more honest and reasonable enterprise AI security discussions that are designed to strategize how this threat—and opportunity—can best be managed. If done right, the result should be a healthy application of AI to the cybersecurity task.

We hope this *Enterprise AI Security Handbook* will provide a useful context. This is exactly why we developed it.

MODELS, SYSTEMS, AND ECOSYSTEMS

The terminology used during our research discussions was hardly consistent. We found terms like "model," "application," "system," and "infrastructure" used without clarity. Even common terms were often mangled. We've seen ChatGPT rendered ChatGTP. Here's a question: Can you spell out what the acronym "GPT" stands for? (Bonus points if you can. We'll wait while you look it up.)

We should start with some concept definitions—and we promise to keep things simple. Let's create a representation of an AI system using three components. First, there is the AI model developed by a large tech firm, often over a long period of time and at great expense. Second, there is the AI application developed by an organization, often using local data. And third, there is the surrounding AI ecosystem for that application. These components are shown in Figure 1.

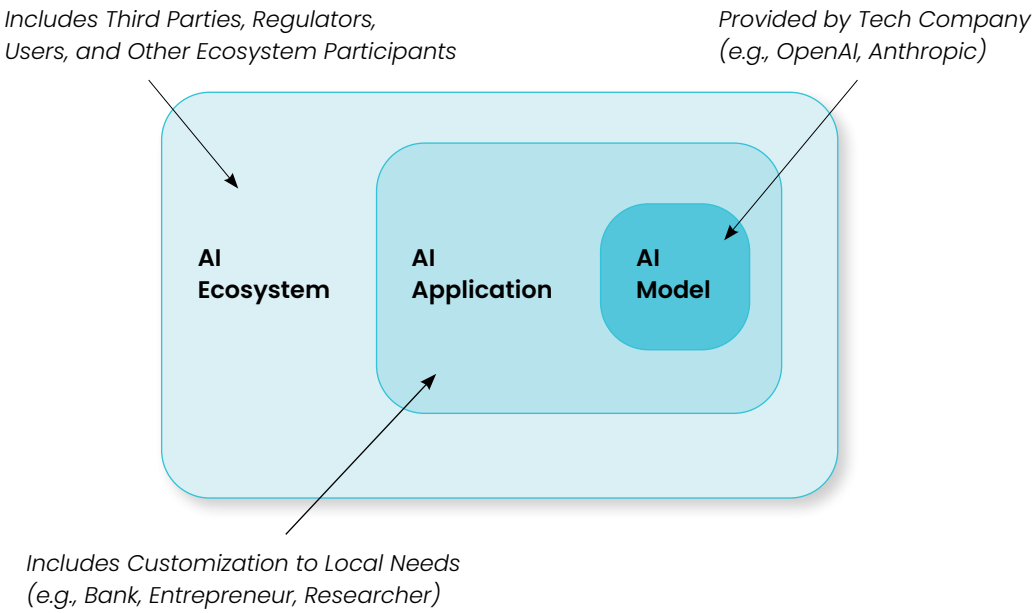


Figure 1-1. AI Model, Application, and Ecosystem

We will refer collectively to these three components, working in tandem, as an AI system. As you might expect, the purpose of this AI system will vary. For example, when it is used to create content like video and text, then we refer to this as a Generative AI (GenAI) system. This is obviously different than the AI that will be used to steer an autonomous vehicle.

And a special component of most GenAI systems involves something called a large language model (LLM) that trains on human language so that it can interact comfortably with human beings. That said, let's dig a bit more into the three components of an AI system.

At the core, the AI model must be built and maintained through trustworthy pipelines, free of data poisoning, bias, or hallucinations. This is a task that we believe should be the responsibility of the model developer. Microsoft and Google, for example, should make the investment to test, tune, and optimize their models to avoid these problems.

The AI application, comprising interfaces, local data, infrastructure, and users, must include security controls to prevent leakage or misuse, perhaps through unauthorized access. These local security controls should be the responsibility of the enterprise security team, and we will spend most of our time here on this task.

The larger AI ecosystem connecting suppliers, third parties, regulators, and end users must be governed with clarity and accountability. This depends on government leaders, regulators, and even analyst groups like TAG to play responsible roles driving good decisions. We offer this in the spirit of improving the global AI security ecosystem.

GUARDRAILS THAT MATTER

One term we hear often is *AI guardrail*. Managers and executives tend to use the term generically to refer to any risk mitigation put in place to protect an AI system from cyber threats. For example, a board member might ask the CISO if the company has sufficient guardrails in place for their AI.

In practice, however, guardrails are usually referenced by vendors and security teams in the context of runtime protection. To that end, guardrails tend to deal with technical threats like prompt injection, model hallucination, or data leakage. This is different than tools that provide visibility of AI in the context of existing controls.

Ultimately, we will reference the functions, procedures, and policies designed to ensure that AI models and applications operate consistent with desired objectives using the blanket term AI Security. And even this gets a little sticky when we reference the use of AI to enhance security protections of non-AI systems. But relax, you need not worry too much about the wording. We'll try to be clear.

FROM PROOF OF CONCEPT TO PRODUCTION

A key theme in this book involves the transition from proof of concept (POC) to production-scale AI security. As mentioned above, many organizations funded AI security projects in 2025 with the sole goal of demonstrating that *something* was being done. This may have satisfied senior leadership's curiosity, but it often did little to address any meaningful risk.

The true test of AI security initiatives and programs in 2026 and beyond, we believe, will be two-fold: First, the programs must successfully connect the objectives of the security practitioner team (to address risk), the leadership team (to drive the business), and the investment-supported vendor community (to make a profit).

But it is also true that POC projects must be associated with a clear path toward scaling to production. Deployment of an AI security tool into a demonstration and test environment makes for nice conversation, but practitioners are too busy for such frivolity. Instead, hard questions must be addressed, such as the following:

- **Can a given AI guardrail be integrated into production AI systems without friction?**
- **Can AI protection measures work against an evolving set of adversary risks?**
- **Can the right AI security tools be obtained amidst security budget constraints?**

These are the right types of questions to be asking if you are a practitioner trying to figure out how to best leverage AI for your security program. And the answers to these questions, which apply to both AI for security as well as security for AI, will no doubt separate successful AI security programs from the endless churn of pilots and demos.

HOW WE INCLUDE VENDORS

Considerable thought went into how we would deal with commercial AI security vendors in this book. It was never a question of whether we would name names from the AI security vendor community. We could not conceive of proceeding without mentioning specific commercial vendors. Without them, this entire project would have been reduced to mere academics.

What we ended up doing was this: From an initial (and enormous) list of AI security vendors—almost 300 in total—we narrowed it down based on frequency of mention in inbound inquiries, the quality of their platforms (determined during briefings), references from customers, interviews with their executives, input from funding teams, and other admittedly arbitrary criteria.

This might seem like an ad hoc process, but it is what industry analysts do. We review the field and then make recommendations. Sadly, too many research firms are swayed by financial considerations in a so-called pay-for-play competition.

Our full disclosure here is that we do have on-going advisory relationships at TAG with some AI security vendors—perhaps 10% of the ones mentioned in this volume. But we do our best to separate that relationship from how we make recommendations to practitioners, who are our main focus at TAG.

The result, we hope, is that we have included some excellent AI security vendors in Chapters 6 and 7 of this book. But we can promise that we have also left out many amazing and capable vendors. Our book is meant to stimulate and inform, but it is by no means a comprehensive list. Readers should keep that in mind as they go through the vendor information.

HOW TO USE THIS BOOK

Because we have targeted our commentary and guidance to enterprise security practitioners, we would expect that most readers will use this book as a roadmap for dealing with AI cyber risk. That, as we have said, is our primary objective. We developed it based on our observation that practitioners were struggling to construct an AI security plan.

But other stakeholders can and should use this volume to gain insight into the types of issues being addressed by the actual buyers who will use the AI security technology funded and developed by the investment and vendor communities. Our goal, as we have said, is to provide greater clarity for all in AI security at a time when we probably need it most. We hope you find this book not just useful but, as you get deeper into it, indispensable.

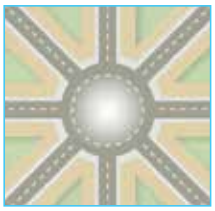
Now that you have some reading momentum, we suggest you flip or click to get started on Chapter 2.

We'll see you there.

CHAPTER 2

ROADMAP FOR SECURING AI

Here are six critical tasks to address AI usage risks as well as leverage its power to advance your security agenda.



Our assumption is that you, the reader, are an enterprise security practitioner, perhaps a CISO, mandated to develop a security roadmap for AI. You are our primary audience. If you are a different kind of stakeholder in AI security (e.g., startup founder, corporate executive, researcher), then you are welcome here as well. So, let's get started.

We will assume that your responsibility, which might come with funding from an executive AI steering committee established in the last couple of years, likely focuses on identifying effective security solutions to protect AI usage within the organization. It might also extend to using AI for improved security operations.

The ecosystem in which you operate, we assume, includes involvement from the cybersecurity team of which you are a part: business unit leaders, senior leadership team (including the CEO), AI committee, and external participants including auditors, regulators, customers, and third-party partners and suppliers. This ecosystem of players will influence development of your enterprise AI security roadmap (see Figure 2-1).

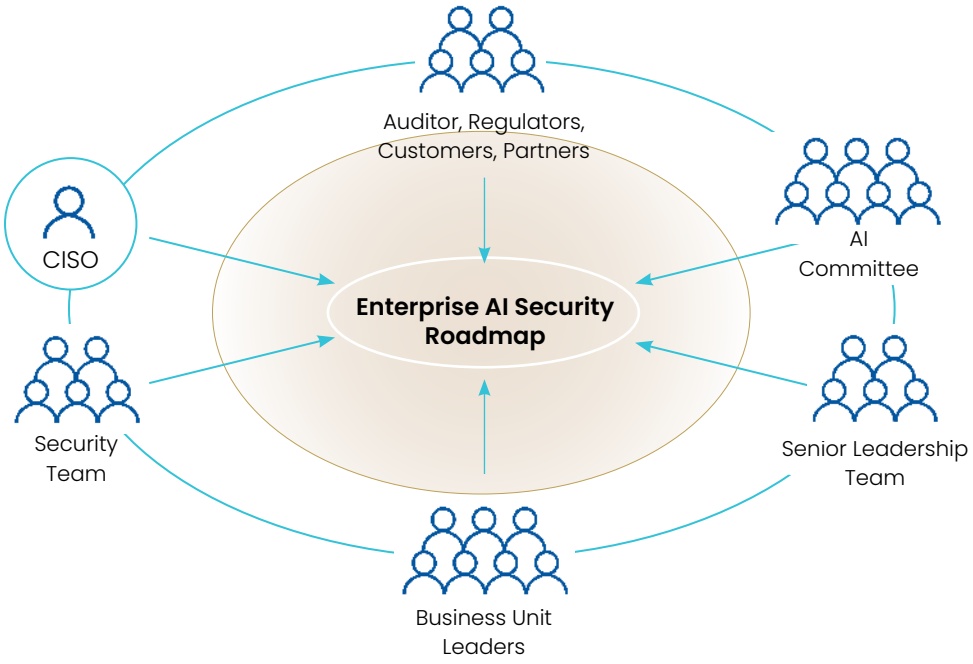


Figure 2-1. Management Ecosystem for AI Security Roadmap

Our goal in this chapter is to provide an initial, high-level, but tailorable management roadmap for securing the use of AI within the enterprise. This includes not just protections for GenAI and LLMs, but also support for AI-enabled security operations center (SOC) automation and detection of deepfakes and misinformation. Our proposed roadmap is intended to help enterprise teams build safeguards around AI systems, and to integrate those safeguards into their broader security architecture.

Special attention is given here to the use of the best available AI security products, services, and platforms—based on our day-to-day TAG research into vendor offerings. Such emphasis should differentiate this work from more academic reports that treat AI risk in a sterile and more theoretical manner. Our approach is to dive into the specifics of how a CISO should (or should not) select and implement an actual commercial solution.

GETTING STARTED

As many readers will know, AI is no longer a novelty in the enterprise. It has evolved into an enabler, usually for expected cost reductions across functions ranging from customer support to internal workflow automation. With such adoption comes the responsibility to ensure that AI systems are deployed in ways that ensure confidentiality, integrity, and availability, while also meeting key AI-related compliance and governance requirements.

As we explained in the introduction, over the past couple of years our team at TAG has conducted workshops with enterprise security leaders across multiple sectors. These discussions have confirmed that while organizations understand the potential of AI, many lack an actionable plan for integrating AI securely into their existing environments. This chapter offers an initial roadmap, including some simple steps for managing risk in an AI deployment.

DIFFERENTIATING AI MODELS, SYSTEMS, AND ECOSYSTEMS

Let's take a moment to remind readers of the three primary components of an AI system—AI models, AI applications, and AI ecosystems—that will comprise a typical enterprise AI use case.

- **AI models** are the foundational components, including large language models (LLMs) and generative AI (GenAI), that are created by the major technology providers such as OpenAI, Anthropic, and Google. Emerging cyber risks at this layer include poisoned training data, backdoors, or embedded Trojans.
- **AI applications** are practical implementations of AI models for business contexts, usually including special interfaces, connectors, and integrations that transform a raw model into a usable service. Here, the threats involve data leakage, application manipulation, and operational disruption.
- **AI ecosystems** encompass the broader environmental context in which AI systems operate, including governance bodies, regulatory frameworks, and interconnected technologies. This layer introduces security concerns around compliance, societal impact, and systemic vulnerabilities.

For enterprise CISOs, our observation is that the focal point for protection should be mostly at the AI application level, since that is where the local responsibility for implementation and defense mostly resides, especially if local data is being used to train the application in a set-up referred to as retrieval augmented generation (RAG).

The AI ecosystem is certainly a consideration, but it is not something the CISO can control. And countering threats in AI models is usually far outside the control and influence of CISOs. This is a key point, because many CISOs are held unreasonably accountable for flaws introduced by an AI model developer such as Microsoft or Anthropic. When a model hallucinates, it is generally the fault of the technology company, even though the CISO might be held to create input or output filters to compensate. (Life is not fair.)

THREATS INTRODUCED BY AI SYSTEMS

The use of AI in enterprise does introduce new types of cyber risks. While some of these risks are extensions of SaaS security (e.g., discovering the nature of usage) or network security (e.g., ensuring continued operation amidst DDOS risk), other AI-related threats emerge directly from the behavior of the new technology. Below are some of the AI system issues that emerge in the context of the well-known CIA triad:

- **Confidentiality Risks:** Prompt injection and related indirect prompt attacks can apparently trick systems into revealing sensitive data. Logs of user interactions may also be misused if not properly governed. Multi-tenant AI platforms risk cross-customer data leakage.
- **Integrity Risks:** Training data, which is essential for AI, can be intentionally and maliciously poisoned to bias outputs or embed hidden triggers. Adversarial inputs can manipulate AI-generated responses, which can undermine trust and enable fraud or misinformation.
- **Availability Risks:** AI inference can be quite resource-intensive (not to mention energy intensive), which can create opportunities for malicious denial-of-service conditions. Presumably, malformed or excessive queries can degrade or disable AI-related services.

We were careful to couch our descriptions above with words like “apparently” and “presumably,” because most AI threats remain largely theoretical. This is not to imply that they are not real, just that it is not yet possible to identify AI attacks that have created conditions commensurate with, say, the notorious Target, Home Depot, Sony, OPM, Uber, Twitter, and other cyberattacks.

Moving beyond the CIA model, our assumption is that AI will bring new types of risks tied to bias, intellectual property misuse, and opaque decision-making, all of which complicate incident response and regulatory compliance. As we have outlined above, however, many of these risks are connected to either models or ecosystems, which place them outside the responsibility of the typical organization’s CISO-led security team.

A LAYERED APPROACH TO ENTERPRISE AI SECURITY

Let's get down to specifics regarding an actual enterprise roadmap. We will assume that your organization already has AI governance, or at least an AI committee created to develop an overall plan for AI. Obviously, your AI security roadmap will have to be integrated into any broader context. That said, based on our work with enterprise teams, TAG recommends a layered approach built around six concurrent tasks.

The origin of these tasks is the insight we gained from our workshops with AI and security teams in 2024 and 2025. We tried to gather the best ideas and approaches from actual enterprise practitioners. None of the tasks below came from what we thought should be done. These were based on what we observed actually being done.

Our view is that efforts should be made to initiate all six of these tasks to begin the journey toward secure implementation and use of AI across your enterprise. With funding constraints, It is possible that you might be forced to start with a subset. But ideally, all would operate in parallel and few would ever reach the point of some logical completion—instead, becoming on-going curation of your AI security (see Figure 2-2).

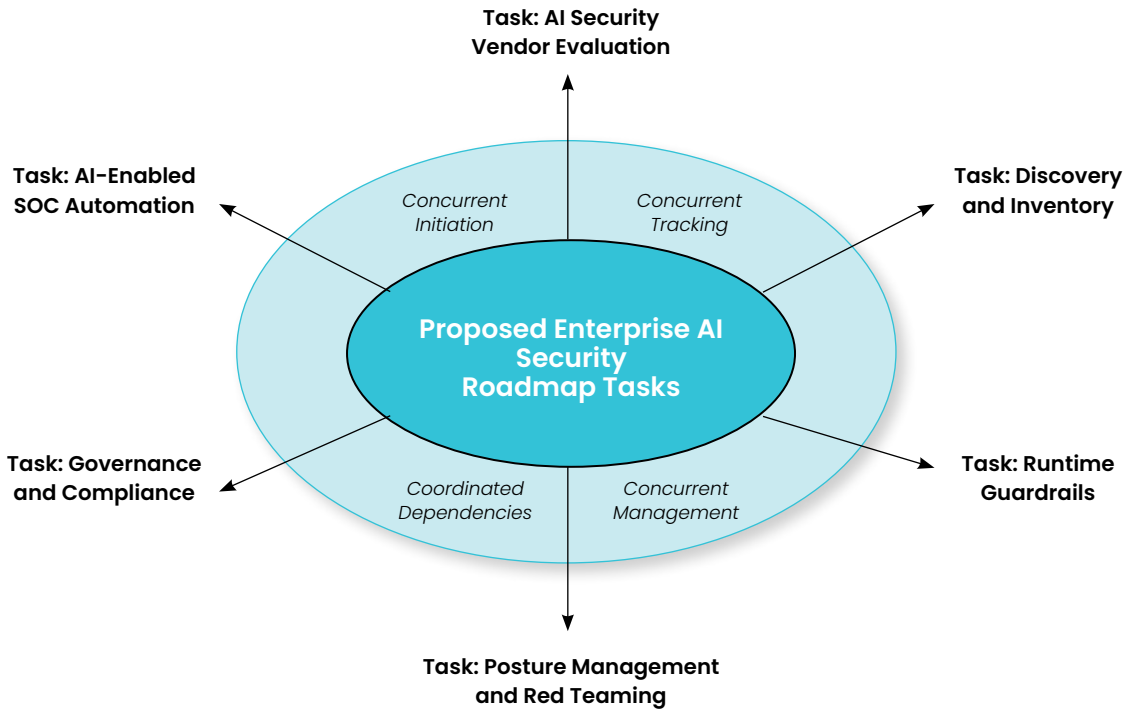


Figure 2-2. Enterprise AI Security Roadmap Tasks

Let's briefly examine these six tasks before we dive into a more detailed discussion on how they would operate in a practical setting (subject, of course, to the usual local management required to meet the actual needs of the organization). Here is a brief summary, numbered for convenience rather than a suggested order of implementation:

- **Task 1:** AI Security Vendor Evaluation: The security team should explore the possibilities for AI security by engaging in discussions with vendors including startups.
- **Task 2:** Discovery and Inventory: Security teams should find a means to discover and catalog all AI systems in use, including shadow AI, to establish a baseline.

- **Task 3:** Runtime Guardrails: Teams should inspect model inputs and outputs to detect evidence of prompt injection, jailbreaks, and unsafe responses.
- **Task 4:** Posture Management and Red Teaming: Plans should be made to continuously test AI systems through adversarial evaluation and to align findings with priorities.
- **Task 5:** Governance and Compliance: Teams should align AI risk oversight with frameworks (NIST, OWASP, EU AI Act) while remaining adaptable to evolving rules.
- **Task 6:** AI-Enabled SOC Automation: Teams should remember to include plans to use AI itself to enhance SOC workflows, from alert triage to incident response.

As should be evident from the diagram in Figure 2-2, and as we have repeated, these six tasks are not intended to be performed sequentially but in parallel. Obviously, vendor evaluation will drive deployment of any vendor solution for security protection, so there might be some implied ordering. But it is healthier to view these tasks as separate and concurrent, and they should be allowed to progress with their natural interdependencies.

Tracking of these tasks will follow whatever management frameworks, processes, and methodologies are used by the organization. Conceptually, we recommend that each task be reviewed for progress and status, and that more individual tracking be provided for each of the six tasks by their respective assigned teams. Below, in Figure 2-3, is a sample conceptual view of what should be tracked at the highest level using a simple spreadsheet:

AI Security Roadmap Task - High Level Tracking	Owner	Status - 2Q26	Notes
Task: AI Security Vendor Evaluation	Ron Smith	Green	Security team in process of meeting with seven vendors recommended by TAG
Task: Discovery and Inventory	Jay Patel	Yellow	Manual reviews ongoing but no plan for automation yet
Task: Runtime Guardrails	Kristy McDonald	Green	Runtime model developed and reviewed (vendor to be discussed with TAG)
Task: Posture Management and Red Teaming	Nick Wong	Green	Draft red team plan in place
Task: Governance and Compliance	Amber Field	Red	Compliance issues identified and being reviewed
Task: AI-Enabled SOC Automation	JR Ambrosini	Yellow	SOC team reviewing Security Co-Pilots with TAG

Figure 2-3. Recommended High-Level Tracking for Six AI Security Roadmap Tasks

TASK: AI SECURITY VENDOR EVALUATION

If you ask ChatGPT to evaluate AI security vendors, you will be provided with a somewhat arbitrary list, not unlike the results of a Google search for AI security vendors. We mention this because we have observed that many security teams begin their investigation in this manner. There is nothing inherently evil about doing this but view it as an imperfect starting point.

Obviously, TAG Research as a Service (RaaS) customers can rely on the TAG analyst team to help with this process, and Chapters 6 and 7 in this book include some practical vendor recommendations. But vendor selection should be viewed as an on-going initiative, one that perhaps never really completes. You should work this with your procurement group.

Complicating matters in AI security, however, is the enormous level of funding (likely a bubble), that result in a confusing mess of vendors and startups, all claiming to have things pretty well solved in terms of AI security risk. We believe many vendors, especially startups, will fail in 2026, mostly because too many of these companies have developed solutions in search of problems. So, be ready to see a rash of startups dissolved or absorbed into larger platforms.

What this implies is that the usual process of making lists, attending demos, running proof of concept (POC) trials, and then selecting a vendor for production, might not work so well in 2026, when it comes to AI security. Instead, we recommend a review that is more focused on advancing learning around AI risk, optimizing integration with existing security tools, and maximizing vendor options are the market changes. Here are some specifics on these three objectives:

1. **Advancing Learning:** This should be your first goal in working with AI security vendors in 2026 and beyond; namely, to advance your team's learning and understanding of AI risks and how they might be mitigated. So much is changing here that practitioners should view AI security as "clay being molded," so to speak, rather than as anything set in stone. Work with vendors who can provide insight and learning.

2. **Optimizing Integration:** You have already made significant investments in your security architecture, including deals with dozens of commercial vendors (if not more). The new use case of AI should not result in any de-emphasis on this existing base, but rather on complementing it where necessary. Focus on vendors that will be good at integrating with what you have, which will be unique to your local environment.

3. **Maximizing Options:** In the spirit of our learning objective above, we strongly recommend that you maximize your options by not locking in long-term deals. Let your AI security vendors earn their business with you and be promiscuous to the degree that your procurement team will allow. The last thing you need is a multi-year deal with a vendor that does not address some new threat that arises needing mitigation.

Special note should be made here, of course, to the tools and platforms using AI to advance automation and autonomy in the SOC. These tools, also referenced in Chapter 6, have demonstrated practical value, often as next-generation options for earlier security orchestration automation and response (SOAR) tools. This implies that good deals can be made here—even multi-year contracts, if desired.

The status tracking (i.e., red, yellow, green) of vendor evaluation should include the three criteria elements mentioned above. We like the approach (see Figure 2-4) of keeping track of how each vendor supports these objectives in addition to the usual types of considerations when selecting vendors (e.g., cost, terms, features, integrations):

AI Security Vendor Evaluation Criteria	Status	Notes
Advancing Learning		
Does this vendor support AI security training sessions?		
Does this vendor provide AI security workshops?		
Does this vendor provide good AI security written articles, reports, and papers?		
Optimizing Integration		
Does this vendor include APIs for data sharing and integration with other tools?		
Does this vendor include connectors to SIEM platforms?		
Does this vendor have a capable and accessible development team to make changes?		
Maximizing Options		
Does this AI security vendor include flexible contract terms?		
Does this Ai security vendor allow for cancellation clauses in the contract?		
Does this vendor have an R&D program to keep up with changes?		

Figure 2-4. Recommended Criteria for Assessing AI Security Vendors

We should emphasize that we strongly recommend focus on learning, integration, and flexible options because we expect that the AI threat model, discussed in detail later in this book, as well as the actual AI use cases in virtually every organization, will change dramatically. This can include, for example, models from Anthropic, OpenAI, and others fixing issues like prompt injection. If they did, then do you see why it would be a mistake to have a multi-year deal with a vendor to fix this problem?

TASK: DISCOVERY AND INVENTORY

Most organizations today operate with incomplete visibility into how AI systems are actually being used. Shadow AI, employee experimentation, and SaaS integrations have blurred the boundaries between sanctioned and unsanctioned model usage. The objective of this task is to establish a dynamic inventory of all AI systems, meaning every model, application, and data pipeline that uses AI within the enterprise. We see three discovery channels as necessary:

1. **Top-down Assessment:** This includes interviews and discussions with business unit leaders to identify their AI-related use cases (e.g., marketing analytics, customer chatbots, internal knowledge assistants).
2. **Bottom-Up Scanning:** This involves using technical discovery tools (using both new and existing platforms) to detect API traffic to common AI providers (OpenAI, Anthropic, Hugging Face, etc.), including indirect integrations through SaaS.
3. **Dataflow Mapping:** This allows for identifying evidence of datasets feeding AI models, including sensitive data, regulated PII, intellectual property, or source code. This can be detected dynamically or in some discovered artifact referencing such data usage.

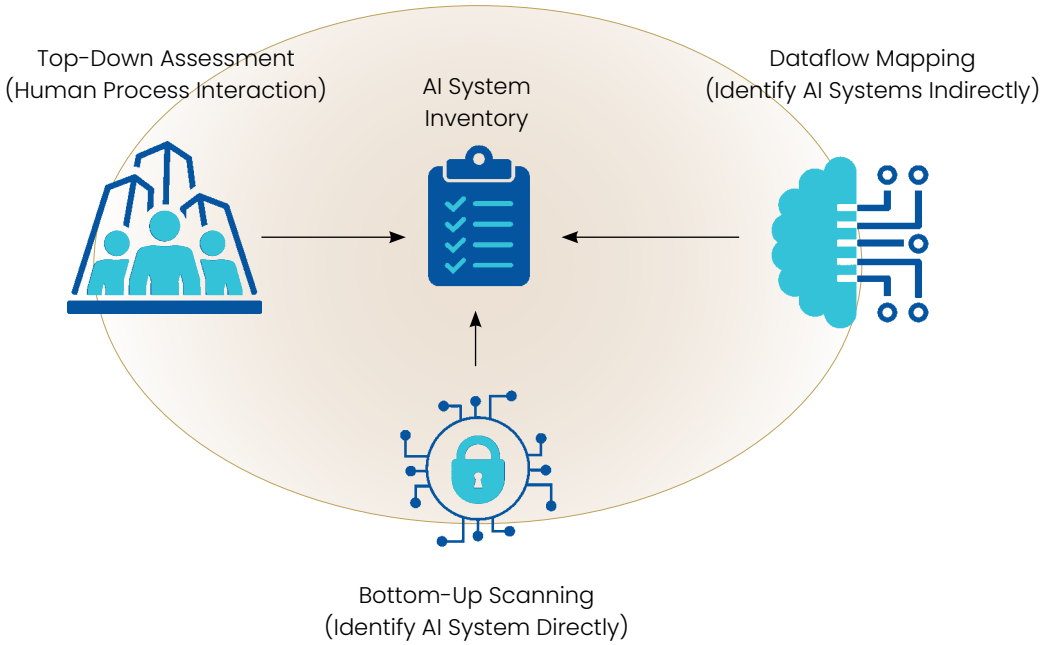


Figure 2-5. Three Strategies to Develop an AI Inventory

Each discovered system should be recorded in an AI configuration management database (AI-CMDB) with attributes such as origin (e.g., proprietary, open source, or API-based), AI security application owner and use case, data sensitivity and retention policy, security and compliance category (e.g., regulated, experimental, internal-only), and relevant integrations with existing IAM, logging and other security platforms or tools.

This inventory will gradually begin to approximate an initial source of truth for AI security oversight, but you should expect it to be incomplete and rapidly changing. It is nevertheless necessary so that you have the context required to apply policies, because not every AI system requires the same level of control. For example, a customer-facing chatbot might need input filtering and logging, while an internal generative assistant may simply require isolation and DLP enforcement.

Over time, you will need to find a way for your AI security inventory to be continuously updated—ideally through automated discovery integrated into data-loss prevention systems and network visibility tools. Just as asset discovery preceded endpoint protection in cybersecurity's evolution, AI discovery will precede AI defense in this new domain. This also goes for the AI-enabled tooling you might be introducing to your SOC.

TASK: RUNTIME GUARDRAILS

As AI systems are discovered, the enterprise security team must find ways to begin enforcing proper behavioral controls during runtime. This task represents one of the key front lines of AI security: namely, the layer where user prompts, model inferences, and generated outputs are mediated in real time. Runtime guardrails serve two fundamental purposes in AI security:

1. **Malicious Input:** This involves protecting the AI system from malicious or unsafe inputs, such as prompt injections or jailbreak attempts that subvert system behavior. Such input can come from humans or from automated workloads, including other AI agents.

2. **Malicious Output:** This involves protecting the enterprise from receiving unsafe or confidential AI outputs, such as the release of proprietary data or toxic responses that could trigger an exposure. Such output can be text-based, or it can involve operational commands affecting systems.

To achieve these two complementary AI runtime security objectives, modern enterprise architectures should deploy proper security gateways, filters, and other controls between users and AI systems. These filters, sometimes called LLM firewalls or safety gateways, analyze input and output tokens to detect the presence of the following types of conditional indicators, which are highly suggestive that some security issue is present:

- unusual prompt or instruction anomalies, such as those that ignore previous rules or reveal hidden system prompts;
- the presence of seemingly sensitive content such as credit card numbers, credentials, and customer identifiers;
- evidence of security policy violations, such as prohibited topics or unapproved external connections;
- certain behavioral signals, such as frequency of requests, entropy of responses, and anomalous session patterns.

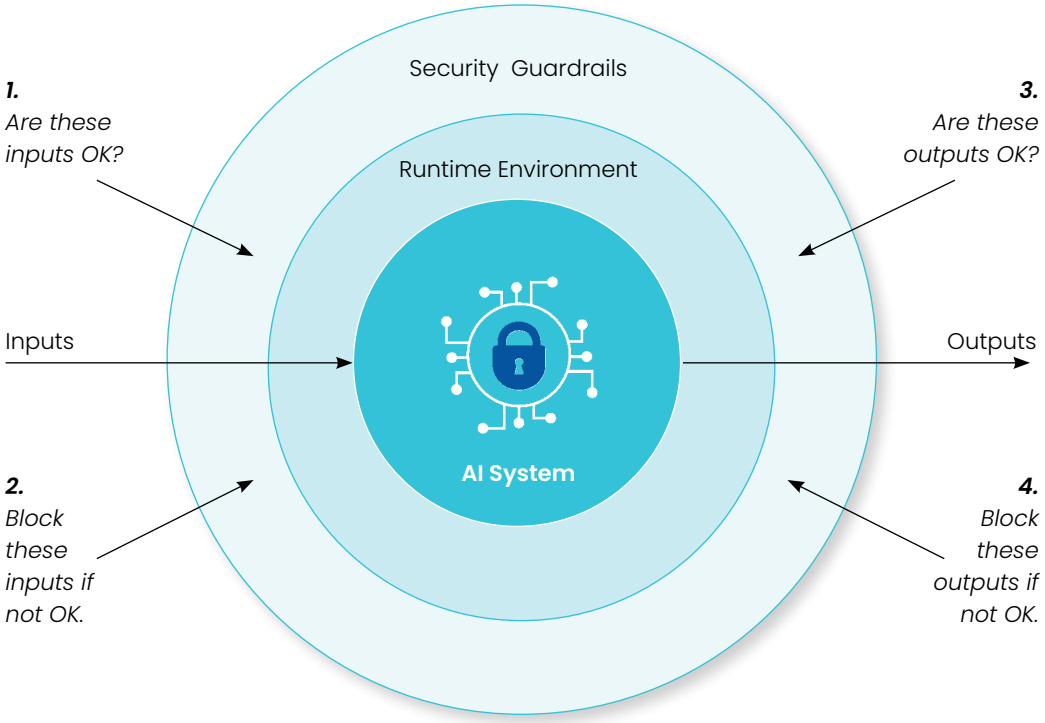


Figure 2-6. Runtime Mediation Schema for AI Security

Beyond simple allow/deny decisions, advanced runtime AI guardrail systems should assign a quantitative score and/or contextualize the interactions. This can include sending the security telemetry to the SOC for further correlation. A prompt that attempts data exfiltration, for example, should not only be blocked but should also trigger an alert correlated with user identity and device posture.

The AI runtime layer should integrate policy enforcement mechanisms from existing secure access service edge (SASE) components (if present), including identity, device trust, and network context, so that any AI usage will be dynamically gated by suitable enterprise risk posture. This runtime approach will result in the conversion of AI security from a static whitelist to a contextualized, adaptive control system.

Finally, AI security guardrail performance can and must be continuously measured. Enterprises should introduce tooling that can track detection rates, false positives, latency, and user experience friction. Over time, the AI security roadmap should treat these metrics the same way SOCs treat detection engineering, which is a discipline of tuning, testing, and continuous improvement.

TASK: POSTURE MANAGEMENT AND RED TEAMING

Assessment of AI security posture via testing and red teaming is pretty essential. This means developing a sustained capability to improve the posture of all AI systems that are either deployed or being considered for deployment. Security teams should design posture assessment in the context of any existing testing and/or red teaming approaches that have already been put in place.

The good news is that many AI posture management platforms and offerings are emerging with libraries of AI threats. These platforms can provide dashboards of configuration and exposure, allowing CISOs to see which AI systems are public-facing, which datasets are unclassified, and which systems fail to comply with internal policy or external regulation. The output of these tools should feed into enterprise risk registers and compliance reviews.

In parallel, enterprises should plan to perform AI red teaming, and this should include structured adversarial testing of AI models, applications, and ecosystems. Unlike traditional penetration testing, AI red teaming requires focus on a different set of threats and issues. These include the following attacks, many of which require some combination of controls from the security team, the AI model provider, and the surrounding ecosystem:

- adversarial machine learning and poisoning attacks, which can create challenges in the outputs generated by LLMs and GenAI systems;
- model inversion and extraction techniques, which degrade the level of trust that users will have in a given AI system;
- prompt engineering for malicious manipulation, which is an extension of the challenges security teams have always had with social engineering attacks;
- hallucination detection and content evaluation, which are usually problems that originate in the model, but which do extend to the AI application or system.

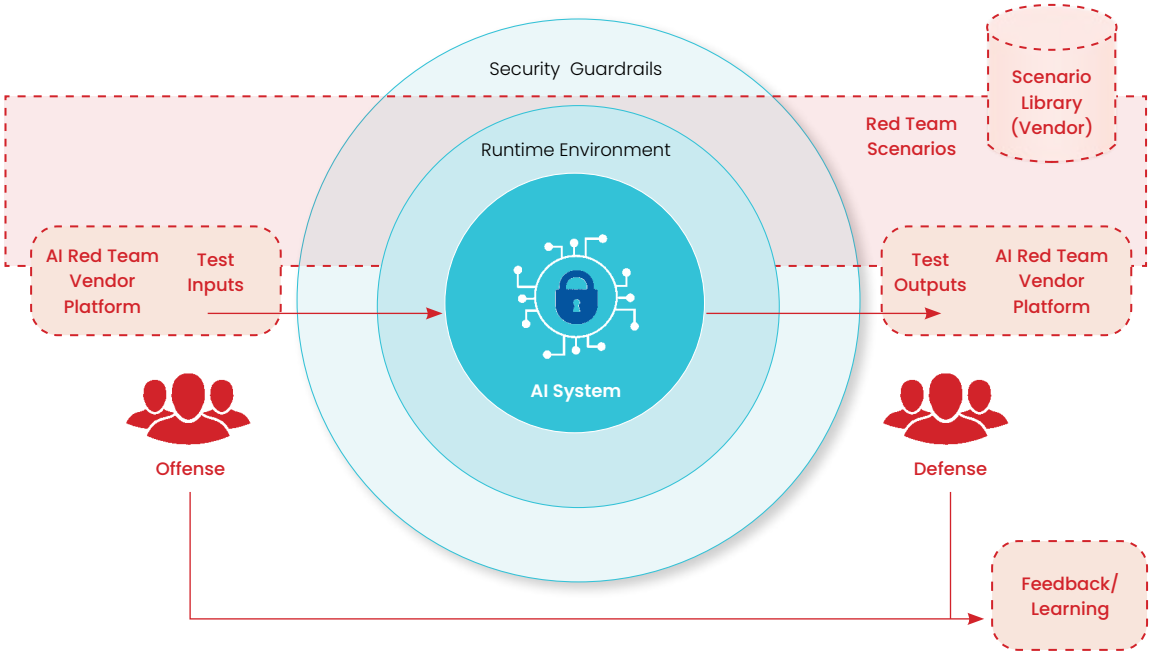


Figure 2-7. AI Security Red Team Focus Areas

TAG recommends that every enterprise running production AI systems begins to plan an appropriate cadence for red team testing and exercises using both internal and external experts. These exercises should test both offensive and defensive maturity. And the focus should not just be on whether the system can be compromised, but whether the organization can detect and respond to such compromise in real time.

Our advice is that each test should influence a posture improvement plan aligned with business priorities. For example, if red team testing reveals that an internal generative assistant leaks fragments of sensitive data, then the response may involve revising DLP policies, modifying prompt templates, and updating employee training. In some more intense cases, the result could be a fundamental change to the use of AI as part of the business model.

TASK: GOVERNANCE AND COMPLIANCE

As any working practitioner knows, just putting in proper functional security controls alone is insufficient. Without clear governance oversight, even technically secure systems can drift into non-compliance—although admittedly, it's not all that clear what this means for AI security (yet). Nevertheless, governance should define who makes decisions, what rules are followed, and how accountability is enforced.

Before we get specific here, we must say that during reviews with AI security vendors, we've seen the term "governance" refer to a variety of different functions ranging from discovery, to guardrails, to even red teaming. This is inevitable in any new discipline, so readers are warned to pay close attention when listening to AI security vendor pitches. That said, here is what we suggest as a reasonable three-tier governance and compliance model for enterprise AI:

1. **Policy Tier:** You should establish an AI usage policy that defines permissible data sources, approval workflows, and security baselines for AI app deployment. If you can classify risk tiers (e.g., internal, customer-facing), then that would be good.

2. **Oversight Tier:** Assuming you already have a corporate AI governance committee with representation from cybersecurity, legal, compliance, risk management, and the business, you should establish a security subcommittee to focus on cyber-related issues.

3. **Accountability Tier:** You would be well-served to assign named owners to each AI system with responsibility for risk acceptance, model retraining approval, and compliance documentation. This will require coordination with the business units.

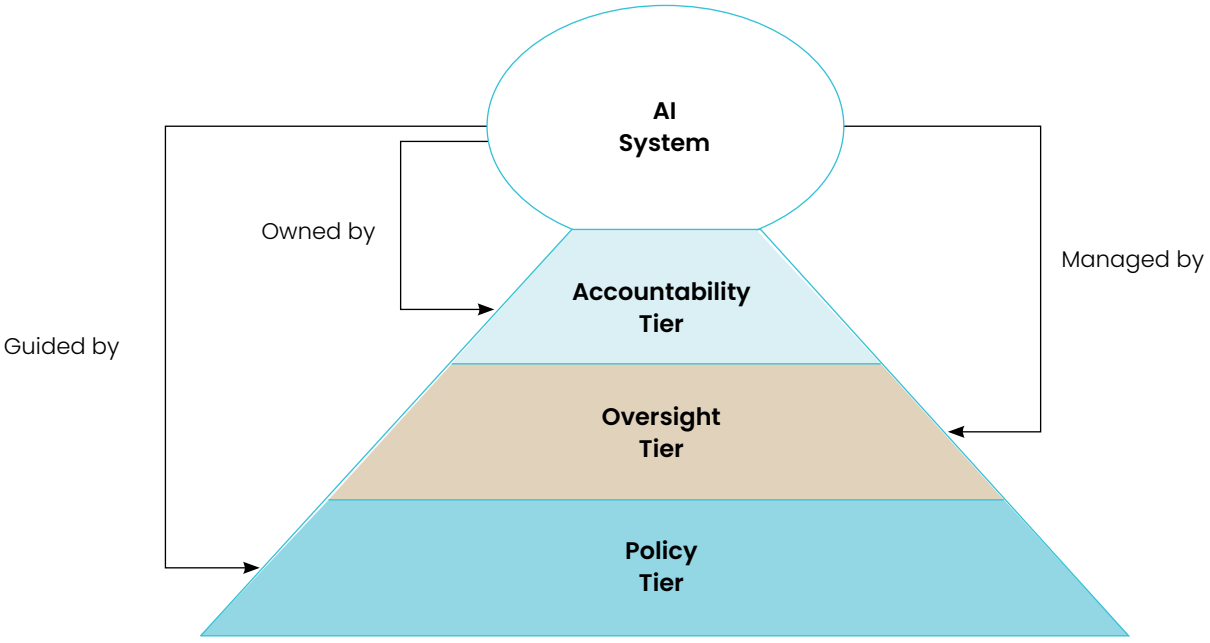


Figure 2-8. Governance Model for AI Security

Once governance is defined, it would be also a good idea to try mapping your policies to frameworks such as NIST AI Risk Management Framework (RMF) for risk identification, measurement, and mitigation; ISO/IEC 42001, for AI management system certification and continuous improvement; and the OWASP Top 10 for LLMs for technical control alignment. This might be supported in your GRC platform—you will have to check.

Compliance reporting should also move beyond checklists. Enterprises should establish AI risk dashboards that quantify exposure using measurable indicators such as the number of AI systems in production, percentage with runtime monitoring, number of red team tests conducted, and response time to Ai-related incidents. These metrics should feed into the same governance channels as privacy and cybersecurity reporting to the board.

Presumably, this is where AI security governance vendors come into play. We like the idea of having a platform that supports policy identification, policy mapping, and other governance-related issues. This can be a new AI governance platform or perhaps it can be your existing GRC tool. What becomes a bit blurry, as we've suggested, is when this support extends to the actual mitigation, either during development or runtime (as a guardrail).

Our analyst team at TAG uses the term "Swiss Army Knife" to refer to the phenomenon of vendors, especially heavily funded startups, that are under pressure to sell and have decided that they literally must do everything AI-related. Our advice is that they would be much better served to focus, but we understand the pressure to add logos to their sales roster. Buyers beware of this approach.

TASK: AI-ENABLED SOC AUTOMATION

The sixth stage of the roadmap reflects a kind of symmetry. While the first five secure AI, the sixth uses AI to secure everything else. AI-enabled SOC automation is the practical expression of this reciprocity, and we see vendors every day in our research at TAG that are focused on this important task, which ultimately replaces many human tasks with AI-enabled automation.

We all know that security operations centers face overwhelming alert volume, analyst fatigue, and skill shortages. AI can relieve these burdens through intelligent triage, context enrichment, and automated incident response. The same underlying AI systems that require protection in business contexts can, when safely applied, power autonomous cyber workflows. We recommend that implementation follow a staged progression, more or less as follows:

1. **Assisted AI Analysis:** Your roadmap should start by using Generative AI copilots or LLMs to summarize alerts, explain vulnerabilities, and generate remediation guidance. This is a relatively easy step, but it allows for cultural acceptance of AI as a useful SOC tool.

2. **Orchestrated Response:** The next step is to review whether integrating AI into your existing SOAR (or comparable) platforms to automate repetitive tasks is an option. This can cover ticket creation, data enrichment, and correlation of threat intelligence.

3. **Autonomous AI Operation:** Now you can begin to consider deploying more domain-specific AI agents that can execute defined playbooks under human supervision, such as isolating endpoints or rotating credentials.

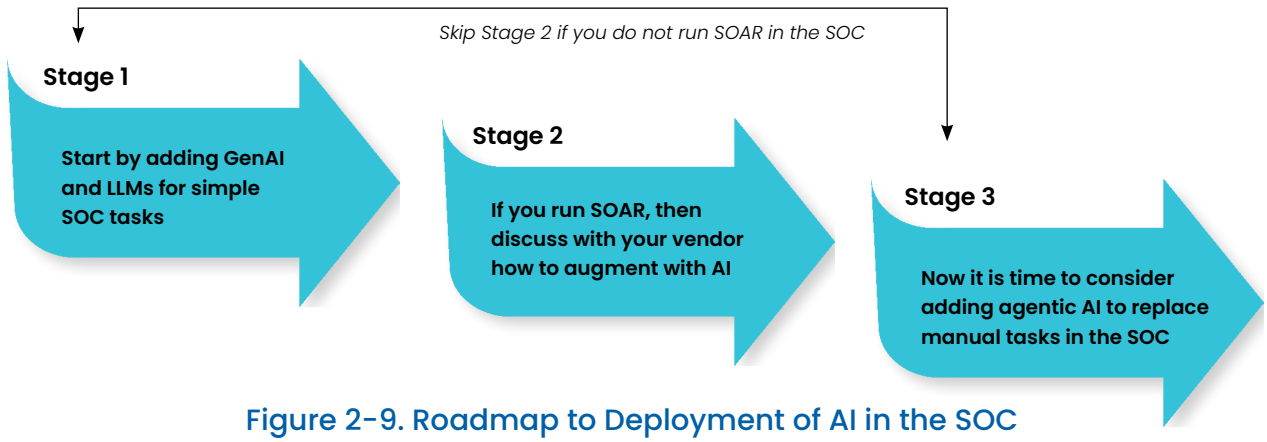


Figure 2-9. Roadmap to Deployment of AI in the SOC

The CISO must ensure security parity between AI-driven automation and the controls protecting production systems. That means clear access boundaries, robust auditing, and explainable reasoning. For example, an AI agent recommending quarantining a server should provide both the rationale and the dataset used for that decision. This is something a human would do today, but an AI agent should have no trouble taking on this task.

Beyond efficiency, AI-enabled SOC automation enhances resilience. During major incidents, when human analysts are saturated, AI agents can continue to process telemetry, detect anomalies, and preserve situational awareness. Over time, this automation layer becomes a force multiplier, allowing organizations to scale defense capabilities without proportional headcount growth.

In our estimation at TAG, SOC automation also acts as a useful feedback mechanism for AI security itself. The same event data used for detection can be analyzed to improve model guardrails, refine prompt filters, and enhance governance metrics. Thus, the roadmap completes a full lifecycle: namely, securing AI, then empowering security with AI in an endlessly improving cycle of adaptation.

NEXT STEPS: FROM GUARDRAILS TO GOVERNANCE TO GROWTH

It's time now to develop an action plan. And when you are doing the planning, recognize that securing AI in the enterprise (or using Ai for security) is not some singular project. Rather, it is a continuous maturation path. The six concurrent tasks outlined in this chapter should provide for you a tailorable, evidence-based structure for progress.

And that's a good place from which to adapt policy rules. As we will see.

CHAPTER 3

AN AI SECURITY POLICY

Adopt Rules that can serve as a baseline for establishing safe and secure use of artificial intelligence in the enterprise.

It is not uncommon for a CISO-led security team, even one with considerable experience and expertise, to be flummoxed when pondering how to create a policy for AI usage. Some start with domain-specific assertions such as "no AI will be used to interact directly with customers," or "no AI can be used to control industrial equipment." But such assertions are hollow, with only theoretical connection to actual threats and little reference to actual business objectives.

The best approach, we believe, is to develop policy that mirrors the existing cybersecurity approaches that are in place already. This implies policy guidance around how AI can (or cannot) be used for application security, authentication, and so on. We believe this is suitable as a baseline method, thus allowing business leaders to sort out when, where, and how AI will be applied to business unit-defined people, processes, and technology.

Additionally, we believe it makes little sense to create entirely new security policies for AI, which we view as a new use case for enterprise. Instead, we strongly recommend that CISOs guide their teams to leverage their existing policy, and AI should be addressed by existing controls. Obviously, new rules will emerge, and perhaps even some new categories of rules, but we propose this approach as the baseline (see Figure 3-1).

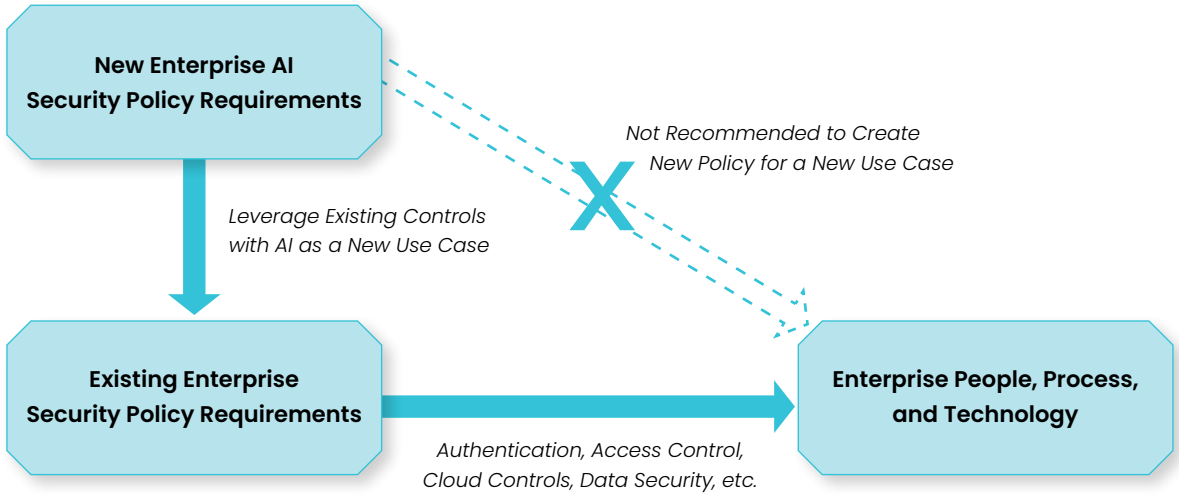


Figure 3-1. Recommended Security Policy Approach for AI

To that end, in this chapter, we suggest an enterprise-wide AI information security policy (AI-ISP) that includes requirements across a set of common enterprise security practices. The AI-ISP covers enterprise use of traditional AI/ML models, LLM/Gen AI-based systems, and AI-based SaaS applications, and is aligned with NIST AI RMF, OWASP Top 10, White House Executive Order on AI, and ISO/IEC 42001. We utilize the TAG Taxonomy as the basis for the AI-ISP.

PRELIMINARY: DEVELOPING INFORMATION SECURITY POLICIES

In the previous chapter, we outlined six tasks that we viewed as essential to getting an enterprise AI security roadmap in place. One shared artifact that will emerge and evolve as these tasks are worked involves the development of AI information security policies. We thus view this task as foundational to virtually everything being done to secure the enterprise for AI usage.

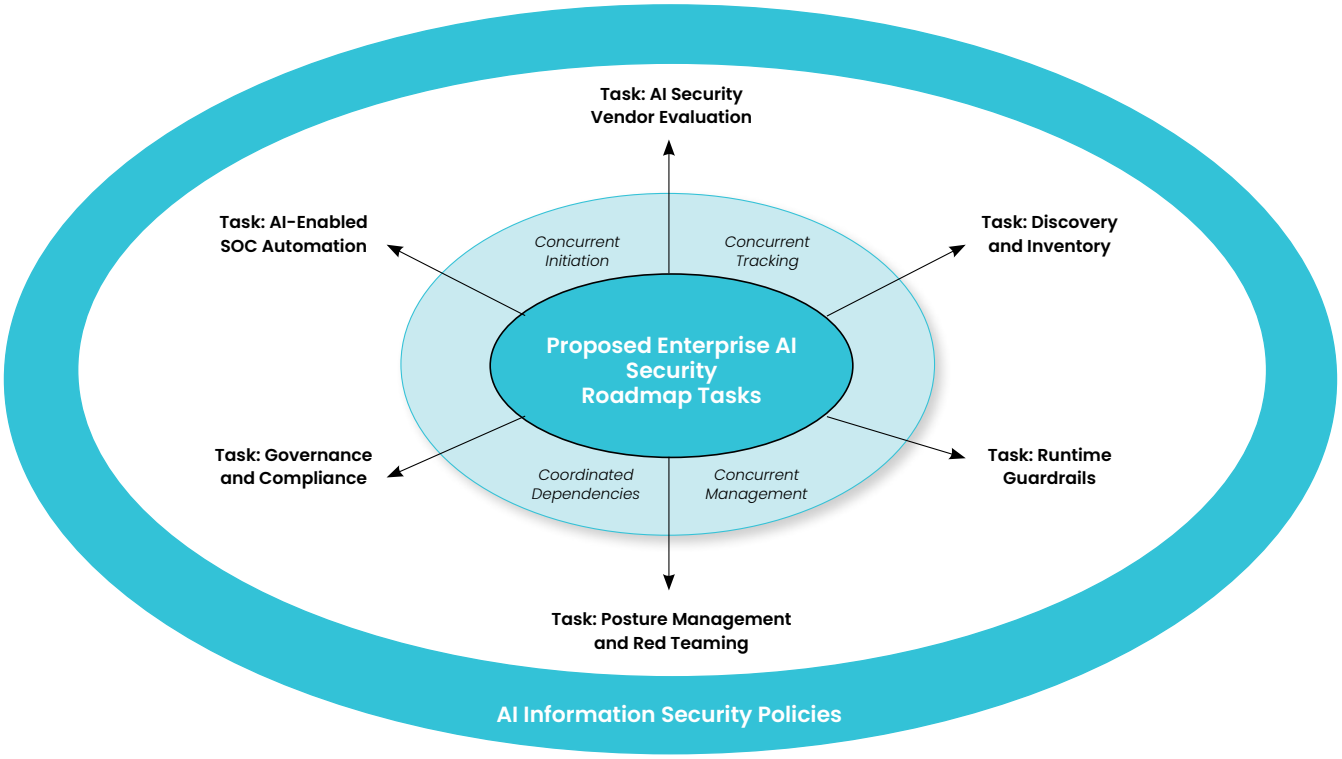


Figure 3-2. Foundational Support from AI Security Policies

Many well-intentioned security teams are now deploying controls for Generative AI and machine learning without first identifying and documenting an organizational policy for securing such systems. At TAG, we believe that an enterprise AI-ISP must complement introduction of AI controls or architectures. This belief is based on decades of experience developing security policies for many different scenarios.

One challenge, however, is that enterprise teams are adopting AI in different ways, with some embedding LLMs into external customer workflows, and others fine-tuning private AI systems behind a traditional firewall. Still others are experimenting with agent-based orchestration or federated learning. This variety, which we observed during our research and which we see every day in our work at TAG, complicates development of a security policy that covers all scenarios.

Nevertheless, one could argue that it is precisely this diversity that underscores the need for a clear AI-ISP. Without a well-structured policy framework that reflects an organization's unique AI ambitions, risk appetite, and data responsibilities, even the most sophisticated security tooling might not solve the correct problem. Our observation is that many proof-of-concept (POC) deployments of AI security platforms suffer from this lack of underlying requirements.

POLICY REQUIREMENTS

We offer here a generic set of proposed AI security policy requirements that can be tailored into an AI-ISP suitable for any enterprise security team. Our team at TAG has designed it to be modular, editable, and aligned to real-world use cases. We use the TAG Taxonomy as the basis for this AI-ISP, since the categories of enterprise security practice will represent a superset of how most modern CISO-led programs are aligned today.

We should start by saying that most of the security frameworks such as MITRE ATT&CK, NIST CSF 2.0, and others are not well-suited, in our opinion, to developing an AI-ISP. MITRE ATT&CK, for instance, lays out a logical collection of attack methods, which we all find useful as a guide to offensive tactics. While this influences an AI-ISP, it is a poor guide for one. (We do explain, however, later in this chapter how the AI-ISP can be mapped to certain frameworks)

What is really needed is a guide that covers the cybersecurity-related methods used in practice by enterprise security teams. Our TAG Taxonomy, which we provide below, is just such a framework. It serves as an easy-to-use and complete list of enterprise security practices in use today (see Figure 3-3).

In the sections below, we offer a proposed functional security policy statement consistent with the 100 categories of security practices in the taxonomy (i.e., from 1.1 to 20.5). The requirements can be viewed collectively as our proposed AI-ISP. Readers are encouraged to use this general policy framework as a baseline on which to tailor, adjust, append, or otherwise improve their existing security policy requirements statements.

1 Application Security 1.1 API Security 1.2 Application Security Testing 1.3 Application Security Posture Mgmt. 1.4 Runtime Application Security 1.5 SBOM/SCA	**6 Email Security** 6.1 Anti-Phishing Tools 6.2 DMARC 6.3 Email Encryption 6.4 Phish Testing and Training 6.5 Secure Email Gateway	**11 Identity and Access Management** 11.1 User and Workload IAM Platforms 11.2 Authentication 11.3 identity, Anti-Fraud, and KYC 11.4 Identity Governance and Admin. 11.5 Privileged Access Management	**16 Operational Technology Security** 16.1 ICS/OT Infrastructure Security 16.2 ICS/OT Network Visibility 16.3 Unidirectional Gateway 16.4 Vehicle Security 16.5 Zero Trust OT
2 Attack Surface Management 2.1 Bug Bounty Services 2.2 External Attack Surface Management 2.3 Automated Pen Testing/Red Teams 2.4 BAS/CTEM 2.5 Security Ratings Platforms	**7 Encryption and PKI** 7.1 Certification Authority (CA) 7.2 Data Encryption 7.3 Secrets Management 7.4 Certificate Lifecycle Mgmt. 7.5 Post-Quantum Cryptography	**12 Security Operations and Response** 12.1 Data Forensics and eDiscovery 12.2 Incident Response 12.3 SIEM Platforms 12.4 SOC/SOAR/Co-Pilot Support 12.5 Threat Hunting	**17 Security Professional Services** 17.1 Penetration Testing 17.2 Security Consulting and Assessment 17.3 Security Industry Research/Advisory 17.4 Security Training 17.5 Security Solution Provider
3 AI Security 3.1 AI Development Lifecycle Security 3.2 AI Runtime Guard Rails 3.3 AI Red Teaming and Testing 3.4 AI Supply Chain Security 3.5 AI Governance, Policy, and Compliance	**8 Endpoint Protection** 8.1 Anti-Malware Software 8.2 Browser Isolation 8.3 Content Disarm and Reconstruction 8.4 Endpoint Detection and Response 8.5 Security Enhanced Browser	**13 Managed Security Services** 13.1 DDOS Security 13.2 Managed Detection and Response 13.3 Managed Security Services Platform 13.4 Network Detection and Response 13.5 XDR Services	**18 Software Lifecycle Security** 18.1 Deepfake Security 18.2 Kubernetes Security 18.3 Container Scanning 18.4 DevSecOps Platforms 18.5 Infrastructure-as-Code Security
4 Cloud Security 4.1 SaaS Security Posture Mgmt. 4.2 Cloud Infrastructure Entitlement Mgmt. 4.3 Cloud Security Posture Management 4.4 Cloud Workload Protection Platform 4.5 Microsegmentation	**9 Enterprise IT Infrastructure** 9.1 Asset Inventory 9.2 Backup Platform 9.3 Infrastructure Resilience 9.4 Insider Threat Protection 9.5 Secure Sharing and Collaboration	**14 Mobility Security** 14.1 IOT Security 14.2 Mobile App Security 14.3 Mobile Device Management 14.4 Mobile Device Security 14.5 Mobility Infrastructure Security	**19 Threat and Vulnerability Management** 19.1 Digital Risk Protection 19.2 Security Scanning 19.3 Third Party Risk Management 19.4 Threat and Vulnerability Platform 19.5 Threat Intelligence
5 Data Security 5.1 Data Security Posture Mgmt. 5.2 Data Access Governance 5.3 Data Discovery and Classification 5.4 Data Leakage Protection 5.5 Data Privacy Platform	**10 Governance, Risk, and Compliance** 10.1 Continuous Compliance 10.2 Cyber Insurance 10.3 Incident Reporting 10.4 GRC Platform 10.5 Risk Management Platform	**15 Network Security** 15.1 Network Access Control 15.2 Next Generation Firewalls 15.3 Secure Access Service Edge (SSE) 15.4 Virtual Private Networks 15.5 Zero Trust Network Access	**20 Web Security** 20.1 Bot Management 20.2 Disinformation Security 20.3 Secure Web Gateway 20.4 Web Application Firewall 20.5 Website Scanning

Figure 3-3. TAG Cybersecurity Taxonomy

1.0 APPLICATION SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences application programming interface (API) security, application security testing, application security posture management, runtime application security, and software bill of materials (SBOM)/software compositional analysis (SCA).

1.1 API Security

All AI-related APIs must be secured using authentication, rate limiting, encryption, and logging. And they must be registered with the enterprise security team.

1.2 Application Security Testing

All AI software and applications must undergo static and dynamic security testing prior to deployment and upon significant update.

1.3 Application Security Posture Management

AI applications must be continuously assessed and monitored for configuration drift, risk scoring, and compliance with secure development policies.

1.4 Runtime Application Security

AI application runtimes must include protections to detect and prevent unauthorized behaviors or malicious inputs during operation.

1.5 SBOM/SCA

Software bills of materials and software composition analysis must be produced and reviewed for all AI components to ensure known vulnerabilities are addressed.

2.0 ATTACK SURFACE MANAGEMENT POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences bug bounty services, external attack surface management (EASM), automated pen testing/red teams, breach and attack simulation (BAS)/cyber threat exposure management (CTEM), and security ratings platform.

2.1 Bug Bounty Services

Public or private vulnerability disclosure programs must be established for externally accessible AI services to identify weaknesses not caught in testing.

2.2 External Attack Surface Management

Enterprise users must ensure all public AI assets are continuously discovered, inventoried, and monitored for exposure or misconfiguration.

2.3 Automated Pen Testing/Red Teams

AI systems must be regularly subjected to automated penetration tests or red team simulations to evaluate their security readiness.

2.4 BAS/CTEM

AI environments must be integrated into breach and attack simulation (BAS) or continuous threat exposure management (CTEM) platforms to assess real-world exposure.

2.5 Security Ratings Platforms

AI suppliers and third-party vendors must be assessed using security ratings or scoring platforms as part of due diligence and risk monitoring.

3.0 AI SECURITY POLICIES

The security policy rules below introduce how the AI use-case introduces new considerations (i.e., they would not exist if the AI use-case was not present) in the area of AI development lifecycle security, AI runtime guardrails, AI red teaming and testing, AI supply-chain security, and AI governance, policy, and compliance.

3.1 AI Development Lifecycle Security

All AI development must integrate security controls and reviews across every stage of the model and application lifecycle.

3.2 AI RUNTIME GUARDRAILS

AI systems must include runtime constraints to prevent undesirable outputs, hallucinations, or unsafe autonomous actions.

3.3 AI Red Teaming and Testing

AI systems must be subjected to adversarial testing and red teaming to identify weaknesses in model behavior, control boundaries, and response to manipulation.

3.4 AI Supply Chain Security

Third-party AI tools, models, and training datasets must be validated for integrity, provenance, and security prior to use in enterprise systems.

3.5 AI Governance, Policy, and Compliance

Use of AI must comply with enterprise AI governance policies and external mandates, including logging, auditability, and transparency.

4.0 CLOUD SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences SaaS security posture management (SSPM), cloud infrastructure entitlement management (CIEM), cloud security posture management (CSPM), cloud workload protection platform (CWPP), and microsegmentation.

4.1 SaaS Security Posture Management

All AI-enabled SaaS platforms must be monitored for misconfigurations, role assignments, and access controls consistent with enterprise standards.

4.2 Cloud Infrastructure Entitlement Management

Cloud-based AI deployments must ensure least-privilege access for all identities and workloads via entitlement reviews and automated enforcement.

4.3 Cloud Security Posture Management

AI cloud environments must be continuously scanned for misconfigurations, compliance issues, and alignment to baseline security standards.

4.4 Cloud Workload Protection Platform

AI workloads deployed in the cloud must be monitored with real-time protection platforms that identify anomalies and block attacks.

4.5 Microsegmentation

AI infrastructure must implement microsegmentation to isolate components, limit east-west traffic, and reduce lateral movement risk.

5.0 DATA SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences data security posture management (DSPM), data access governance (DAG), data discovery and classification (DDC), data leakage protection (DLP), and data privacy platform.

5.1 Data Security Posture Management

All datasets used in AI must be inventoried and monitored for security posture, including access control, classification, and encryption.

5.2 Data Access Governance

Access to AI training, inference, or sensitive datasets must be governed by enterprise-approved access policies and logged.

5.3 Data Discovery and Classification

Data used for AI training or inference must be automatically discovered, classified, and labeled according to sensitivity.

5.4 Data Leakage Protection

AI systems must be configured to prevent unauthorized transmission or leakage of sensitive data through outputs or logs.

5.5 Data Privacy Platform

AI systems that handle personal data must integrate with enterprise privacy platforms to enforce consent, redaction, and data minimization.

6.0 EMAIL SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences anti-phishing tools, DMARC, email encryption, phish testing and training, and secure email gateway.

6.1 Anti-Phishing Tools

AI-based communications or summarization systems must include or integrate with phishing detection tools to prevent propagation of scams.

6.2 DMARC

All enterprise email domains used for AI systems must enforce DMARC, SPF, and DKIM to prevent spoofing and impersonation.

6.3 Email Encryption

AI tools that interact via email must use end-to-end encryption where feasible to protect sensitive input/output.

6.4 Phish Testing and Training

Users interacting with generative AI systems via email must undergo training and periodic testing to identify phishing risks.

6.5 Secure Email Gateway

AI-generated or -consuming email must transit secure gateways that scan for malware, spoofing, and policy violations.

7.0 ENCRYPTION AND PKI POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences certification authority (CA), data encryption, secrets management, certificate lifecycle management, and post-quantum cryptography.

7.1 Certification Authority (CA)

All AI systems using certificates must rely on trusted enterprise-approved CAs and enforce certificate validation.

7.2 Data Encryption

AI data stores, models, and communication channels must use strong encryption aligned with enterprise cryptographic policies.

7.3 Secrets Management

AI systems must store API keys, tokens, and credentials in secure secrets management solutions with rotation policies.

7.4 Certificate Lifecycle Management

All certificates used in AI systems must be tracked, renewed, and revoked as needed to prevent lapses or misuse.

7.5 Post-Quantum Cryptography

Planning for quantum-safe cryptography must be incorporated into AI systems expected to handle long-lived sensitive data.

8.0 ENDPOINT PROTECTION POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences anti-malware software, browser isolation, content disarm and reconstruction, endpoint detection and response, and security enhanced browser.

8.1 Anti-Malware Software

Endpoints used for AI development or model execution must run enterprise-approved anti-malware solutions with real-time protection.

8.2 Browser Isolation

Web-based AI services must be accessed through browser isolation where feasible to prevent drive-by attacks or session hijacking.

8.3 Content Disarm and Reconstruction

AI systems accepting document inputs must use CDR technologies to sanitize embedded content and prevent malware execution.

8.4 Endpoint Detection and Response

All devices executing AI models or tools must have EDR agents that detect malicious activity and report it centrally.

8.5 Security Enhanced Browser

AI users must access sensitive data or services only through hardened enterprise browsers with policy-enforced settings.

9.0 ENTERPRISE IT INFRASTRUCTURE SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences asset inventory, backup platform, infrastructure resilience, insider threat protection, and secure sharing and collaboration.

9.1 Asset Inventory

All AI-related systems, including model endpoints, APIs, and storage, must be inventoried and tracked in the CMDB.

9.2 Backup Platform

Critical AI models, datasets, and metadata must be regularly backed up using enterprise solutions with secure recovery procedures.

9.3 Infrastructure Resilience

AI hosting environments must be architected for resilience against outages, attacks, and system degradation.

9.4 Insider Threat Protection

AI systems must include controls to detect and respond to malicious or negligent insider behavior, especially data exfiltration.

9.5 Secure Sharing and Collaboration

Any collaboration on AI projects must use secure enterprise platforms that enforce identity verification and content controls.

10.0 GOVERNANCE, RISK, AND COMPLIANCE (GRC) POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences continuous compliance, cyber insurance, incident reporting, GRC platform, and risk management platforms.

10.1 Continuous Compliance

AI systems must be continuously assessed for compliance with internal policies and external regulations, with automated alerts on deviation.

10.2 Cyber Insurance

All AI systems supporting critical functions must be included in cyber insurance coverage assessments and meet insurability criteria.

10.3 Incident Reporting

Security incidents involving AI systems must be reported through established enterprise channels and include model behavior context and logs.

10.4 GRC Platform

All AI security policy controls and exceptions must be tracked and reviewed within the enterprise GRC platform for risk oversight.

10.5 Risk Management Platform

AI-specific risks must be documented, categorized, and prioritized within the enterprise risk platform to support mitigation efforts.

11.0 IDENTITY AND ACCESS MANAGEMENT (IAM) POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences user and workload IAM platforms, authentication, identity/anti-fraud/KYC, identity governance and administration, and privileged access management.

11.1 User and Workload IAM Platforms

Access to AI systems must be managed through centralized IAM platforms supporting human, service, and agent-based identity types.

11.2 Authentication

All access to AI infrastructure and interfaces must require strong, multifactor authentication aligned with enterprise policy.

11.3 Identity, Anti-Fraud, and KYC

AI systems that handle user onboarding or identity validation must adhere to enterprise KYC and fraud prevention requirements.

11.4 Identity Governance and Administration

Access to AI tools and data must be regularly reviewed and governed using role-based controls and automated provisioning.

11.5 Privileged Access Management

Administrative access to AI environments must be managed through privileged access controls, with session logging and just-in-time access where possible.

12.0 SECURITY OPERATIONS AND RESPONSE POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences data forensics and eDiscovery, incident response, SIEM platforms, SOC/SOAR/co-pilot support, and threat hunting.

12.1 Data Forensics and eDiscovery

All AI-related data—inputs, outputs, prompts, and logs—must be retained and searchable for legal, regulatory, and forensic review.

12.2 Incident Response

Incident response plans must explicitly address AI-specific risks, such as model manipulation, prompt injection, or data leakage.

12.3 SIEM Platforms

Logs from AI systems must be forwarded to enterprise SIEM platforms with tagging to distinguish model activities and inference calls.

12.4 SOC/SOAR/Co-Pilot Support

Security operations must integrate AI system telemetry into SOC visibility and automate responses where appropriate via SOAR platforms.

12.5 Threat Hunting

Threat hunting teams must include AI infrastructure and models as part of their hypothesis-driven search for anomalies or compromise.

13.0 MANAGED SECURITY SERVICES POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences DDoS security, managed detection and response, managed security services platform, network detection and response, and XDR services.

13.1 DDoS Security

AI inference services exposed to external users must be protected with enterprise DDoS mitigation solutions and monitored for volumetric abuse.

13.2 Managed Detection and Response

Where MDR services are used, AI system logs and activities must be included in the monitored scope and reviewed regularly.

13.3 Managed Security Services Platform

Outsourced security providers managing AI assets must adhere to internal policy requirements and provide clear SLAs on AI monitoring.

13.4 Network Detection and Response

AI workloads must be visible to NDR platforms with detection tuned for lateral movement, exfiltration, and anomalous traffic.

13.5 XDR Services

AI endpoint, cloud, and data signals must be correlated in extended detection platforms to support holistic threat visibility.

14.0 MOBILITY SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences IOT security, mobile app security, mobile device management, mobile device security, and mobility infrastructure security.

14.1 IOT Security

Where AI is deployed to edge or IoT systems, device security controls and firmware validation must be enforced.

14.2 Mobile App Security

AI functionality integrated into mobile apps must be protected against tampering, reverse engineering, and unsafe permissions.

14.3 Mobile Device Management

Enterprise mobile devices used for AI data collection or interaction must be enrolled in MDM and policy enforced.

14.4 Mobile Device Security

AI models deployed to mobile devices must be sandboxed, monitored, and able to be remotely disabled if compromised.

14.5 Mobility Infrastructure Security

AI-related APIs and services must be secured when accessed over mobile infrastructure, with VPN or private access when appropriate.

15.0 NETWORK SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences network access control, next generation firewalls, secure access service edge (SASE), virtual private networks, and zero trust network access.

15.1 Network Access Control

AI infrastructure must participate in NAC enforcement, with endpoint posture verification prior to joining enterprise networks.

15.2 Next Generation Firewalls

All AI traffic must pass through NGFWs that can inspect application-layer behavior and enforce policies.

15.3 Secure Access Service Edge (SSE)

Remote or distributed access to AI environments must use enterprise SSE platforms for access control and threat protection.

15.4 Virtual Private Networks

AI development and administration activities conducted remotely must use enterprise VPN solutions with logging and MFA.

15.5 Zero Trust Network Access

AI systems must be accessed via ZTNA architecture, enforcing identity-aware, context-sensitive, and least-privilege access decisions.

16.0 OPERATIONAL TECHNOLOGY (OT) SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences ICS/OT infrastructure security, ICS/OT network visibility, unidirectional gateways, vehicle security, and zero trust OT.

16.1 ICS/OT Infrastructure Security

AI applied in ICS/OT environments must not compromise physical safety or integrity and must follow hardened deployment protocols.

16.2 ICS/OT Network Visibility

AI systems connected to operational networks must be monitored via OT visibility platforms to ensure protocol and behavior compliance.

16.3 Unidirectional Gateway

When AI inference data needs to be moved from sensitive OT systems, unidirectional gateways or data diodes must be used.

16.4 Vehicle Security

AI systems embedded in vehicles must be tested for safe model operation and secured against remote or unauthorized manipulation.

16.5 Zero Trust OT

Zero trust principles must be applied to AI agents interacting with OT systems, especially those with autonomous decision-making capabilities.

17.0 SECURITY PROFESSIONAL SERVICES POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences penetration testing, security consulting and assessment, security industry research/advisory, security training, and security solution providers.

17.1 Penetration Testing

AI components must be included in enterprise penetration testing scope, including LLM endpoints, APIs, and model logic.

17.2 Security Consulting and Assessment

External security assessments of AI deployments must be performed annually or upon major update to identify gaps and risks.

17.3 Security Industry Research/Advisory

Teams developing or deploying AI must stay informed on emerging security practices and adversarial research affecting model integrity.

17.4 Security Training

All developers and users of AI systems must complete training on AI-specific risks, threats, and secure development practices.

17.5 Security Solution Provider

Third-party AI security tools must be evaluated and procured according to enterprise procurement and solution assessment policy.

18.0 SOFTWARE LIFECYCLE SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences deepfake security, Kubernetes security, container scanning, DevSecOps platforms, and infrastructure-as-code security.

18.1 Deepfake Security

AI-generated media must be assessed for authenticity and watermarked or labeled where possible to prevent misuse or disinformation.

18.2 Kubernetes Security

AI workloads running in Kubernetes must use hardened configurations, limit privileges, and monitor for pod or API abuse.

18.3 Container Scanning

AI-related containers must be scanned pre-deployment for vulnerabilities, embedded secrets, and outdated libraries.

18.4 DevSecOps Platforms

AI CI/CD pipelines must integrate DevSecOps tools to enforce policy gates, artifact signing, and secure dependency management.

18.5 Infrastructure-as-Code Security

IaC used to deploy AI environments must be scanned and validated for security posture, including access, network, and storage settings.

19.0 THREAT AND VULNERABILITY MANAGEMENT POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences digital risk protection, security scanning, third party risk management, threat and vulnerability management, and threat intelligence.

19.1 Digital Risk Protection

AI-related brand impersonation or misuse must be monitored externally via DRP platforms, with takedown capabilities where needed.

19.2 Security Scanning

AI code, configurations, and models must be scanned continuously for vulnerabilities and weaknesses as part of the DevOps lifecycle.

19.3 Third Party Risk Management

Vendors providing AI capabilities or datasets must be evaluated under third-party risk frameworks with contractual security obligations.

19.4 Threat and Vulnerability Platform

Threat intelligence platforms must integrate AI-specific threats such as model evasion techniques or prompt injection strategies.

19.5 Threat Intelligence

Teams must collect and analyze AI threat intelligence to inform defensive controls, adversary simulation, and model hardening.

20.0 WEB SECURITY POLICIES FOR AI

The security policy rules below introduce how the AI use-case influences bot management, disinformation security, secure web gateway, web application firewall, and website scanning.

20.1 Bot Management

AI services exposed to web users must distinguish between human and automated traffic and mitigate malicious bots.

20.2 Disinformation Security

AI-generated content must be subject to review and labeling processes to prevent unintentional propagation of disinformation.

20.3 Secure Web Gateway

Access to external AI tools must pass through secure web gateways that enforce category-based filtering and malware inspection.

20.4 Web Application Firewall

All AI-driven web applications must be protected with WAFs configured to block injection, enumeration, and evasion attempts.

20.5 Website Scanning

Public-facing AI applications must be regularly scanned for vulnerabilities and misconfigurations as part of routine hygiene.

POLICY ALIGNMENT WITH COMMON USE CASES

Our view is that the AI-ISP presented above aligns well with the enterprise AI adoption patterns we've observed across the industry. Whether an organization is embedding LLMs into customer-facing workflows, fine-tuning foundation models behind a firewall, experimenting with orchestration via autonomous agents, or deploying federated learning at the edge, the policy framework addresses applicable risk vectors.

The AI-ISP can be selectively tailored or prioritized depending on where the organization sits along its AI maturity curve. A team exploring Generative AI in sandboxed environments, for example, should find that basic policies on SaaS posture and access control are sufficient to begin. More advanced teams integrating AI into production systems, however, will need deeper controls on AI system integrity, supply chain verification, and behavior monitoring.

The AI-ISP thus allows for tiered alignment, giving CISOs and governance teams the flexibility to adapt requirements based on mission-criticality, regulatory context, and risk appetite. In short, the AI-ISP does not enforce a one-size-fits-all standard. Rather, it provides a practical, use-case-aware policy foundation that any enterprise can operationalize across diverse AI implementation scenarios. We hope it is helpful to you.

POLICY ALIGNMENT WITH IMPORTANT AI FRAMEWORKS

Despite our comments above regarding alignment with MITRE ATT&CK and other attack-oriented frameworks, we do understand the need for practitioners to map their policies to certain requirements, usually dictated by auditors or certifying agencies in their review efforts. Below, we outline the type of alignment one will obtain from the AI-ISP requirements listed above, which can be integrated into a GRC platform.

Specifically, the AI-ISP aligns with the NIST AI Risk Management Framework (AI RMF) by enabling identification, measurement, management, and governance of AI-specific risks. Multiple policy statements map directly to the core NIST functions, including requirements around transparency (e.g., SIEM processing), robustness (e.g., penetration testing), and accountability (e.g., policy enforcement through GRC platforms).

The AI-ISP also reflects principles from the OWASP Top 10 for LLMs, covering key attack vectors such as data leakage protection (DLP). Specific deployed controls consistent with the AI-ISP such as runtime guardrails, output monitoring, red teaming, and prompt hygiene directly mitigate the threats articulated by OWASP and serve to reduce AI misuse.

In addition, the AI-ISP supports guidance from the Executive Order on Safe, Secure, and Trustworthy AI, by emphasizing supply chain transparency, privacy safeguards, and accountability in the use of AI. The policy also aligns with ISO/IEC 42001, which codifies AI Management Systems, by ensuring that technical safeguards, governance structures, and stakeholder obligations are all embedded into a single comprehensive policy framework.

THE BOTTOM LINE

In closing, we believe that the AI-ISP presented above offers a defensible and standards-aligned policy foundation upon which organizations can build assurance, demonstrate compliance, and confidently scale AI capabilities.

Our advice is to cut-and-paste the policies above into your own document and then begin tailoring the specific policies to your local context. And, dare we say, you might even engage ChatGPT (or equivalents) to help. (Just make sure you go through the output carefully.)

Good luck.

CHAPTER 4

DEVELOPING AN AI RISK-TIERING APPROACH

Here is a model that can help to establish a better understanding of the challenges that emerge with use of artificial intelligence.

The goal in this chapter is to provide an information security risk-tiering model for the most common AI use cases. We believe this will be timely, since many organizations are beginning to be asked by their leadership to identify the likelihood of attacks to AI models, applications, and systems, as well as the true consequences of such attacks. And, as all security practitioners will know, risk is measured as the confluence of likelihood and consequence.

There are generic AI threat models, such as **MITRE ATLAS**. These tend to propose collections of breach signatures, which play a useful role in designing AI tests or in comparing the efficacy of different AI security vendors. But interpreting lists of threat tactics in the context of real business environments can be difficult. This has led us to follow a different approach.

Specifically, we will focus on the elements of risk-tiering in the context of AI-related business tasks and activities that we would expect to see in a typical business setting. For example, the handling of sensitive data is different than the handling of mundane, non-sensitive data. This suggests the introduction of AI in these two scenarios would have different consequences. This type of analysis is the basis for the risk-tiering approach we will explain.

A WORD ON USING RISK FOR FUNDING

To create an accurate AI security risk model, the enterprise security team must take the time to understand and document the true purpose of AI usage in their organization. Technically, there are four possibilities:

1. **Lower Costs:** AI usage can provide cost reductions by replacing expensive humans with less expensive neural networks.
2. **New Offerings:** AI usage can lead to new service introductions through the extra-human capabilities of AI.
3. **Research Ideas:** AI usage can drive R&D investigations to identify new ideas and innovations enabled by AI.
4. **Market Value:** Companies can introduce AI to create stakeholder value through perception of future promise.

Let's start by being honest: While new offerings, research ideas, and market value are all reasonable objectives, virtually every AI initiative that we've reviewed of any substance has been squarely focused on reducing expense. It has been our observation that the primary purpose for the deployment and use of AI is lower cost (see Figure 4-1).

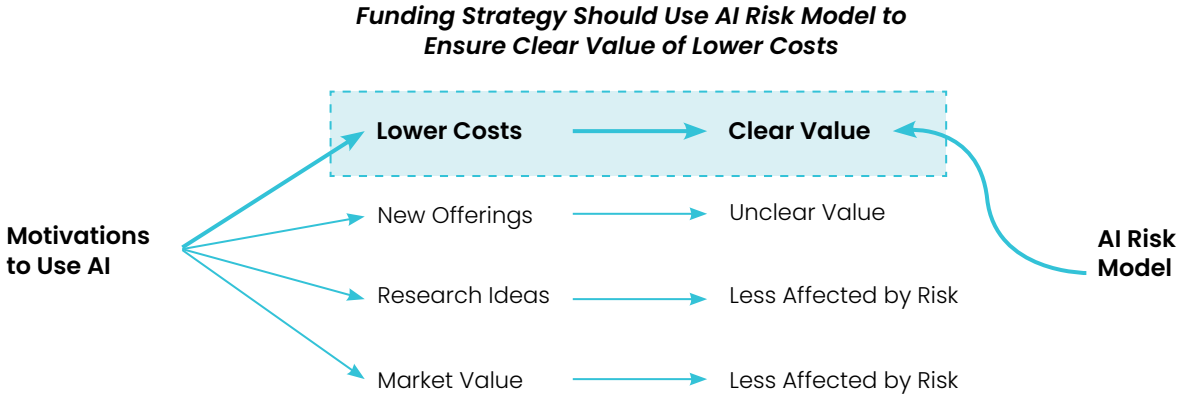


Figure 4-1. Risk Focus to Obtain Funding for AI Security

What this implies is that if CISOs and their teams intend to use AI risk models such as the tiering approach shown in this chapter, then they will have to connect such risk to the possibility that the cost reduction objectives would not be met. Stated otherwise: Cost reduction objectives are balanced by the need to address the risks associated with AI, which could subvert the successful deployment and use of the technology.

CASE STUDY: BLACK LIGHT DATA RISK

Many CISOs will point to a significant risk that emerges in the context of AI usage in their enterprise: specifically, the introduction of co-pilots to help employees locate information across their network. What CISOs have noticed, however, is that when users are over-permissioned, which is common, then the risk becomes clear that they might find information that they otherwise would not have noticed that they could access.

The canonical example involves an employee using a co-pilot to ask what their supervisor makes as a salary, which is obviously something companies will keep confidential. What happens with AI technology is that it is entirely possible that the tools will locate one or more files (e.g., ExecPay.pdf, ExecBonus.pdf) that include this information, which is then shared back with the requesting user.

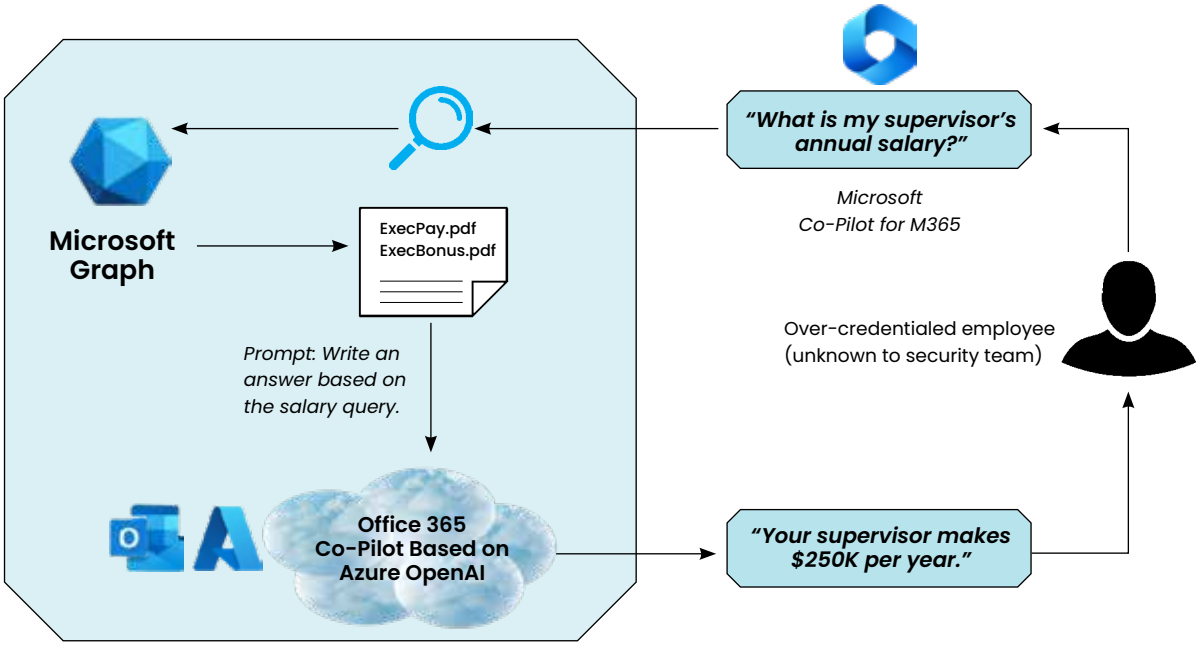


Figure 4-2. Risk of Data Exposure from AI Co-Pilot Tools

At TAG, we refer to this risk as a black light problem, because it is technically not the AI technology that is the problem, but rather an underlying data security weakness. For this reason, we strongly recommend that enterprise security teams ensure that they have engaged a world-class data security partner to address it. Here are some thoughts on this topic from Varonis.

VARONIS AN INTERVIEW WITH YAKI FAITELSON, CEO, VARONIS

We recently had the opportunity to discuss this data security exposure issue with Yaki Faitelson, CEO of Varonis. During our discussion, he focused on the issue of over-permissioned users exposing sensitive data when they have access to AI co-pilots in the enterprise. It was clear that many Varonis customers are addressing this risk today. We include brief excerpts from his comments.

TAG: Yaki, enterprises are adopting AI co-pilots and agents, but many still struggle with data governance. How does having over-permissioned users make this problem worse in the AI era?

FAITELSON: AI behaves exactly as designed, and that's the problem. Humans may technically have access to thousands of files they've never opened, never noticed, and would never realistically consume. AI doesn't work that way. It ingests everything it can see, instantly and indiscriminately.

When an AI agent inherits over-permissioned access, it doesn't just skim data, it absorbs it at hyperspeed, connects the dots humans never would, and operationalizes it. In the AI era, excessive access isn't dormant risk anymore. It's active risk by default. And that's why governance failures suddenly matter so much. AI doesn't wait for intent, curiosity, or mistakes. It uses all of its access, every time.

Continued

VARONIS *Continued*

TAG: What foundational approach should security teams take to mitigate those risks, and how does Varonis address it?

FAITELSON: Most security programs are still built on a comforting illusion: that you can prevent bad things from happening if you just add enough tools. That model is breaking.

The foundational shift is accepting that access – not malware, not exploits – is the real attack surface. If identity is the new perimeter, then agency is the new weapon. Security teams need both sides of the equation: the ability to detect and respond to threats in seconds and the discipline to proactively minimize the blast radius before anything happens. AI makes that proactive stance non-negotiable.

Varonis is built around that reality. We automatically and continuously reduce excessive access so when something goes wrong, and it will, the impact is fundamentally limited. When an insider goes rogue, an AI agent misbehaves, or identities are stolen, the impact is constrained by design. Automation is how you keep up with machines.

TAG: Can you give an example of how Varonis helps protect real enterprise data in AI contexts?

FAITELSON: We see companies respond to AI risk with blunt instruments: block the tool or allow it. The problem is, without understanding the data, those controls are guesswork.

In one case, a customer's AI copilot began indexing sensitive HR and legal data, not because it was malicious, but because access was wide open. Varonis recognized the sensitivity of the data, understood which identities and use cases were inappropriate, and automatically removed excess access without shutting AI down.

And that's the difference. You can't secure AI without data context. Only a true data security platform can set – and enforce – the granular rules of the road, instead of relying on all-or-nothing controls while hoping for the best.

A THREE-TIERED SECURITY RISK MODEL FOR AI USE CASES

We propose now a simple classification of enterprise AI use cases (e.g., the black light data problem examined above) into high, medium, or low information security risk tiers. The objective is to determine the level and depth of pre-deployment security assessments and review what's required to identify the proper controls that can mitigate threats to confidentiality, integrity, and availability. Readers should tailor the criteria in each tier to their local context.

TIER 3: HIGH INFORMATION SECURITY RISK

This highest risk tier is designed to include those use cases where insecure use of AI is either unusually high in likelihood or can lead to significant consequences. Obviously, the criteria must be adjusted for different environments (e.g., industrial control systems would need additional focus on life or safety critical consequences), but the general idea here should be evident. Here are the inclusion criteria for the AI use case of interest:

1. **Data Sensitivity:** This involves AI processing sensitive personal data (PII, PHI), financial data, regulated data (e.g., GDPR, HIPAA, GLBA), or trade secrets.

2. **System Exposure:** This includes externally facing AI systems that are accessible by customers, vendors, or the public.

3. **Autonomy & Control:** In this case, the AI system has significant automation or decision authority without human review ("autonomous AI" or agentic AI).

4. **Data Storage:** This involves the AI system storing or transmitting outputs and inputs across untrusted networks or with third-party vendors (e.g., via API).

5. **Downstream Access:** Here, AI outputs may directly or indirectly trigger automated actions in critical systems (e.g., security operations, financial transactions).

6. **Model Customization:** This involves fine-tuning or training of models using proprietary enterprise datasets.

7. **Compliance Impact:** This is where a failure in the AI system may result in regulatory violations or legal liability.

Specific assessments, reviews, and controls will vary between different organizations, often differentiated by the size of their revenue. Accordingly, the types of cyber preventions and mitigations that we would expect to see for use cases in this high information security risk category include the following, many of which are being covered by commercial vendors developing AI security solutions:

- **Threat modeling and red team assessment (e.g., jailbreak and prompt injection testing).**
- **Security Architecture Review and penetration testing**
- **Model input/output logging and auditability controls**
- **Data loss prevention (DLP) and encryption in transit and at rest**
- **Privacy impact assessment and model explainability checks**

TIER 2: MEDIUM INFORMATION SECURITY RISK

The inclusion criteria we've identified for this category of use cases is derived from our estimation that either the likelihood of attack is lower or the consequences of a successful AI-related breach are lower. Again, as with the previous grouping, local tailoring by security teams is recommended to ensure alignment with specific business objectives and conditions. Here are the inclusion criteria:

1. **Data Sensitivity:** This involves internal business data that is not regulated but proprietary or confidential.

2. **System Exposure:** In this case, the AI application is internally facing but used across departments or business units.

3. **Automation Scope:** Human-in-the-loop is present, but the AI significantly augments or accelerates internal workflows (e.g., AI co-pilots for code or policy generation).

4. **Integration Depth:** The AI is integrated into shared productivity platforms or enterprise infrastructure (e.g., Microsoft Copilot).

5. **Model Interaction:** This case uses public foundation models via API, but with safeguards (e.g., no fine-tuning on sensitive data).

6. **Data Sharing:** This involves some interactions with third-party vendors or tools (e.g., through SaaS platforms with access to enterprise data).

As suggested above, while specific assessments, reviews, and controls will vary, the types of preventions and mitigations we would expect to see for use cases in this medium information security risk category include the following—again, highly correlated to the types of commercial offerings we see in the AI security marketplace (covered in Chapter 6):

- **Security configuration review of third-party integrations (API, SaaS)**
- **Threat assessment for prompt misuse or LLM overreliance**
- **Role-based access control (RBAC) and identity protection**
- **Basic PII redaction or filtering mechanisms**
- **Risk acceptance and business owner sign-off for model use**

TIER 1: LOW INFORMATION SECURITY RISK

The inclusion criteria we've identified for this category is derived from our estimation that either the likelihood of attack is low or negligible, or the consequences of a successful AI-related breach are low or negligible. Again, as with the previous groupings, local tailoring by security teams is recommended to ensure alignment with specific business objectives and conditions. Here are the criteria:

1. **Data Sensitivity:** Here, AI systems use only public, anonymized, or synthetic data with no identifiable attributes.
2. **System Exposure:** This involves internal or sandboxed use by a restricted set of users (e.g., AI research sandbox).
3. **Model Scope:** In this case, no fine-tuning or training is required but this assumes off-the-shelf model use only.
4. **Automation Impact:** This assumes read-only or advisory functions with no decision-making authority.
5. **Business Impact:** This assumes that the outputs are not being used to drive operational or financial decisions.
6. **Data Flow:** In this case, data remains in enterprise-controlled environments with no third-party transmission.

Again, as suggested above, while specific assessments, reviews, and controls will vary, the types of preventions and mitigations we would expect to see for use cases in this low information security risk category include the following:

- **Basic data governance checks and usage policy enforcement**
- **Endpoint control and monitoring**
- **Risk documentation with minimal approval requirements**
- **Use limited to pilot or sandbox environments with expiration timelines**

HOW TO USE THIS TIERING SYSTEM

This risk model can and should be applied by enterprise information security teams as a pre-deployment classification mechanism for any new AI project, integration, or vendor evaluation. Once the tier is assigned, a predefined level of assessment (light, moderate, or full) can be initiated, ensuring that enterprise risk remains proportionate to the scale and sensitivity of the AI use case.

The diagram in Figure 4-3 below should be helpful in summarizing how best to utilize the risk-tiering. We would expect that security teams would use this structure to tailor the risk tiers to the specifics of the local environment.

Risk Tier 3 (High)

1. Data Sensitivity: Involves processing of **sensitive personal data** (PII, PHI), financial data, regulated data (e.g., GDPR, HIPPA, GLBA), or trade secrets.
2. System Exposure: **Externally facing** AI systems accessible by customers, vendors, or the public.
3. Autonomy & Control: AI system has **significant automation or decision authority** without human review ("autonomous AI" or agentic AI).
4. Data Storage: Stores or transmits outputs and inputs across **untrusted networks** or with third-party vendors (e.g., via API).
5. Downstream Access: AI outputs may directly or indirectly trigger automated actions in critical systems (e.g. security operations, financial transactions).
6. Model Customization: Fine tuning or training of models using proprietary enterprise datasets.
7. Compliance Impact: A failure in the AI system may result in regulatory violations or legal liability.

Risk Tier 2 (Medium)

1. Data Sensitivity: Involves **internal business data** that is non-regulated but proprietary or confidential.
2. System Exposure: AI application is **internally facing** but used across departments or business units.
3. Automation Scope: **Human-in-the-loop present,** but AI significantly augments or accelerates internal workflows (e.g., AI co-pilots for code or policy generation).
4. Integration Depth: Integrated into shared productivity platforms or enterprise infrastructure (e.g., Microsoft Copilot).
5. Model Interaction: Uses public foundation models via API, but with safeguards (e.g., no fine-tuning on sensitive data).
6. Data Sharing: Some interactions with third-party vendors or tools (e.g., through SaaS platforms with access to enterprise data).

Risk Tier 1 (Low)

1. Data Sensitivity: AI systems use only **public, anonymized, or synthetic data** with no identifiable attributes.
2. System Exposure: **Internal or sandboxed use** by a restricted set of users (e.g., AI research sandbox).
3. Model Scope: **No fine-tuning or training is required** but this assumes off-the-shelf model uses only.
4. Automation Impact: This assumes **read-only or advisory functions** with no decision-making authority.
5. Business Impact: This assumes that the outputs are not being used to drive operational or financial decisions.
6. Data Flow: Data remains in enterprise-controlled environments with no third-party transmission.

Figure 4-3. Utilizing the Risk-Tiering

FINAL THOUGHT ON RISK MODELS

Keep in mind that generic risk models, such as the one we've presented above, should never be applied blindly to a specific environment. Furthermore, in practice the consequences will have to be connected to some sort of tangible financial value. The FAIR model, for example, is a popular option for this task.

It will be interesting to see whether actual threats, live attacks, and verifiable consequences to real organizations can be identified to validate this risk model, which is admittedly based largely on prediction and conjecture.

We will keep watch to see how this fairs.

CHAPTER 5

MANAGING AI-RELATED IDENTITIES

This chapter provides a high-level introduction to the issues enterprise teams must address to integrate their identity and access management with emerging AI deployment and use.

As enterprises move from experimentation to operational deployment of AI, identity emerges as a foundational security concern. AI systems are no longer limited to isolated analytics or model inference tasks. They increasingly act on behalf of the enterprise, interact with sensitive data, invoke tools, and make decisions that have real operational consequences. This shift forces security teams to reconsider long-standing assumptions about identity, access, and trust.

To understand how identity applies in AI environments, we must first clarify what kinds of entities are participating. Unlike traditional enterprise systems, AI introduces a mix of human users, software workloads, autonomous agents, and fully agentic systems. Each of these requires identity support, but not all require the same level of assurance or privilege control.

A useful starting point is the distinction between passive and active AI systems. Passive systems, such as LLMs responding to prompts, operate in a reactive mode. Active systems, by contrast, are designed to make decisions, initiate actions, and adapt behavior with limited or no human involvement. As autonomy increases, so does the importance of strong authentication, fine-grained authorization, privilege boundaries, and continuous monitoring.

Within this spectrum, it is also important to distinguish between AI agents and agentic AI. An AI agent is scoped to a specific task, such as triaging alerts or enriching data. By contrast, agentic AI systems operate with broader autonomy, chaining tools together, coordinating with other agents, and pursuing goals dynamically. From an identity perspective, agentic AI systems behave less like traditional software and more like independent digital actors.

AI Agent	Agenic AI
Repetitive tasks	Autonomous decisions
Needs instructions	Selects from different options
Predefined objectives	Creates new approaches
Human interactions	Minimal human interactions
Passive or Active	Mostly Active

Figure 5-1. AI Agents versus Agentic AI

However, to avoid getting tangled in distinguishing between (and nitpicking definitions of) Ai agents and agentic AI, we will use "AI agent" broadly throughout this chapter to describe any AI-enabled entity that acts within enterprise systems. Whether passive or autonomous, these agents must be assigned identities, governed by policy, and held accountable through security controls.

EXTENDING IDENTITY SYSTEMS FOR AI

Enterprises have already made substantial investments in identity and access management (IAM), identity governance and administration (IGA), and privileged access management (PAM) platforms. Any realistic AI identity strategy must therefore build on these investments rather than attempt to replace them. Introducing parallel identity systems specifically for AI would increase complexity, fragmentation, and risk.

At the same time, AI agents often replace or augment human roles. They analyze alerts, approve transactions, trigger workflows, and access sensitive resources. Functionally, they behave like members of the workforce, but without the natural guardrails of human intent, judgment, or fatigue. This makes identity extension unavoidable. That is, AI agents must be brought inside the same control framework that governs human and machine access.

This is where non-human identity (NHI) security becomes central. AI agents are not users, but they are also not static applications. They are non-human identities with dynamic behavior, evolving permissions, and variable context. Treating them as first-class identities allows enterprises to apply lifecycle controls, risk assessment, and auditability in a consistent way.

BUILDING AN AI IDENTITY PROGRAM

In today's environments, the most practical approach is to model AI agents similarly to existing non-human workloads such as service accounts or machine identities. This is an advantage, not a limitation, because enterprises already understand many of the risks associated with non-human identities, including credential sprawl, excessive privilege, and lack of visibility.

Modern cloud platforms address these issues through workload identity mechanisms that emphasize short-lived credentials, federation, and elimination of static secrets. Inside service meshes and distributed systems, cryptographic identity frameworks provide mutual authentication and workload attestation. These approaches enable AI agents to participate in zero-trust architectures, inherit enterprise policies, and be monitored like other automated entities.

However, AI agents differ from traditional workloads in one critical way: autonomy. Unlike conventional services that follow predefined call paths, agents discover tools, invoke APIs dynamically, and communicate with other agents. This behavior introduces new attack surfaces and requires more advanced identity threat detection and response capabilities tailored to agent activity rather than static code execution.

As AI agents become more common, identity programs must evolve from simple federation toward interaction-aware security. This includes understanding who an agent is, what it is allowed to do, why it is acting the way it is, and under what conditions its authority should be limited or revoked.

WORKLOAD IDENTITY FOUNDATIONS WITH SPIFFE AND SPIRE

One of the most relevant foundations for securing AI agents comes from the workload identity domain, particularly through the secure production identity framework for everyone (SPIFFE) and SPIFFE runtime environment (SPIRE) frameworks. These technologies were designed to address how to assign strong, cryptographic identities to non-human workloads operating across heterogeneous and dynamic infrastructure.

SPIFFE defines an identity format for workloads, independent of network location, IP address, or underlying platform. Each workload gets a cryptographically verifiable identity that can be authenticated using mutual transport labor security (TLS). SPIRE is the implementation that handles identity issuance, attestation, rotation, and revocation. Together, they provide an identity control plane that works across public cloud, private data center, container, and virtual machine environments.

This model aligns with early-stage AI agent deployments. Like microservices and batch workloads, AI agents are software entities that need to authenticate to APIs, access internal services, and communicate securely with other components. Using SPIFFE/SPIRE allows enterprises to issue short-lived credentials, eliminate static secrets, and enforce mutual authentication without introducing new trust assumptions.

From a zero-trust perspective, SPIFFE/SPIRE provides several advantages. For instance, identity is decoupled from underlying infrastructure, credentials are ephemeral by design, and authentication occurs continuously rather than at login time. These characteristics are well suited for AI agents that may scale dynamically, move across environments, or be instantiated on demand.

However, while SPIFFE/SPIRE is a nice starting point, it doesn't cover all the challenges introduced by AI. Traditional workloads operate within predefined call graphs and static trust relationships. AI agents, by contrast, may autonomously discover tools, chain actions, and collaborate with other agents in ways that were not programmed in advance. This shows that workload identity solves who a software entity is, but nothing about its purpose or authority.

NEW IAM, IGA, AND PAM REQUIREMENTS INTRODUCED BY AI

AI does change the requirements for identity systems. IAM platforms must authenticate agents continuously, not just at session start, and must evaluate access decisions based on changing context rather than static roles. IGA systems must govern agents across their lifecycle, including onboarding, capability changes, and retirement. PAM systems must address privileged actions executed by agents, including delegation from humans and escalation across tools.

These requirements blur traditional identity boundaries. That is, because an AI agent may simultaneously resemble a user, a service account, and an automation framework, identity platforms must converge around capability-based access, policy-driven delegation, and auditable decision trails rather than fixed entitlements alone.

Looking ahead, we expect that emerging identity systems will need to support verifiable credentials and signed capability tokens for AI agents. These mechanisms allow agents to prove what they are allowed to do without exposing unnecessary information, aligning with zero-trust principles in autonomous environments. This will likely be an area in which startups will create capabilities that should be acquired by the bigger IAM and IGA players. Time will tell.

CONTEXT, IDENTITY, AND THE ROLE OF MCP

A recurring concept in AI security involves something that has become known as context. Technically, context includes who initiated a given action, what goal is being pursued through that action, what data sources are involved in the action, and what constraints apply. For humans, much of this context is implicit. For AI agents, it must be explicitly conveyed and enforced.

The model context protocol (MCP) created by Anthropic defines how AI systems communicate in a structured, identity-aware manner. MCP provides a mechanism for agents to declare intent, request access, and receive responses in a way that can be inspected, authenticated, and authorized. When combined with enterprise identity controls, MCP identifies why an agent is calling an API and under whose authority it is doing so.

This use of MCP is a critical shift, because identity becomes more than a credential. Rather, it becomes a carrier of context. Agents can present identity-bound context to security controls, enabling more precise enforcement, better auditability, and safer delegation between humans and machines. Luckily, many vendors are now providing excellent support for practitioners to leverage this protocol in their infrastructure.

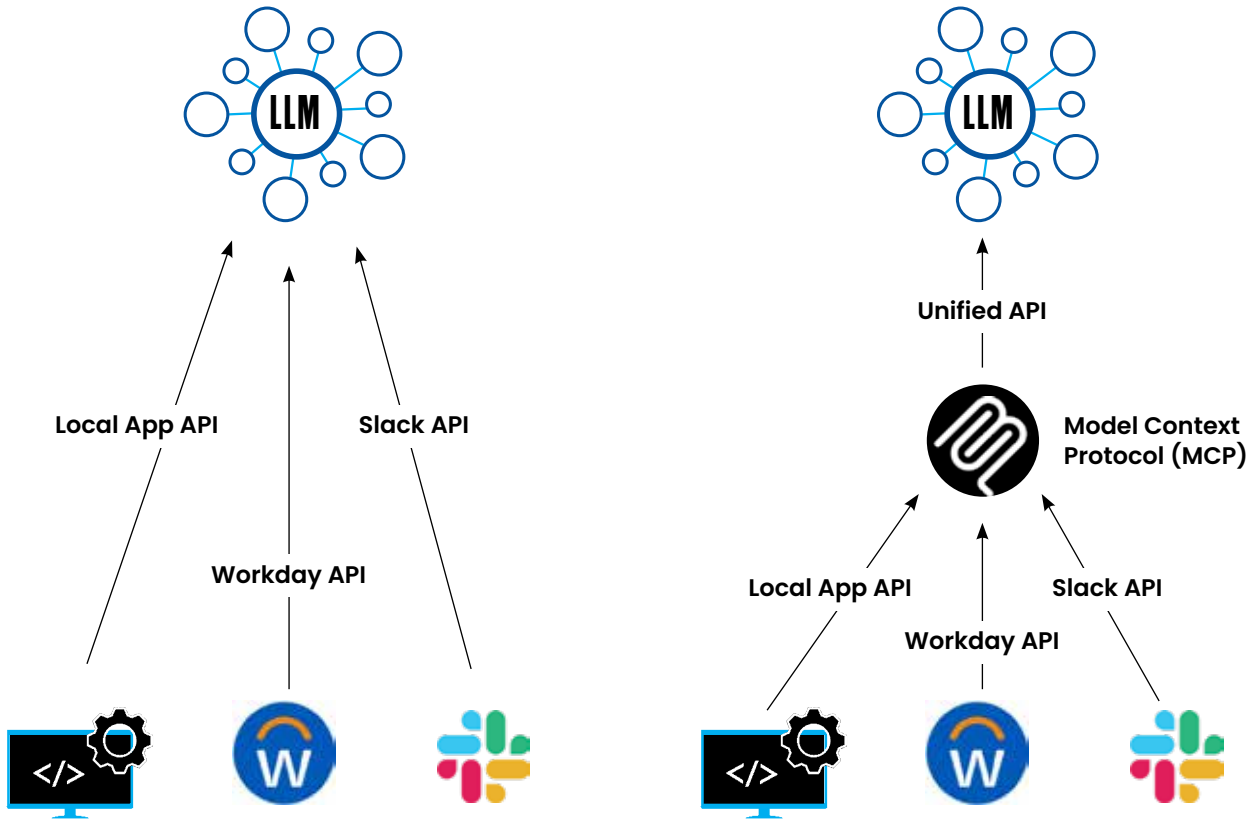


Figure 5-2. Before and After Use of MCP for AI APIs

Teleport AN INTERVIEW WITH EV KONTSEVOY, CEO OF TELEPORT

We spoke with Ev Kontsevoy, CEO of Teleport, to discuss how Infrastructure Identity is becoming the foundation for securing model context protocol (MCP) and agent-driven systems. As enterprises begin deploying agentic AI that relies on MCP and other data sources in production infrastructure for tool use, data access, and system coordination, Teleport's approach highlights why identity — not monitoring or perimeter controls — is emerging as the primary security control for the AI era .

TAG: Why do you see MCP and AI initiatives as an infrastructure identity problem rather than just another API or protocol security challenge?

KONTSEVOY: MCP and agentic AI fundamentally change how systems interact. You are no longer dealing with a single application calling a known service. You are dealing with autonomous and semi-autonomous AI agents that dynamically discover tools, request context, and operate across systems, often without direct human involvement. In that environment, identity fragmentation and secrets sprawl introduce unmanageable risk. The first step, before deploying any agentic workflow, is to establish a unified identity layer, based on strong identity rather than digital credentials. With identity as the basis for trust, you can then control how everything interacts, properly securing these new actors across infrastructure.

TAG: How does Teleport's approach differ from traditional IAM or API security when applied to agentic and MCP-based environments?

KONTSEVOY: Traditional IAM was built around humans logging into systems, and API security was built around static service-to-service calls. MCP sits in between and goes beyond both. AI agents are non-deterministic, continuously operating identities that need short-lived credentials, fine grained access, and strong provenance. Teleport focuses on cryptographic identity and just-in- time access so that every AI interaction is authenticated, authorized, and auditable . That allows organizations to secure dynamic agent behavior without relying on legacy perimeter-based controls or siloed identity systems.

TAG: As AI agents become more autonomous, what role do you see Teleport playing in long- term MCP security strategies?

KONTSEVOY: Our role is to establish the unified identity layer that delivers infrastructure trust. With this foundation, we can then provide zero trust-based access, governance, and identity security capabilities. As agents gain more autonomy, the potential blast radius of a compromised identity grows exponentially . With Infrastructure Identity, that radius is minimized because agents only have access to the resources needed for the task at hand, with authorization that expires.

That creates accountability and control in systems that would otherwise be vulnerable and opaque. In an MCP-driven world, identity must be the foundation of infrastructure trust , and that is exactly where Teleport is focused.

RECOMMENDED STEPS FOR AI AGENT IDENTITY

In the near term, we recommend that enterprises treat every AI agent as a non-human identity with a defined lifecycle. This includes registration, attestation, credential issuance, monitoring, and retirement. Existing workload identity services and cryptographic identity frameworks can provide a solid foundation.

Second, enterprises should expect the emergence of AI access gateways that broker interactions between agents, models, and tools. These gateways will likely enforce identity, inspect prompts, track delegation, and serve as policy enforcement points for AI activity.

Third, when exposing enterprise resources to agents, organizations should standardize on MCP-based interfaces with strong authentication and authorization overlays. Maintaining an internal inventory of agents, permissions, and context sources will become as important as traditional asset management.

Finally, enterprises should plan pilot efforts in 2026 to explore secure agent-to-agent communication, recognizing that standards and tooling remain immature. These pilots will inform future production architectures as identity systems evolve to support autonomous software at scale.

LOOKING AHEAD

Identity for AI is still in its early stages. As of early 2026, standards, protocols, and commercial solutions are evolving rapidly. Security leaders should communicate clearly to executive stakeholders that AI identity is not a solved problem but a developing discipline. The goal is not perfection, but controlled progress grounded in zero-trust principles, strong governance, and continuous learning.

In the next chapter, we examine how commercial vendors are responding to these challenges and where enterprise teams can expect practical support as AI systems move from experimentation to mission-critical deployment.

CHAPTER 6

AI SECURITY VENDORS

This chapter introduces a new AI security taxonomy for enterprise and includes mappings of select commercial AI security vendors to the nine categories included in the taxonomy.

We have now reached the point in our narrative where it makes sense to examine what is commercially available today in the AI security market. As one would expect, the range of companies operating in this space ranges from tiny start-ups with bootstrapped or pre-seed funding all the way to the largest technology firms and their backers who are investing hundreds of millions (or more) to build offerings in this rich and fast-emerging domain.

Others have tried to make sense of this market, including legacy analysts such as Gartner with its TRiSM model. This approach begins with *AI Governance* (and we have comments on the word "governance" below). Gartner supports this with layers called *AI Runtime Inspection* and *Enforcement and Information Governance* (there is that word again). Our observation is that these groupings do not match up well with emerging AI security solutions.

Our goal here is to help enterprise security and AI leaders make better sense of the AI security market by providing a simple tiered taxonomy (see Figure 6-1) for source selection, planning, and alignment. To illustrate the model, we reference actual vendors whose current offerings are good examples of a given category. Such mappings are never perfect, but we hope our attempt is helpful to practitioners.

TAG TWO-TIERED TAXONOMY

The top tier in our taxonomy involves three groupings: Posture, which is focused on determining whether the AI has been properly identified, configured, set up, and readied; Execution, which is focused on ensuring that during runtime the AI is operating safely, securely, and within policy; and Assurance, which is focused on providing evidence that the AI has been selected well and has sufficient levels of integrity, compliance, and trust.

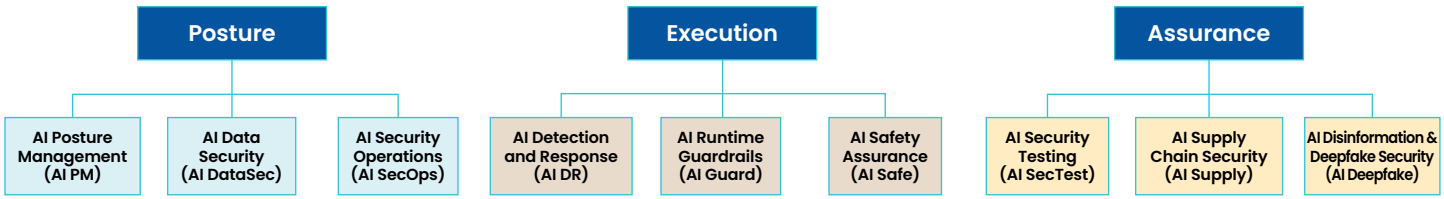


Figure 6-1. Two-Tier Taxonomy of AI Security Focus Areas

In our day-to-day work at TAG, we refer to the taxonomy groupings as focus areas. This wording was selected to leave open some room for local interpretation. Some, for example, will use the taxonomy to categorize vendors. Others, however, will use these focus areas to help define major domains within the company that deserve staffing and budget. Still others might map the focus areas to compliance controls. Readers are invited to share other interpretations.

Several of the enterprise partners who have helped us with this taxonomy are making local modifications to the groupings that support better alignment with their own organizations. This includes their funding models. This is not only a welcome use of the taxonomy but is encouraged. Our goal is not to introduce a rigid structure, but rather to create a baseline tool that helps practitioners with their AI security-related work.

An advantage of working with enterprise partners, of course, is that rather than relying on vendors to self-define their grouping, or to allow the pay-for-play analysts to define categories that support their own sales, we developed something that we believe is more practical and that we provide for anyone to adopt and use with no need of license or permission. We view our work as an "open source" taxonomy of commercial products in cybersecurity.

OBSERVATIONS ON THE AI SECURITY MARKET

Readers are warned, however, that this market will shift quickly. By the time any taxonomy such as this one is published, the commercial landscape will already have changed. We will do our best to future-proof the taxonomy against evolution, but this is an area of considerable flux. Readers should thus resist the temptation to get too comfortable with any single set of labels, marketing descriptors, or vendor category claims.

During our research, for example, we noticed that many vendors use the term governance, perhaps because it is general and avoids being boxed into anything specific. Governance can mean proxies, red teaming, runtime guardrails, configuration posture, compliance—almost anything. When a vendor tells you they do "AI security governance," they are leaving open their options. This is neither bad nor good – but you should be aware.

A second pattern emerged repeatedly during vendor briefings. It involves something that we call the "Swiss Army Knife" approach, in which a given AI security vendor claims to do pretty much everything. This can include red teaming, usage monitoring, security compliance, platform posture, runtime guardrails, user privacy, and more. The danger is not that the aspiration is wrong, it's that the breadth often masks shallow execution.

Granted, there are vendors with platforms that possess the engineering depth to sustain true multicategory excellence. However, "we do it all" can sometimes translate to "we do many things, but none at a level you should rely on." Our advice to the practitioner community is simple: when a vendor claims comprehensive coverage, this can be an excellent opportunity for pre-integrated support, but be sure to do your analysis.

One more thing: The last couple of years have seen aggressive acquisitions of AI security startups by larger vendors, including the cloud hyperscalers. This introduces the sometimes awkward option of either obtaining a function—say, red teaming—–from a tiny startup, or obtaining similar capability from a massive company that has just purchased a tiny startup. The latter might ease procurement, but the former is usually more fun.

AI SECURITY TAXONOMY WITH VENDOR MAPPINGS

In this section, we go through the TAG AI Security Taxonomy in detail, with representative vendor mappings. This approach is intended to help provide a realistic reference architecture for buyers to interpret vendor claims, identify gaps, avoid redundancy, and align investments to risk and compliance needs. The layers reflect both directions of the market: the use of AI for security and the use of security to protect AI.

Of course, the risk in naming actual vendors in this chapter is obvious. Good ones may be left out, some may be misunderstood, and many will be miscategorized to some degree. All that said, we did not want our work to ignore what is being used in the field. Our goal is to offer practitioner value, and practitioners buy products, integrate platforms, and deal with the reality of vendor messaging. So, we needed to include real vendors.

AI SECURITY POSTURE

This first grouping focuses on AI Security Posture. This involves addressing what AI is being used, how it has been set up, whether issues exist in the configuration, whether data is being handled properly, and so on. These objectives are captured in three sub-categories which we refer to as *AI Posture Management (AI PM), AI Data Security (AI DataSec)*, and *AI Security Operations (AI SecOps)*, as shown in Figure 6-2 below.

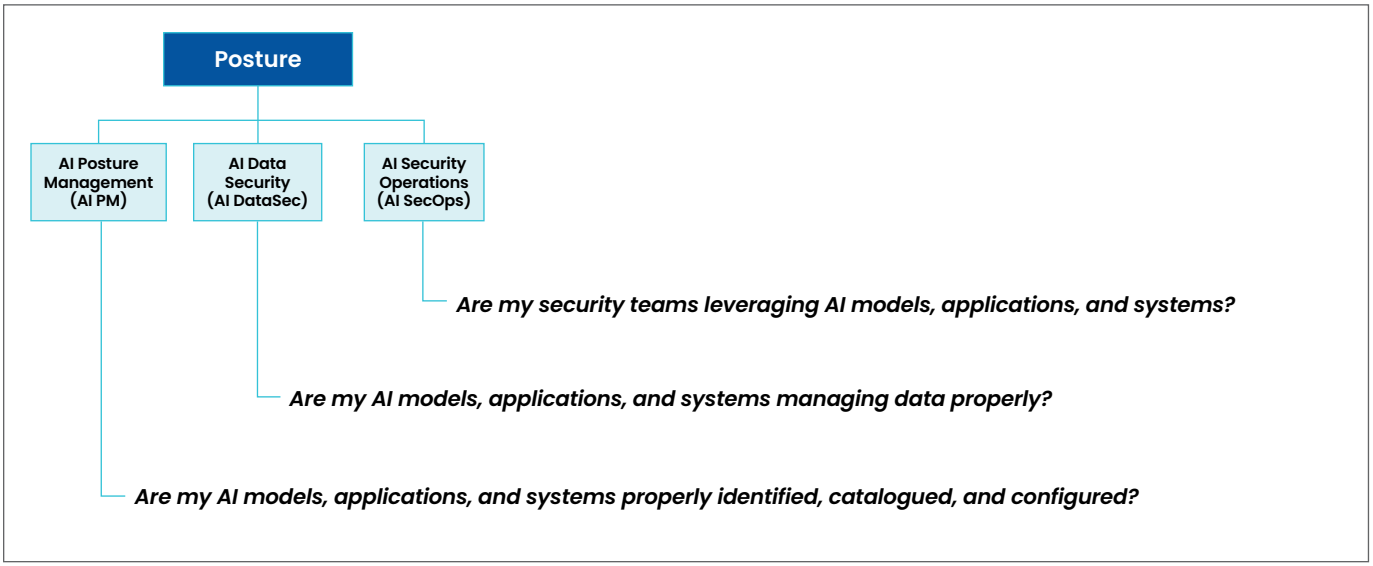


Figure 6-2. AI Security Posture

AI PM

(Are my models, applications, and systems properly identified, catalogued, and configured?)

The objective of AI Posture Management is to help enterprises understand what AI they are actually running, where it resides, and how it has been configured relative to policy, risk tolerance, and regulatory expectations. This category addresses the growing visibility gap created by decentralized AI adoption, where teams experiment with models, agents, plugins, and fine-tuning pipelines outside traditional IT, and security oversight.

Our experience at TAG is that without a clear inventory of AI assets, including models, prompts, vector stores, APIs, and orchestration layers, security teams are unable to reason effectively about exposure, ownership, or accountability. Posture management, in this context, includes identification and discovery tasks, which have often been traditionally separated in many other posture tools such as for cloud or data.

From a practical standpoint, AI PM solutions focus on discovery, configuration assessment, and posture analytics for AI systems in development and production. Capabilities typically include identifying unmanaged or shadow AI usage, evaluating configuration drift against internal standards, and mapping AI components to business units and data domains. It is not uncommon for AI PM to be among the first tasks performed by enterprise AI security teams.

AI PM VENDORS

The commercial vendors we associate with AI PM support a range of capabilities, but all have one thing in common: They are designed to offer visibility and insight into the AI models, applications, and systems in use across the enterprise. Since that is such an important task, often referenced in proof of concept (POC) and proof of value (POV) solicitations, excellent venture capital funding has been available for companies in this area.

Representative vendors in this category that we view as strong include **Cranium**, which is a recent KPMG spin-off startup (highlighted below in an interview with its CEO). We also like the solutions from **Protect AI,** which was acquired by Palo Alto Networks; **Hidden Layer**, which was a recipient of $56 million in Series A funding; **Orca Security,** which includes AIPM in its cloud security arsenal; and **Zenity**, which also offers an AIPM solution.

CRANIUM

AN INTERVIEW WITH JONATHAN DAMBROT, CEO, CRANIUM

We recently sat down with Cranium CEO Jonathan Dambrot to discuss his views on AI discovery, inventory, and configuration assessment. As enterprises struggle to understand which AI models, agents, and pipelines are actually deployed across development and production, Cranium addresses the foundational visibility gap that often undermines downstream security and governance efforts. Here is an excerpt from our discussion:

TAG: Based on your many recent interactions with enterprise security teams, why do you suppose that posture management is emerging as such a critical first step in securing AI?

DAMBROT: Most of the organizations that we have spoken with are further along with AI than many observers would realize. Teams are experimenting with models, agents, fine-tuning pipelines, and third-party APIs outside traditional controls, and

Continued

CRANIUM *Continued*

this creates blind spots very quickly. Without a clear inventory of what AI exists, where it runs, and how it is configured, it's impossible to make informed security or risk decisions. AI posture management gives enterprises that baseline understanding so every other control can be applied rationally.

TAG: How do you see AI posture differing from traditional network, cloud, or application posture management?

DAMBROT: AI introduces a whole new class of asset types that don't map cleanly to existing tools. This includes things like prompts, vector databases, orchestration logic, and model lineage. Configuration risk in AI isn't just about network exposure. It's about how models are trained, what data they reference, and how they're accessed. We designed Cranium specifically to treat AI systems as first-class assets rather than forcing them into cloud or application security abstractions that miss critical context.

TAG: In addition to what you support now, what should enterprises realistically expect from AI posture tooling in the next couple of years?

DAMBROT: Traditionally, security and governance teams have lived in silos—security teams focused on the "how" (technical vulnerabilities) while governance teams focused on the "why" (legal and ethical frameworks).

Today, enterprises should realistically expect AI posture tooling to act as the connective tissue between these two functions. Right now, many tools provide a "snapshot" of a model's risk. In the very near future, we are moving to Runtime Governance. This means security and governance policies will be enforced at the "last millisecond" before a model acts. For example: Imagine an AI agent attempting to access a database it wasn't trained for. The posture tool won't just alert you; it will dynamically revoke the agent's "keys to the kingdom" in real-time, treating it as an autonomous insider threat.

In addition, as we shift more to Agentic AI, posture tools must manage the "Digital Workforce". Within the next 12-24 months, your posture tool will likely be managing more AI identities than human ones. It will enforce "least-privilege" access for every agent, ensuring they operate within defined guardrails and ethical boundaries without human oversight.

AI DATASEC

(Are my models, applications, and systems managing data properly?)

The objective of AI Data Security is to ensure that sensitive data used or exposed by AI systems is properly identified, protected, and governed across training, tuning, inference, and retrieval workflows. AI does introduce new data exposure points such as prompt injection, sensitive leakage, and unintended memorization that complicate conventional controls. AI DataSec addresses these risks by explaining how data flows into and out of AI systems.

In practice, however, this category encompasses traditional data discovery, classification, access enforcement, and leakage prevention tailored to AI usage patterns. This includes understanding which data sources are feeding models, whether regulated or proprietary information is being exposed via prompts or outputs, and how retrieval-augmented generation (RAG) pipelines are governed.

This category goes one step further, however, by including the data security problem that emerges when AI tools such as copilots expose sensitive data to individuals who are over-permissioned. In this sense, we view the AI as a sort of "black light" on the data security posture of an organization. This implies that the best control for dealing with such exposure would be a more traditional DSPM solution.

The result is that this category is a bit more complex than others, because it is probably not going to result in—at least for most organizations—significant investment in new vendors or startups. This is because data security tools are generally in place for the more complex environments, and our view is that these existing tools will usually be sufficient to include coverage of this black light issue.

AI DATASEC VENDORS

Based on our research and discussion with enterprise security teams, as well as our recognition of the black light issue, the vendors we have chosen to line up with AI DataSec include mostly traditional data security companies. This reflects the holistic approach we recommend for handling data-related issues that emerge when AI exposes data to over-permissioned users, or when AI causes data leakage as a result of uncontrolled or insecure processing.

Representative vendors in this category that we view as strong include **Varonis**, which has been a strong leader in data security and has extended this capability for AI-related data risk. We also very much like the solution from **BigID**, which has expanded its coverage from data privacy to include coverage for data security and AI-related data risk. We also like the **Cyera** solution, and we highlight the views of its CEO in the interview below.

CYERA AN INTERVIEW WITH JASON CLARK, CHIEF STRATEGY OFFICER, CYERA

We asked our longtime colleague and industry icon Jason Clark, who runs strategy at Cyera, to chat with us about AI security and why CISOs need to extend their data security strategy to best address AI and agentic architectures . As AI systems increasingly surface sensitive data to users through copilots and RAG pipelines, data exposure risk is increasing. Here is a portion of what Clark shared with us:

TAG: How do you believe that AI changes, or perhaps doesn't change, the data security conversation for enterprises?

CLARK: Our view is that AI acts like a force-multiplier on existing data issues. It doesn't create bad permissions or poor classifications, but it exposes them instantly and at scale. When a copilot or agent can surface sensitive data in seconds, the cost of data hygiene mistakes becomes much higher. That's why AI data security isn't about inventing new controls, but about making sure foundational data protections actually work in AI contexts. Another way to think about it is that AI is exacerbating the DLP problem we've been struggling with forever: Controls and policy fragments are embedded in different parts of an ecosystem, but data itself travels through them all.

TAG: How should enterprise security teams and their CISOs be thinking about AI-specific data leakage risks?

CLARK: They should focus on understanding data flow, not just data location. AI introduces new pathways such as prompts, embeddings, and outputs that

Continued

CYERA *Continued*

traditional DLP often doesn't see, and the new question is one of intent. What's the goal of this agent? Is it accessing resources appropriately? The goal should be to maintain consistent visibility and control as data moves through AI workflows. When data security platforms can observe, classify, and interpret that movement, AI becomes safer without slowing innovation – and in turn, organizations feel more confident in scaling their AI deployments.

TAG: Do you expect AI data security to drive net-new tooling purchases, or will most enterprise teams just leverage their existing data security platforms?

CLARK: I might be a bit biased here, but based on my many years of experience dealing with CISO vendor selections, I'd have to say that in most cases they will not need to make new purchases for AI data security. Enterprises already own data security platforms, and none of this works without understanding your data. The priority should be to first, make sure your data security platform can classify data automatically, with high precision, speed, and scale. If your current platform can't move at the speed of AI, then it won't be a stable foundation for the next step of your journey. The second priority is to extend those investments to cover AI usage rather than creating a parallel stack. Vendors like Cyera that can bridge traditional data security with AI workflows will win because they align with how enterprises actually buy and operate, while providing immediate results and tangible business value.

AI SECOPS

(Are my security teams leveraging AI applications, systems, and models?)

The objective of AI Security Operations is to implement AI security controls within day-to-day security workflows, ensuring that risks are monitored, triaged, and responded to with the best available AI technology for defense. As AI attacks increasingly behave like autonomous actors, SOC processes must accommodate their speed, scale, and ambiguity. AI SecOps seeks to close this gap by embedding AI-specific telemetry and alerts into existing operational systems.

Solutions in this category emphasize alert correlation, playbook execution, and incident response that leverage AI technology. This may include the use of AI to extend or enhance pretty much anything that is being done today in the SOC, including data collection, correlation, and processing. Enterprises should view AI SecOps not as a replacement for existing SOC capabilities, but as a necessary extension.

AI SECOPS VENDORS

The commercial vendors we associate with AI SecOps are different than the vendors included in other categories of the taxonomy, because AI SecOps involves the use of AI for security as opposed to the use of security for AI. In this sense, the vendors here are much more suited to selection and use directly by security teams for their own use in protection, detection, and response, versus being put in place to enable an AI system or application.

Representative vendors in this category that we view as strong include **Qevlar AI,** which includes strong support for AI-based SOC tasks; **Swimlane**, which has done an excellent job leveraging its work in SOAR to develop world-class AI-enablement for the SOC; **Torq**, which provides world-class AI-based SOC enablement; and **Vectra**, which uses AI to support the SOC threat hunter. We highlight the excellent work of **Anomali**, which is an agentic SOC platform, in the interview below.

ANOMALI AN INTERVIEW WITH AHMED RUBAIE, CEO, ANOMALI

We recently spoke with Anomali CEO Ahmed Rubaie about the role of AI in modern security operations and how SOC teams are adapting to increasing speed, volume, and complexity of threats. Anomali delivers an intelligence-native agentic SOC platform, unifying data, contextual threat intelligence, and agentic AI decision systems, to help cybersecurity teams detect, respond, and act at machine speed to stop AI-automated attacks. Below is a portion of our discussion.

TAG: Where is the Macro Economy Headed?

RUBAIE: One of the latest large scale new economic wave adoptions by humanity was the shared gig economy – some thought it was a tech shift and in fact, it was a business model disruption. Next comes AI native economics (the efficient/productive economy) – once again, not a tech shift but rather a business model disruption. AI is fundamentally reshaping how revenue is created, how value is perceived, and what companies sell – a more horizontal transformation across every angle of the economy than anything humanity has seen yet. In tandem, it is creating the fear of job losses.

TAG: How Should Security Vendors Evolve?

RUBAIE: Security vendors have two areas to evolve – the interoperability of data accessibility (visibility to Big Data), context (intelligence) and orchestration of agentic autonomous decision making. The agentic layer sits on top of the data and context layer – this composition can be in conjunction with the Big Data players or in place of where the Security vendor evolves into "all things" enterprise. Whichever way, the result is the same for enterprise customers – an agentic SOC is part of an agentic enterprise framework – deliver better outcomes with new levels of efficiency and productivity.

TAG: How should Enterprise Customers Rethink Risk Management?

RUBAIE: Enterprise customers should revisit Risk Management – this is too big of a business model transformation to leave to legacy thinking. Does the recent move to single platforms give rise to significantly more risk in the age of AI? Big Data visibility, context and interoperability with agentic orchestration of AI will likely split from the single platform concept for prudent risk management. As noted above, the dust will settle on an enterprise brain with security DNA.

AI SECURITY EXECUTION

This second TAG top-tier taxonomy grouping focuses on AI Security Execution. This involves addressing whether AI runtime indicators or attacks are being detected and responded to, and whether guardrails are in place to avoid attacks. These objectives are captured in three sub-categories which we refer to as *AI Detection and Response (AI DR)*, *AI Runtime Guardrails (AI Guard)*, and *AI Safety Assurance (AI Safe)* as shown in Figure 6-3 below.

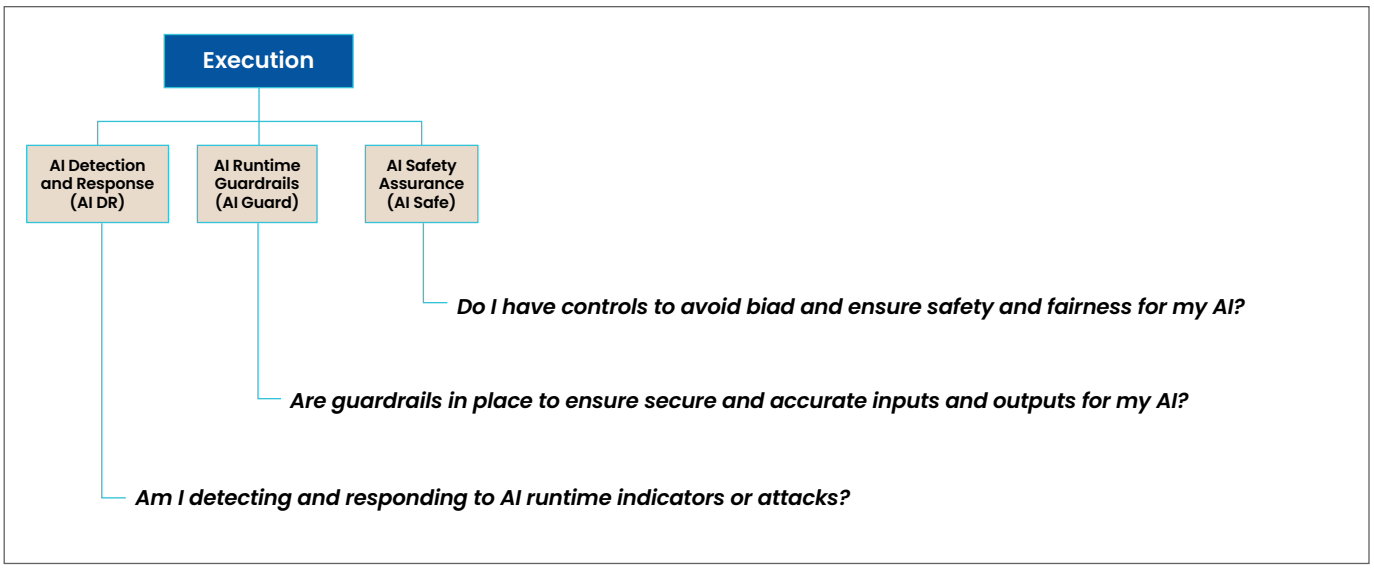


Figure 6-3. AI Security Execution

AI DR

(Am I detecting and responding to AI runtime indicators or attacks?)

The objective of AI Detection and Response is to identify active threats targeting AI systems and to support mitigation during runtime. As adversaries probe models through adversarial prompts, data poisoning, and abuse of exposed APIs, AI DR focuses on recognizing malicious behavior as it occurs rather than relying on pre-deployment controls. This mirrors the evolution from static signatures toward more dynamic behavioral approaches in cyber defense.

In fact, AI DR capabilities do often include behavioral analysis of prompts and responses, detection of anomalous model usage, and correlation of AI-specific indicators with broader security signals. Some solutions extend to automated or semi-automated response, such as throttling access, blocking abusive sessions, or triggering human review. For buyers, AI DR is most relevant where AI systems are externally accessible or business-critical.

It is worth mentioning, however, that some buyers might be forgiven for feeling weary of the endless detection and response (DR) capabilities as evidenced by MDR, NDR, EDR, ADR, CDR, and even the wildcard XDR. It is thus reasonable to demand clarity around why there needs to be a new set of vendors for AI DR when so many detection and response options already exist, including those that may already be deployed into the local environment.

AI DR VENDORS

As suggested above, the commercial vendors we associate with AI DR should be held to define clearly how they differentiate from existing detection and response vendors. We do, in fact, believe that some

vendors make this case reasonably well, pointing to new attack vectors from AI that require a unique response. But the specter of existing XDR or MDR vendors covering this use case, perhaps even through an AI DR startup acquisition, seems strong.

Representative vendors in this category that we view as strong include **Cisco**, (which we highlight in the interview below) Cisco recently acquired **Robust Intelligence**; **CalypsoAI**, which was recently acquired by F5; **Fortinet**, which is a leading security vendor for pre-integrated solutions and includes AI detection and response; and, of course, **Microsoft**, which can pretty much fit their range of solutions into every category.

AN INTERVIEW WITH DJ SAMPATH, VP AI SOFTWARE AND PLATFORM, CISCO

We spoke with our colleague DJ Sampath, VP AI Software and Platform, about AI security following Cisco's acquisition of Robust Intelligence. We aimed to better understand how AI detection and response fits into an already mature enterprise security portfolio. Cisco's position offers an interesting lens on how AI-specific threats are integrated into broader detection and response strategies rather than treated as a standalone problem.

TAG: From a leadership perspective, should AI security be treated as a fundamentally new risk category—or as an extension of existing cybersecurity programs?

SAMPATH: AI brings a new class of risk because it is fundamentally different from traditional software. Traditional cyber-attacks exploit technical weaknesses and are typically addressed by patching these vulnerabilities.

On the other hand, AI risk is distributed across the various components of a complete AI system—compromised models, poisoned datasets, malicious agents and tools. In fact, AI failures can often arise from misuse of intended functionality. A system can behave exactly as designed and still cause harm.

This distinction matters at the executive level. AI risk is new and complex, but the core concepts of existing cybersecurity programs still apply. The right approach is not to create a parallel security program, but to extend existing ones with tools and strategies purpose-built for AI.

TAG: Many CISOs worry that AI detection and response could become yet another silo. Since your team designed AI Defense natively inside Cisco's security portfolio, what architectural decisions did you make to prevent that outcome?

SAMPATH: We started with the assumption that security teams don't want another console. They want better signal fidelity inside the platforms they already trust. So we designed AI Defense to integrate at two levels.

First, it aligns with existing SOC workflows—alerts, severity scoring, and investigations look familiar to analysts. Second, it treats AI risk as correlated risk. A suspicious model interaction might not matter on its own, but if it coincides with abnormal API usage or compromised credentials, the combined signal is powerful.

The goal is not to create an "AI SOC," but to make AI behavior observable and actionable inside the same operational fabric that already protects the enterprise.

Continued

Continued

TAG: Looking ahead, how do you see AI detection and response evolving over the next few years, especially as autonomous agents and model-to-model interactions become more common?

SAMPATH: We're moving from securing individual models to securing AI systems which are increasingly capable and complex. As agents make decisions, call tools, and interact with other models, the attack surface becomes systemic.

The evolution of AI detection and response will grow to include greater context awareness and more adaptive controls. This means greater focus on intent drift—whether an AI system is behaving consistently with its intended purpose. Response will also become more dynamic; rather than simply blocking requests, systems may restrict capabilities, retrain models, or isolate specific components or interactions within the AI workflow.

Ultimately, AI security will look less like signature-based defense and more like continuous governance: understanding what your AI is allowed to do, what it is actually doing, and intervening when those diverge.

AI GUARD

(Are guardrails in place to ensure secure and accurate inputs and outputs for my AI?)

The objective of AI Runtime Guardrails is to constrain AI behavior during operation so that models and agents act within defined safety, security, and policy boundaries. Unlike detection-centric approaches, guardrails are designed to prevent undesirable outcomes such as data leakage, policy violations, or unsafe responses before they reach users or downstream systems. This category implies that not all AI risk can be mitigated through training or testing alone.

In practice, AI Guard solutions implement controls such as prompt filtering, output moderation, policy enforcement, and contextual access checks at inference time. These guardrails may be static or adaptive, and may vary by user role, application, or risk profile. Enterprises adopting AI at scale increasingly rely on runtime guardrails as a compensating control, particularly when using third-party or foundation models that cannot be fully customized or audited internally.

AI GUARD VENDORS

As suggested above, the commercial vendors we associate with AI Guard provide runtime support, which is quite different than static posture assessments. When we see vendors who do both static posture and runtime guardrail support, we always seek to learn more about what's going on behind the scenes. These are different types of controls, and they demand different types of operational support and functionality. Keep this in mind during source selection.

Some vendors that we view as strong include **Witness AI,** which includes guardrails in its portfolio; **Alice**, which also includes guardrails for AI; **AWS**, which is one of the big hyperscalers that supports this type of security for AI workloads via its Bedrock capability; and, of course, **Google**, which has been a leader in all aspects of AI security for many years now. **Noma Security**, which we highlight below, also includes guardrails in its end-to-end offering.

NOMA AN INTERVIEW WITH NIV BRAUN, CEO, NOMA SECURITY

We sat down with Noma Security CEO Niv Braun to discuss the emerging need for runtime guardrails that constrain AI behavior during operation. What we learned was that Noma's platform serves as an excellent example of not only runtime support for AI, but as an end-to-end lifecycle solution that offers a fully integrated AI security experience for enterprise teams. Here are some key points we learned from the discussion.

TAG: What are the top priorities you see currently for enterprise security teams when it comes to AI security controls?

BRAUN: The top priority is to adopt a comprehensive AI Security Framework that provides the enterprise the ability to implement in a practical way the security and governance policies around AI, and this way to enable wider adoption.

This framework starts with discovery and full visibility to all the different AI assets in the organization - models, agents, datasets, MCP and more. Continues with proactive risk management, checking the agent blast radius control and posture management, ensure safe use of AI supply chain, and red team for the AI application and agents. And completed with real time monitoring of all the AI and agent communication, to enforce the enterprise policies in prompt, responses, and agent communication.

Another priority is that enterprise want to make sure that this framework covers all types of AI and agents in the enterprise - from homegrown AI built by engineers (cloud, repositories, AI platforms like Databricks) to SaaS agent platforms (e.g. Copilot Studio, Salesforce Agentforce) and up to local agents (e.g. Cursor, Claude code) and all the MCP spread around the enterprise.

TAG: How do you respond to concerns that posture checking and runtime guardrails might limit innovation?

BRAUN:The goal of an AI Security Platform is to enable innovation in the enterprise and accelerate adoption, not to restrict it. It provides the trust for wider adoption, quicker deployment, the confidence to embed AI and agents in more sensitive business cases, and saves time from self-building of controls. If the AI security controls in the enterprise slow it down, it most times means that they were not implemented properly. Main emphasis for effective adoption of an AI Security Platform that boosts the AI transformation:

First, granular, configurable and adaptive for the enterprise business cases - each enterprise has its needs, and different areas inside the enterprise have different use cases. To be able to support these use cases the AI security controls, from posture to guardrails, should be granular and easily configurable to adjust to each use case.

Second integrate into existing operations - while AI security requires new tools, the most efficient implementation would be to embed the operations as part of the existing security operations. That means to use AI Security Platform that integrates into the existing vulnerability management processes, to operations such as DevSecOps, and the existing SOC processes and tools.

Continued

NOMA *Continued*

Third, contextualization - to avoid alert fatigue of new tools contextualization is critical. To have not only posture management, testing (red team), and AI Detection and Response (guardrails), but to use the context of each of these tools to feed the others.

Once these principles are being implemented, an AI Security Platform should only boost AI innovation and not limit it.

TAG: What types of AI security innovation and evolution do you expect to see from the marketplace, or from practitioners, over the next couple of years?

BRAUN: We are excited to see that our end-to-end AI Security Platform has become a standard approach, and the next phase is the contextualization of the tools. In addition, we expect to see agentic behavior detection becoming an important part of the AI Detection and Response - Detecting anomalies of agents that point to risks of malicious activities as well as rogue agents which expose the enterprise to risks unintentionally.

AI SAFE

(Do I have controls to avoid bias and ensure safety and fairness for my AI?)

The objective of AI Safety Assurance is to provide confidence that AI systems behave in ways that are consistent with organizational values, ethical principles, and external expectations. While often conflated with governance or compliance, this category focuses on validating that AI outputs and behaviors do not introduce unacceptable harm, bias, or misuse. AI Safe reflects the growing convergence of technical assurance with broader trust considerations.

Capabilities include bias detection, harmful output analysis, robustness evaluation, and documentation of safety-related testing and controls. These functions are especially relevant for enterprises in regulated sectors, where AI behavior is subject to heightened scrutiny. Buyers should recognize that AI Safety Assurance is not a one-time exercise, but an ongoing validation effort that must evolve alongside models, data, and use cases.

AI SAFE VENDORS

As suggested above, the commercial vendors we associate with AI Safe must keep in mind that the large technology companies developing the models being used for AI applications and systems are well-positioned to implement safety and ethics-based functions directly into their models. It will be interesting to see if this becomes an embedded feature, either natively developed by OpenAI, Anthropic, and others, or installed via startup acquisition.

Representative vendors in this category that we view as strong include **Holistic AI,** which supports this area of AI security (and which we highlight in our interview below). In addition, we like **FairNow**, recently acquired by **Optro**. We also like the solutions from **OneTrust** and **Vanta**, which will allow existing customers to obtain this type of support without having to introduce a new vendor.

Holistic AI

AN INTERVIEW WITH ADRIANO KOSHIYAMA, CO-CEO, HOLISTIC AI

We spoke with Adriano Koshiyama, Co-CEO of Holistic AI, about why AI safety assurance is becoming a core enterprise security responsibility as AI systems increasingly influence decisions with real-world consequences. Rather than treating AI safety as a documentation or compliance exercise, Holistic AI focuses on providing continuous validation of fairness, robustness, and trusted behavior across the AI lifecycle. Here is a portion of the discussion.

TAG: Based on your experience dealing with enterprise security teams, why do you think that AI safety moved beyond a purely technical discussion?

KOSHIYAMA: AI systems are no longer confined to experimental labs or isolated analytics use cases. They are making or influencing decisions that affect people directly, such as credit approvals, hiring recommendations, fraud detection, medical prioritization, and more. When those systems exhibit bias, instability, or unsafe behavior, the consequences extend beyond model performance metrics. They become safety, legal, reputational, regulatory, and trust issues. That reality pulls AI safety firmly into enterprise risk management, where CISOs, risk officers, and compliance teams all have a stake.

TAG: How does this shift in focus of AI safety and trust change the role of security teams specifically?

KOSHIYAMA: Our observation is that security teams have been accustomed to thinking in terms of assurance. That is, they have had to focus their efforts on proving that systems behave as expected under real operating conditions. As AI becomes embedded in critical workflows, those same teams are being asked new questions. For example, executives might ask whether a given model is operating fairly across populations. Or they might ask whether the security team can detect harmful drift or unintended behavior before leads to a negative impact. Holistic AI gives these security and risk teams the familiar type of technical evidence that they need to answer these types of questions with confidence at scale, rather than relying on self-attestations or one-off static model reviews.

TAG: How do you distinguish AI safety assurance from AI security governance, or would you say that they are the same?

KOSHIYAMA: No, there are certainly differences. AI governance focuses on issues such as intent. For example, it captures policies, principles, risk tolerance, and regulatory obligations, with emphasis on what an organization wants its AI systems to do and not do. Assurance, in contrast, is much more about proof. It provides measurable, testable evidence that deployed AI systems actually behave in line with those expectations, continuously and at scale. Without assurance, governance remains aspirational. Our Holistic AI platform operationalizes that gap by translating high-level governance requirements into concrete tests, metrics, and monitoring that can be enforced across models and use cases.

AI SECURITY ASSURANCE

This third grouping focuses on AI Security Assurance. This involves addressing whether the AI is being subjected to security testing and red teaming, whether information exists regarding the supply chain for the AI, and so on. These objectives are captured in three sub-categories which we refer to as *AI Security Testing (AI SecTest), AI Supply Chain Security (AI Supply)*, and *AI Deepfake Support (AI Deepfake)* as shown in Figure 6-4 below.

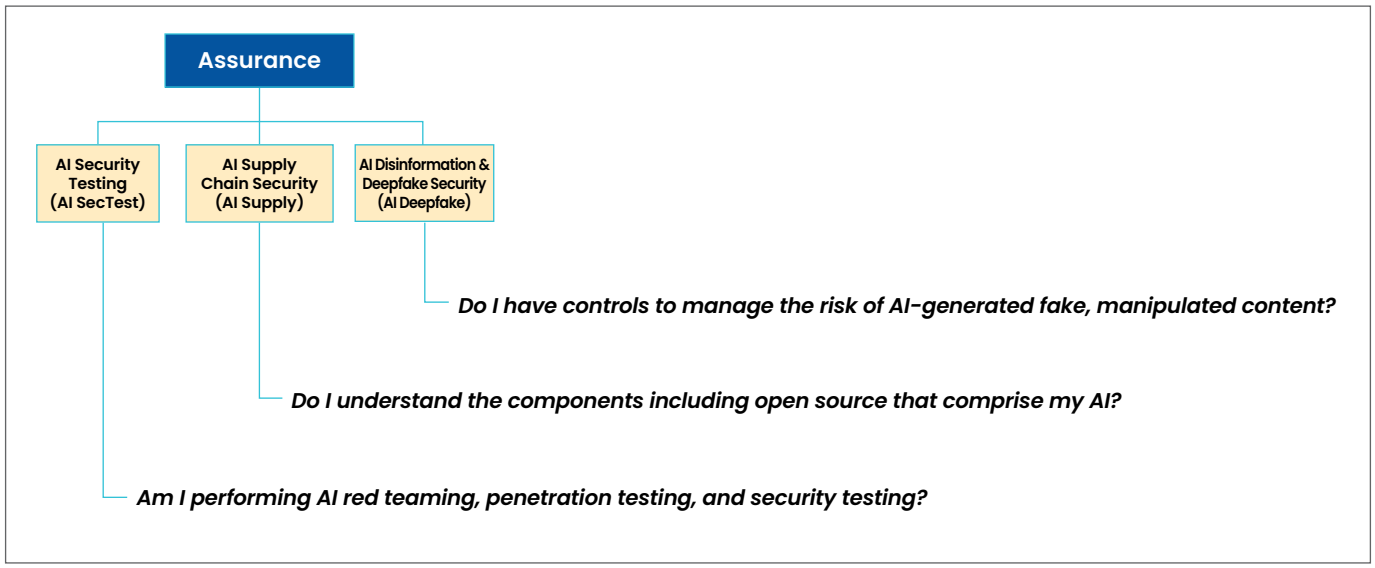


Figure 6-4. AI Security Assurance

AI SECTEST

(Am I performing AI red taming, penetration testing, and security testing?)

The objective of AI Security Testing is to evaluate AI systems for vulnerabilities prior to, and increasingly during, production deployment. This includes identifying weaknesses such as prompt injection susceptibility, training data leakage, model inversion risks, and unsafe emergent behaviors. Much like application security testing in earlier eras, AI SecTest aims to surface flaws before adversaries do.

Solutions in this category often draw from red teaming, adversarial testing, and simulation techniques adapted for AI systems. These tests may be manual, automated, or hybrid, and may target models, agents, pipelines, or integrated applications. For practitioners, AI SecTest provides a critical feedback loop that informs posture, execution, and assurance decisions, and helps prioritize remediation efforts based on observed weaknesses.

AI SECTEST VENDORS

The commercial vendors we associate with AI SecTest usually point to their libraries of attacks as a major value proposition. It will be interesting to see how this advantage stays (or goes) as organizations such as MITRE drive a common view of signatures to be tested here. And it should go without saying that AI testing must move from signatures to behavioral—and then, eventually, to AI testing AI. Stay tuned.

Representative vendors in this category that we view as strong include **Splx**, which was recently acquired by Zscaler, and **Mindguard** (which we highlight in the interview below). **Mend.io** and **Group-IB** are also well-positioned to provide support for buyers in this area. We do feel obliged to emphasize that enterprise teams can use existing relationships with technology firms such as **Trail of Bits**, **Pentera**, or **Bishop Fox** to obtain AI penetration testing and red teaming.

MINDGARD AN INTERVIEW WITH JAMES BREAR, CEO, MINDGARD

We spoke with Mindgard CEO James Brear about why securing AI systems requires a different, and often expanded, mindset compared to traditional application security testing and red teaming. Mindgard specializes in attacker-aligned AI security testing that focuses on system behavior, abuse scenarios, and emergent risk across models, agents, and surrounding infrastructure. In this excerpt, Brear explains why many established AppSec assumptions break down in AI environments and how enterprises should rethink ownership and cadence for AI security testing.

TAG: Why doesn't traditional application security testing translate directly and cleanly to AI systems?

BREAR: Traditional AppSec assumes determinism. If a piece of code receives a given input, it should produce the same output, and vulnerabilities tend to live in predictable places such as input validation, authentication logic, memory handling, or configuration errors. AI systems behave fundamentally differently . The same prompt or input can yield different outputs depending on context, prior interactions, model state, or even subtle changes in phrasing.

That non-determinism undermines static analysis, signature-based scanning, and even conventional dynamic testing. From a security perspective, the risk isn't just whether the system can be broken, but how it behaves under misuse, manipulation, or adversarial pressure. Testing for these behaviors must go beyond known vulnerability classes to simulate adversarial intent .

TAG: How should enterprise teams think about the frequency and cadence of AI security testing in the context of their overall security program?

BREAR: AI security testing has to be continuous by design. Models change frequently through retraining, fine-tuning, prompt updates, or exposure to new data sources – and each change can introduce new risks. Even when the underlying model remains the same, user behavior and usage patterns evolve, which can surface entirely new misuse scenarios.

Treating AI testing as a one-time pre-production activity creates a false sense of security. The better approach is scenario-based, ongoing testing that reflects real-world usage of AI systems and evolving attacker techniques . This allows organizations to measure risk continuously rather than assume it remains static after deployment.

TAG: Who should own AI security testing inside the enterprise, and this includes managing budget and selecting vendors?

BREAR: No single team can own AI security end-to-end, and that's one of the biggest organizational challenges we see today. Security teams bring adversarial thinking and understand how attackers probe systems for weakness. Engineering and data science teams understand how models are built, tuned, and deployed, and what trade-offs exist between performance and control. Risk, legal, and compliance teams frame the business, regulatory, and reputational impact if something goes wrong. Effective AI security testing sits at the intersection of the se groups. It requires shared ownership, common tooling, and a shared understanding of risk.

AI SUPPLY

(Do I understand the components including open source that comprise my AI?)

The objective of AI Supply Chain Security is to address risks introduced by the complex and often pretty opaque ecosystem of components that comprise modern AI systems. These components include pre-trained models, open-source libraries, datasets, APIs, plugins, and third-party services, often sourced from multiple vendors with varying levels of transparency. AI Supply recognizes that trust in AI cannot exceed trust in its dependencies.

Capabilities in this category involve provenance tracking, integrity verification, dependency analysis, and risk assessment across the AI lifecycle. This may include understanding where models originated, how datasets were curated, and whether components have been tampered with or deprecated. Enterprises with mature third-party risk programs should view AI Supply as a natural extension of those efforts.

Unfortunately, we also must address here the uncomfortable supply chain issue of U.S. versus Chinese AI models. It would be our wish and dream for the world that these two nations would cooperate on AI, but it doesn't seem like that is happening. As such, many supply chain efforts are geared toward providing assurance that AI from the U.S. is not beaconing back to China for support or processing. Such a shame.

AI SUPPLY VENDORS

As suggested above, the commercial vendors we associate with AI Supply will need to demonstrate sufficient value for AI specifically to warrant purchase by enterprise teams already in possession of tools that do provenance tracking, including creation of bill of materials. The guidance we always provide at TAG is to leverage what you've already bought, so this could be an uphill battle for some of the weaker vendors.

Representative vendors in this category that we view as strong include **Legit Security**, which empowers developers to safely adopt AI; **Endor Labs**; **CrowdStrike**, which acquired start-up **Protect AI**; and **Reversing Labs** (which we feature in our interview following), which we have always liked in terms of software supply chain risk management, and which extends their support to AI models and apps.

REVERSINGLABS

AN INTERVIEW WITH MARIO VUKSAN, CEO, REVERSINGLABS

We recently had the great pleasure to speak with Mario Vuksan, Chief Executive Officer of ReversingLabs, about how the rapid adoption of AI and software automation is changing the nature of software trust. As enterprises increasingly rely on compiled software, container images, firmware, and AI-enabled components sourced from third parties, traditional security controls are often blind to what is actually being delivered. ReversingLabs focuses on deep inspection and analysis of software artifacts to help organizations understand whether the software they deploy is safe, authentic, and free from hidden malicious behavior.

TAG: Many organizations talk about software supply chain risk today. How has that risk changed as AI and automation become more embedded in enterprise systems?

VUKSAN: The impacts on end user organizations of large language models and generative AI are huge. Software vendors and organizations are embracing the development and use of AI powered applications at an unprecedented level, creating new transparency and risk challenges. Both sides struggle to adequately understand the relationship between risk and liability.

Continued

REVERSINGLABS *Continued*

On the development end, there's also unprecedented growth in the code base. AI-driven coding increases the speed, scale, and the volume of software changes frequently trusted without being fully inspected. This also applies to the development of AI applications themselves. For example, ReversingLabs researchers identified two Hugging Face models containing malicious code that evaded detection by HuggingFace's "Picklescan" security tool. Separately, we discovered malicious packages targeting users of Alibaba AI Labs with Infostealer malware impersonating a Python SDK that interacts with Aliyun AI Labs services.

And then there are the threats posed by AI-powered offensive cyber operations. Those include highly targeted phishing campaigns, bots, and more recently fully AI-generated malicious software.

TAG: What do you see as the biggest gap in how enterprises currently manage software supply chain security?

VUKSAN: The biggest gaps are in both the assumption that reputation or vendor trust alone is sufficient. Many organizations implicitly trust software because it comes from a known supplier, a popular repository, or is delivered as a signed package. But trust should be continuously verified, not assumed. Malicious code does not always come from unknown sources. We routinely see legitimate software packages that have been tampered with, Trojanized, or subtly modified. Without the ability to analyze the actual binary or artifact itself, organizations cannot confidently determine what software will do once deployed, and whether it should be accepted in the first place, regardless of the reputation of its authors.

TAG: How should enterprise teams think differently about establishing trust in software?

VUKSAN: Trust must be based on transparency and evidence. Organizations need software to pass inspection at a deep technical level to be considered fit for consumption, like inspecting the food we buy in our grocery stores. We may want to review software's "ingredients" - a software bill of materials (SBOM) - and analyze binaries, containers, installers, and firmware to detect hidden malware, suspicious capabilities, or unexpected changes between versions.

Establishing trust in software also means establishing long-term confidence in software vendors' commitment to quality and safe software. This approach shifts supply chain security from a policy exercise ("Please complete this questionnaire") to a disciplined verification grounded in technical truth and evidence-based monitoring software vendor's commitment to continuing partnership.

TAG: What misconception do you most often encounter when discussing this topic with security leaders?

Vuksan: Over the last four decades, our infosec industry has been built around the concept of securing things: desktops, laptops, servers, embedded devices - not to mention the networks they inhabit and the applications they run. Software supply chains, on the other hand, were assumed to be trustworthy.

Surprisingly, in 2026, that is often still the case. Malware detection is only relevant at the endpoint, and falls to EDR solutions. Network security is about access control and network monitoring, but fails to verify the integrity of the software infrastructure it is deployed on. And so on. In reality, the game has shifted substantially. Supply chain attacks have gone from "unicorn" type events conducted by top-tier state actors to commonplace occurrences carried out by

Continued

REVERSINGLABS *Continued*

cybercriminal groups -and even individual actors. They occur long before software ever reaches an endpoint, with malicious functionality introduced during build, packaging, or distribution, cloaked in the reputation of the software package they came with.

Security leaders and the industry are finally being forced to accept that software supply chains are not trustworthy, and that it is their responsibility to ensure that software must be assessed and evaluated before deployment, not after a compromise. This will benefit not only end user organizations, but also the entire interconnected community. The earlier trust is established, the more effective risk reduction becomes.

TAG: Looking ahead, how do you expect software supply chain security to evolve?

VUKSAN: We believe organizations will embrace the concept of software transparency and continuous software verification in the coming months, with every piece of software carefully evaluated throughout its life cycle. And, as AI-driven development increases the velocity of code creation, automated and scalable inspection - from vulnerability scanning to complex binary analysis - will become essential. Ultimately, enterprises will learn to treat the software artifacts pushed to their IT environments the same way they treat identity- or financial transactions: with nothing trusted implicitly, and everything verified based on observable evidence.

AI DISINFORMATION AND DEEPFAKE SECURITY

(Do I have controls to manage the risk of AI-generated fake, manipulated content?)

The objective of AI Deepfake Support is to help organizations detect, analyze, and respond to synthetic media used for fraud, impersonation, or disinformation. As generative models make it increasingly easy to fabricate convincing audio, video, and imagery, enterprises face new threats to trust, brand integrity, and decision-making processes. This category addresses both external threats and internal misuse scenarios.

Solutions in this area typically focus on detection of manipulated content, validation of authenticity, and integration with investigative or response workflows. Use cases range from executive impersonation fraud to manipulated evidence and social engineering campaigns. While not all enterprises will prioritize this category, AI Deepfake Support is becoming relevant for organizations with high public visibility or real reputation risk.

AI DISINFORMATION AND DEEPFAKE SECURITY VENDORS

The commercial vendors we associate with AI Deepfake are not always included in AI security taxonomy groupings, but we believe they should be here. The creation of deepfakes and misinformation is a direct consequence of the existence of AI tools, so their prevention and detection are also best done with tools that leverage AI. We expect to see this become a more commonly budgeted item

Representative vendors in this category that we view as strong include **Alethea**, which provides excellent support in the detection and mitigation of disinformation and misinformation, **GetReal Security**, which is an industry leader in the accurate and consistent detection of deepfake images and video, and **Blackbird AI**, which detects manipulated information (and is highlighted in the interview below).

BLACKBIRD.AI AN INTERVIEW WITH WASIM KHALED , CEO AND CO-FOUNDER, BLACKBIRD.AI

We spoke with Blackbird.AI CEO and Co-Founder Wasim Khaled about the accelerating threat posed by AI-generated Narrative Attacks created by disinformation, impersonation, and synthetic media. As AI-based technology becomes cheaper, faster, and more accessible to threat actors, cybercriminals and nation states, organizations are confronting a new class of security risk that directly targets trust, reputation, and executive decision-making, often outside the boundaries of traditional cybersecurity controls.

TAG: Why should security leaders care about narrative attacks today?

KHALED: Because AI-based narrative attacks are no longer experimental or niche. Instead, they are actively being weaponized. We're already seeing synthetic audio, video, and text used to impersonate executives, manipulate employees, mislead customers, and influence public perception. Unlike traditional fraud techniques, narrative attacks exploit human trust directly, bypassing many technical controls by tapping psychology, authority, and urgency. For security leaders, this represents a shift in the threat model. Narrative attacks enable fraud and manipulation at scale, often with minimal cost to the attacker and very high potential impact. A single convincing impersonation can trigger financial loss, reputational damage, regulatory scrutiny, or even physical safety concerns for executives.

TAG: How does AI change the defensive side of this problem?

KHALED: The scale and speed of synthetic content generation make manual detection unrealistic. Human review simply cannot keep pace with the volume, variety, and sophistication of AI-generated media. Effective defense requires actionable AI-driven Narrative Intelligence detection and analysis that can identify subtle signals of manipulation, such as patterns in language, audio artifacts, visual inconsistencies, and contextual anomalies that humans often miss. But detection alone isn't enough. Defensive AI also has to provide context, including what harmful narratives are targeting an executive or organization, how the content is spreading across networks, is it bot influenced, and what the likely intent is. If you have that context, you can make informed strategic decisions to significantly reduce risk

TAG: Who typically owns this risk inside organizations?

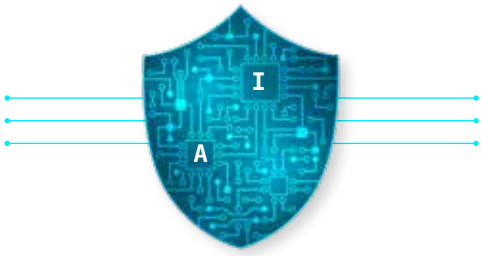
KHALED: More and more, CISOs and the security operations center is taking the lead as this is considered an attack on the company, especially with recent physical attacks on executives. Security teams understand threats and incident response and when they monitor a new threat, they share it with important groups within the company including the communications and brand teams who understand reputation and public trust andlegal teams who assess liability and regulatory exposure. Executives themselves are often the direct targets. Narrative atatcks cut across all of these domains simultaneously. Effective organizations recognize that Narrative Intelligence is critical and requires shared ownership and coordinated response. That means aligning security, communications, legal, and leadership around common detection signals, escalation paths, and response playbooks in what many are now calling a Fusion Center. When those groups operate in silos, attackers exploit the gaps.

ADDITIONAL AREAS OF AI SECURITY SUPPORT

Our taxonomy, as we have acknowledged, is not perfect, so coverage gaps between the various categories and subcategories are inevitable. Some of the more prominent areas in which emphasis exists in the marketplace (meaning that vendors have been funded, or inbound requests come in from enterprise teams from time to time) but for which support in our taxonomy might be somewhat lacking, include the following:

- **Third-Party Risk Management for AI Usage:** We like the idea of using specific controls to deal with third-party Ai usage risk. Vendors like **PromptArmor** appear to be doing good work in this area.
- **Compliance Support for AI Usage:** Most teams already have implemented compliance automation and should be able to extend this for AI models and apps, but vendors such as **Scytale** offer a more AI-focused approach for interested buyers.
- **Identity Controls for AI:** We view identity and related non-human identity (NHI) as largely separate from AI and include them as separate taxonomy categories. That said, vendors such as **Linx** bring these together (including AI) under a common umbrella.
- **AI Communications including MCP Security:** One area that we were tempted to include as its own category was MCP Security, but we still deem this to be a developing area. We like how **Teleport** integrates MCP security into its solution, and we admire the work being done at **Runlayer** to secure MCP. We will keep an eye on this area as it develops.

The TAG AI Taxonomy should be viewed as a tool rather than as some canonical reference point for categorizing vendors. That approach does not match the fluidity and dynamic nature of modern cybersecurity or AI usage. That said, we hope the material in this chapter has been useful for you.



CHAPTER 7

VENTURE CAPITAL STRATEGIES FOR AI SECURITY

This chapter provides an overview of investment strategies for AI security and highlights the approaches being taken by four of the most prominent venture capital firms in this area.

Few if any modern technologies reach scale without venture capital. This has certainly been true for general cybersecurity, and it now seems to be proving true for AI-based cybersecurity. One of the clearest signals that AI security is becoming a durable category is not vendor marketing, analyst commentary, or regulatory pressure, but rather the movement of venture capital toward AI security-centric platforms and away from many traditional cyber segments.

It is worth noting that our focus and attention on venture capital-based startups were fueled by the discussions and research required for "Reaching the Chasm" (Columbia University Press, 2025), a book written by our lead analyst at TAG, Ed Amoroso. The intense work required to gain insights into the startup funding process exposed clear signals that AI security was heating up as a topic.

PROGRESS FROM 2024 TO 2026

Specifically, during 2024 and 2025, our team at TAG observed a notable shift in investor conversations. Where once late-stage cybersecurity infrastructure, endpoint tools, or incremental cloud security solutions dominated funding discussions, venture capital firms increasingly redirected attention toward AI platforms, AI-native infrastructure, and companies focused on securing AI systems.

This reallocation was not subtle, nor was it temporary. It reflected a broad investor belief that AI represents a generational platform transition, one that will simultaneously reshape software, security, and enterprise operations. Accordingly, we saw 2024 as a year of planning by both practitioners and investors, followed in 2025 as a year of proof of concept (POC). The challenge, of course, involves how to reach production in 2026 and beyond.

Figure 7-1. AI Security Roadmap from Planning to POC to an Uncertain Future

In this chapter, we examine how leading security-focused venture capital firms are approaching AI investment, why capital is flowing away from certain legacy cybersecurity segments, and what this means for practitioners, vendors, and buyers of AI security technology. In some cases, private equity teams have also gotten into the game, although their focus is different, as we will see.

THE REALLOCATION OF CAPITAL: FROM CYBERSECURITY TO AI

It would be misleading to suggest that venture capital firms are abandoning cybersecurity altogether. Rather, what we are seeing is a gradual reprioritization within security investing.

Capital is being redirected from mature or saturated categories toward AI-driven opportunities that promise asymmetric growth. We say this based on our observation rather than formal research, but we have confidence that it is correct.

The fact is that traditional cybersecurity markets, including endpoint protection, basic SaaS posture management, and standalone vulnerability scanning, have grown crowded, feature-convergent, and increasingly difficult to differentiate. Revenue growth in these areas is often linear, dependent on competitive displacement rather than market expansion. Venture capital firms, particularly those with deep cybersecurity expertise, are acutely aware of this reality.

By contrast, AI introduces new attack surfaces, operational dependencies, and governance challenges, many of which cannot be addressed by retrofitting legacy security tools. Investors recognize that enterprises will spend on AI security not because of compliance mandates, but because AI systems increasingly sit in the critical path of cost reductions that organizations have made to Wall Street, investors, and other stakeholders.

WHY VENTURE CAPITAL SEES AI AS A GENERATIONAL GROWTH MARKET

As a result, venture funding is flowing toward companies that support the taxonomy categories listed in Chapter 6. In many cases, the same venture capital firms that once led cybersecurity investing are now explicitly branding themselves as AI or AI-security investors. Getting more specific, we can say that from these investors' perspectives, AI differs from prior security waves in three important ways.

1. **Cost Reduction Funding:** First, AI adoption is budget-attached to transformations driven by CEO-supported committees and working groups. Enterprises deploy AI to reduce cost and automate workflows. They understand fully that security is required to enable this AI adoption. CISO teams therefore participate indirectly in driving these cost and staff reductions, rather than merely acting as insurance against loss.

2. **Dependency Risk:** Second, AI systems create compounding dependency risk. Once embedded into business processes, AI models and agents cannot simply be turned off without operational disruption. This dependency creates what can be viewed as an emerging long-lived demand for platforms that ensure availability, integrity, and trust, characteristics that investors favor when evaluating durable markets.

3. **Emerging Market:** Third, the AI ecosystem is still immature. Tooling for governance, posture management, runtime protection, and agent security remains fragmented. Venture capital thrives in precisely these conditions, where category formation is incomplete and architectural winners have yet to emerge. Obviously, there is risk when these markets form, but that's the whole point when it comes to venture funding.

HOW LEADING SECURITY VCS ARE INVESTING IN AI SECURITY

Before we continue, we should pause to emphasize something mentioned above in passing— namely, that enterprise AI adoption has been for cost reduction, not revenue growth. This has simply been our observation in working with over 100 different enterprise teams deploying AI. None—and we mean zero—are doing this to create new jobs, new opportunities, and new capabilities. In every case, the AI is all about streamlining and automating.

To make the discussion in this chapter more concrete, we will examine how four well-known security-focused venture capital firms are positioning their investment portfolios around AI and AI security. Our goal is not to catalog individual deals exhaustively, but to highlight patterns of intent and strategic emphasis. We offer a callout to these fine firms for helping us learn their investment patterns and approaches. Let's dive into how they are working in this area.

Ballistic Ventures

The approach at Ballistic Ventures involves focus on AI-Native security as a so-called first-principles problem. To that end, the firm has been explicit in framing AI security as a fundamentally new problem space rather than an extension of traditional cybersecurity. Their investments reflect a belief that AI introduces novel threat models, including theft, poisoning, inference abuse, and agent manipulation that require purpose-built platforms.

Across its portfolio, Ballistic has backed companies focused on protecting machine learning pipelines, defending models at runtime, and securing AI infrastructure. The common thread is deep technical differentiation and AI-native architecture, rather than incremental enhancements to existing security. This aligns with Ballistic's broader philosophy of funding technically ambitious teams willing to redefine categories rather than defend incumbents.

The Ballistic ventures group includes an all-star team of cybersecurity industry icons and investors too numerous to mention. Readers can learn of the team by visiting the Ballistic website. Key AI-related investments at Ballistic have included Pangea, which was acquired by CrowdStrike, and Reach Security, which uses an AI platform to detect misconfigurations and close security gaps. Witness AI and Nudge Security (highlighted in the interview below) are also important investments.

AN INTERVIEW WITH RUSSEL SPITLER, CEO, NUDGE SECURITY

We recently spoke with Russ Spitler, CEO of Nudge Security, to discuss why SaaS security has become an essential foundation for understanding and governing AI-driven SaaS application usage. As enterprises rapidly adopt AI-embedded SaaS tools across business functions, Nudge's approach highlights how visibility, context, and behavioral insight are becoming prerequisites for effective AI security. Here is a portion of the interview.

TAG: Why has SaaS security become so central to the AI security conversation, especially for enterprise teams trying to understand their exposure?

SPITLER: AI is now deeply embedded inside the SaaS layer, where real work happens. Security teams are no longer just protecting infrastructure or endpoints. They need to understand how employees are using hundreds or thousands of SaaS applications that increasingly embed AI features by default. That is where sensitive data is accessed, transformed, and sometimes unintentionally exposed. Without clear visibility into SaaS usage patterns, permissions, and behavior, it's not possible to fully manage enterprise AI risk.

TAG: How does SaaS security for AI differ from traditional CASB or cloud access controls that enterprises may already have in place?

SPITLER: Traditional controls focus on access and configuration, which are necessary but incomplete for securing AI-driven SaaS environments. AI introduces new behaviors such as autonomous actions, content generation, data enrichment, and cross-application workflows. Nudge focuses on understanding how SaaS applications are actually used in practice, who is using AI features, what data is involved, and whether that usage aligns with policy and business intent. That insight is what allows security teams to manage AI risk without slowing the business.

TAG: From your perspective as a CEO, what role does Nudge play in helping enterprises prepare for a future where AI SaaS usage is the norm rather than the exception?

SPITLER: Our role is to give organizations clarity and control. AI in SaaS is not going away, and trying to block it outright is unrealistic. Instead, enterprises need continuous insight into what is happening across their SaaS ecosystem so they can guide AI usage responsibly. By surfacing risk, misuse, and unexpected behavior early, we help security teams stay ahead of AI-related issues while still enabling innovation. That balance is what makes SaaS security so critical in the broader AI security landscape.

Evolution Equity Partners

The theme at Evolution Equity involves operationalizing trust in AI systems. The well-known investment firm approaches AI security from a risk-centric and operational standpoint. With a strong background in scaling enterprise security companies, Evolution Equity has gravitated toward startup platforms that help organizations operationalize trust, assurance, and resilience in AI deployments.

Their AI-related investments tend to focus on model risk management, adversarial testing, and infrastructure protection, reflecting a belief that enterprises will demand measurable assurance before allowing AI systems to operate autonomously or at scale. The firm's thesis implicitly rejects the idea that AI security will remain a niche concern. Rather, they believe it will be embedded into mainstream enterprise risk and compliance workflows.

Evolution Equity boasts an experienced team, including founder and managing partner Richard Seewald as well as industry icon Taker Elgamal. Key AI investments have included a successful funding and acquisition of Protect AI by Palo Alto Networks in July 2025. Evolution Equity has also funded and supported a startup called Defined.ai that processes and enriches training data for machine learning.

ForgePoint Capital

This iconic cybersecurity investment house tends to treat AI security as an extension of enterprise data risk. ForgePoint Capital has long emphasized the intersection of security, data, and regulation. Its AI investment strategy naturally reflects this orientation. Rather than treating AI as a standalone technology, ForgePoint views AI systems as new consumers, processors, and amplifiers of enterprise data risk.

As a result, ForgePoint has invested in AI security companies focused on data leakage prevention, misuse detection, and governance controls for generative AI and LLMs. This perspective resonates with regulated industries, where AI is constrained as much by data exposure and compliance risk as by technical vulnerability. ForgePoint's approach suggests that AI security will converge with data security posture management and privacy engineering.

ForgePoint Capital has been led for many years by two industry icons—Alberto Yepez and Don Dixon—both of whom have decades of experience in cybersecurity. Key AI-related investments at ForgePoint include Nudge Security and Databahn.ai, which collects, monitors, and optimizes data ingestion—key functions for AI implementations. GetReal Security (highlighted below) is also an investment.

GetReal

AN INTERVIEW WITH MATTHEW MOYNAHAN, CEO GETREAL SECURITY

We spoke with CEO Matthew Moynahan at GetReal Security about why AI-driven impersonation, synthetic media, and deepfake threats are now attracting serious venture capital attention. Backed by ForgePoint Capital, GetReal illustrates how AI security investment is expanding beyond traditional infrastructure to include trust, identity, and human-centric risk.

TAG: Why are investors increasingly viewing deepfake and impersonation risk as a core security problem rather than a niche concern?

MOYNAHAN: Because the fundamentals of trust in the enterprise have changed. Remote work, global teams, deepfakes, and the replacement of in-person interaction with voice and video have changed how identity is established and trusted. As a result, the attack surface has shifted from systems to people. Deepfakes exploit trust, authority, familiarity, and urgency, but they're just

Continued

GetReal *Continued*

one manifestation of the broader digital identity and trust problem. Investors recognize that generative AI lowers the barrier to creating convincing synthetic content, which makes social engineering cheaper, faster, and more effective. It's a new world where adversaries bypass existing controls without malware or exploiting a single software vulnerability. And, this is no longer a future risk. It's already impacting executive communications, account takeover, financial approvals, and hiring.

TAG: How does AI change both the offense and defense in this area?

MOYNAHAN: It accelerates both sides. Attackers can generate high-quality fake audio, video, images, and identities at scale. Defenders cannot rely on manual review or intuition. The only viable defense is AI-driven detection that identifies subtle artifacts, behavioral patterns, and contextual inconsistencies across media and identity signals. At GetReal, we focus on providing defenders with forensic evidence of content or identity manipulation in real time. And that evidence enables confident decision making, supports investigation and response, and delivers actionable intelligence rather than just alerts that breed uncertainty, inefficiencies, and user friction.

TAG: Who ultimately owns this type of risk inside large organizations?

MOYNAHAN: That's part of the challenge in this area of risk. That is, identity and deepfake risk spans security, fraud, legal, communications, and executive leadership. Investors see value in platforms that can bridge those silos and support coordinated response. The companies that win here will be those that help enterprises operationalize trust across both technical and human dimensions, addressing not only deepfake images and videos, but also impersonation, fraud and insider threats that corrupt business decision making.

Crosspoint Capital

The approach at Crosspoint Capital reflects an interest in platform-level defenses for AI infrastructure. Its investment posture reflects a platform-oriented view of AI security. Rather than isolating individual threats, Crosspoint has shown interest in companies that secure AI infrastructure holistically, including compute environments, APIs, agent communications, and model access controls.

This mirrors the firm's emphasis on backing leaders capable of scaling into large, durable platforms. In the AI context, this means focusing on vendors that can evolve alongside complex agentic architectures and distributed-AI systems. Crosspoint's investments suggest a belief that AI security winners will resemble cloud security platform leaders of the prior decade with emphasis on being broad, integrated, and deeply embedded into enterprise operations.

Crosspoint Capital is well-known for its ownership of the RSA Conference, as well as its iconic leader, Greg Clark. Key AI-related investments include Calypso AI, which provides security for AI applications and agents and was acquired by F5. Crosspoint Capital has also funded Knostic, which offers a range of tools to protect enterprise users, data, and AI tools. Command Zero (highlighted below) is also funded by Crosspoint Capital.

AN INTERVIEW WITH DOV YORAN, CEO, COMMAND ZERO

We asked the CEO of Command Zero, Dov Yoran, to share how he thinks about AI security from a platform and infrastructure perspective, particularly in light of Crosspoint Capital's emphasis on scalable, enterprise-grade defenses. Command Zero reflects an emerging view that AI security must integrate deeply into how organizations build and operate modern systems. Here is a brief portion of that discussion.

TAG: Why do you think platform-oriented AI security is attractive to investors right now?

YORAN: Because AI is the single most potent innovative push we've had in decades. With soaring adoption, it is becoming a distributed system of models, agents, APIs, and workflows. Securing that environment requires platform-level visibility and control, not point solutions. Investors understand that enterprises prefer integrated platforms that evolve alongside their infrastructure rather than isolated tools that solve narrow problems. This is especially true for the high potential of AGI in the near future.

TAG: How does AI infrastructure security differ from traditional cloud or application security?

YORAN: AI introduces new building blocks. Models, embeddings, agents, and inference pipelines behave differently than traditional workloads. They create new pathways for misuse and new forms of dependency. Security platforms have to understand those building blocks natively in order to enforce access, monitor behavior, and prevent abuse at scale.

TAG: What does long-term success look like for AI security platforms in this category?

YORAN: Success means becoming invisible but indispensable. AI security platforms will become catalysts to drive AI adoption, consequently enabling business growth. When AI security is embedded directly into how enterprises deploy and operate AI, it becomes part of the fabric rather than an overlay. Investors favor companies that can drive revenue growth, because it signals durability, defensibility, and long-term relevance.

WHAT THIS MEANS FOR PRACTITIONERS AND VENDORS

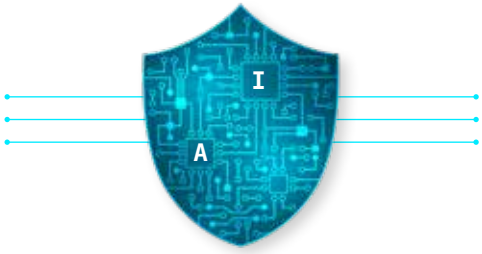
For enterprise security practitioners, the implications of this investment shift are practical rather than theoretical. As should be clear from our examples above, major venture capital allocation strongly influences which vendors survive, which features mature, and which platforms receive sustained R&D investment. Understanding where capital is flowing can therefore inform more resilient vendor selection strategies.

For startups and other commercial vendors, the investment message should be even clearer. AI security companies that merely repackage legacy controls with AI branding are unlikely to attract sustained investment. Venture capital firms are signaling a preference for AI-native architectures, credible threat models, and platforms that align with long-term enterprise dependency on AI systems.

CLOSING THOUGHTS

We can probably all agree that venture capital is not always right, but it does help explain which way the wind is blowing. The movement of security-focused investors toward AI, and away from many cybersecurity categories, reflects a conviction that AI represents the next major enterprise platform shift. AI security sits at the center of this transition, not as an optional add-on, but as an enabling condition for safe adoption at scale.

For readers of this handbook—whether practitioners, founders, or investors—the lesson should be to watch the capital flow, but also to understand the reasoning behind it. AI security is no longer a speculative edge case. It is becoming one of the most strategically important investment arenas in enterprise technology, a reality that will shape the security landscape for years to come. We will stay close to this evolution at TAG, and we will keep you informed.



A TAG ROUNDTABLE

TRYING TO GET A READ ON THE MOVING STATES OF AI SECURITY

Ed Amoroso

Kris Lovejoy

Bashar Abouseido

Ed Amoroso: Our topic is AI security, and there's a lot to cover: both securing AI and also using AI for security. Bashar Abouseido and Kris Lovejoy, welcome to the discussion. Really wonderful to have you both here. Bashar, we'll start with you. Love to hear just a couple of minutes about your wonderful career.

Bashar Abouseido: Thank you, Ed. Glad to be here. I represent about 30 years of technology and cyber management, both on the P&L side and on the corporate side. I led multiple Fortune 200 international cyber programs and technology programs across the board. I recently retired after 13 years of being the CISO of a Fortune 200 financial institution. In the last 10 to 12 years of my career there, we invested heavily in AI to help maintain trust and give cyber defenders an advantage And I have very practical experience in what I found to be working versus not, and where you can get ROI, mainly in the cyber and fraud space.

Amoroso: That's why you're here. You're perfectly well suited to comment on this. It's going to be fun to hear your perspective. And now my good friend Kristen, tell us a little bit about your wonderful career.

Kris Lovejoy: Thank you. My career has been, yes, long and "storied." For my 30 years or so in the industry, I've been kind of on all sides: pen tester, consultant, CISO for IBM for a while. I have recently moved into a different role as the global head of strategy for Kyndryl, a large corporation in the technology services space. We do a lot of infrastructure management, and my background has been in security and grappling with the trust issues, particularly around AI. And I say that because I've

also been on the entrepreneurial side. One of my startups was actually an AI company that we were selling into the intel and defense community. And so I had an opportunity to learn quite a lot about the positives and the negatives of that technology. How it needs to be implemented, integrated, managed. What the challenges are. And that has, weirdly, prepared me for a new role in strategy.

Amoroso: I'm going to guess when you were conceptualizing this strategy position, AI was a big piece of the discussion.

Lovejoy: It was the biggest component. There's a lot of disruption happening in the market right now because of AI. It is inextricable from the question of digital trust. I think the two are inexorably intertwined. And so a lot of organizations are really looking for people who have the two experiences.

Amoroso: I want to start with something broad, because we've seen different waves of innovation. AI seems to be the thing now, but we remember others. Bashar, when you see these new ideas, these new waves of—I think innovation is the right word here for AI—waves that can change, disrupt, modify processes or technology, what comes to mind? Do you immediately think, "Wow, we've got to adjust. We've got to ride the wave. We've got to look for guardrails"?

Abouseido: First thing that comes to mind is we have an opportunity. Innovation brings opportunity. Change also brings an opportunity. But change and innovation always come with risk, and we are kind of in the field of managing and helping manage risk. Therefore, our job is to support innovation, to look for ways to make sure that we understand the risk of taking on that innovation and producing the appropriate desired outcome for our businesses. And use that innovation to help make the world a better place. I've been fighting cyber crime bad guys all my life, and I want any opportunity I can get to help the defenders gain the advantage. We have an explosion of digital transformation that started before the pandemic. It continues, and I think we're more interconnected than ever before. And we're here to build and maintain the level of trust that people need in order for this to thrive and continue to expand.

Amoroso: It's funny, when I'm talking to my grad students at NYU, we use the word "invention" for "innovation." I think they're really the same word. Kris, I want to ask you about the word "disruption." In our normal, everyday lives, if we talk about disruption, it strikes me as a somewhat pejorative thing. But in our world, disruption is kind of the whole idea, isn't it? And we'll get into the AI thing, but when you see disruption, and you do strategy for a living, is disruption generally the whole idea? is that kind of what we're trying to do here?

Lovejoy: I think so. I think that's kind of the market force. I think we're all chasing disruption. We're chasing that new idea. What is the new thing that is going to change the world? I know you said let's pause on the AI conversation, But I do think that this is a disruptive phase.

Abouseido: I agree with Kristen. I think it's disruptive evolution. It's not a revolution by any means.

Amoroso: Evolution is slower than revolution.

Abouseido: Yes, exactly, It is a disruptive evolution of all the technologies and innovation that we have accumulated over time.

Amoroso: When I ask people, "What are you disrupting?", they usually say two things. But I think it's only one. Let me tell you what they are and see if you guys agree with this. And let's talk about AI specifically. They'll always start with: "I'm going to use AI to create new capabilities and services and products." And then they'll say, "And also, we're using it to streamline and reduce costs." And then when I look at what they're actually doing, it's one little jigger of the first thing in about 20 ounces of the second. It's a lot of cost reduction. We're all practitioners. You want to be positive and build

"I LISTEN TO THESE AI STARTUPS TELL ME THAT THEY'RE WORRIED ABOUT THE PROBLEMS OF PROMPT INJECTION AND MODEL BIAS. BUT THAT JUST SEEMS LIKE HOUSEKEEPING. I'M WORRIED ABOUT AUTONOMOUS ATTACKS."

Ed Amoroso

things, but man, most of what I've seen in the application and use of AI recently has been more about streamlining as opposed to creating something. With my phone I'm using facial recognition and neural network. That's a new capability. But more recently, I see less capability, more cost saving. Kris, we'll start with you. Am I getting this wrong? Am I being too negative?

Lovejoy: No. I was just thinking about the comment about evolution not revolution. I've been thinking a lot about this particular subject, and what I would liken to the period of time that we're in. It's like when we first invented electricity. If you think about how people experienced electricity, they'd go to a fair, they'd go to a carnival. Somebody would hold up a light bulb. It would go on. "Oh, wow." It was like it was the Age of Experimentation. It didn't become something we could use in a profound way until we built the transformers, we built the transmission, we built the distribution, we built the infrastructure to enable it to work. I think what you're seeing right now is we're coming out of the Age of Experimentation. We've got a lot of POCs (proofs of concept). We're using it. We're lighting up a building, if you will. We're lighting up a street. But we don't have it full scale, integrated into our environments. And I think we're going into a period of transition right now, which is important. I tend to think about AI—it's not a product, it's infrastructure. It's cognitive infrastructure. More specifically, I'm beginning to think about it more like cognitive middleware. That cognitive middleware has to be integrated into our environments. We have to have some fabric, or some capability that enables us to integrate this in a standard way so that organizations can take advantage of the value of AI. I don't think we're there yet. I think we're moving into a transitionary period. And I think in this transitionary period we're going to be focusing a lot on of the integration of AI before we can full-scale use it to the advantage that we all foresee. Now that period, I don't how long it's going to be. Five months? Five years? Five hundred years? Arguably, some people are in the Trough of Disillusionment. I feel more optimistic. I think we're in that transition period. But I do think that we are going to be living through this for the next few—I don't know, ergs—before we get there.

Amoroso: So maybe you can't quite answer the question yet whether the real benefit is going to be streamlined costs or fundamentally new capabilities.

Lovejoy: I think it's going to be both. I think it's easiest right now because we don't trust it. Going back to the security and resiliency piece, when you look at the things that are hindering people from actually using it, it's the inability to scale. And the inability to scale is largely due to the inability to trust at scale. And so there's no auditability, there's no real compliance around it. If you're using agents that are polymorphic, how do you know that they're not evolving beyond their parameters? There's a lot of stuff that's happening that we haven't figured out the answer to. Once we figure that out, then you can unlock it, unleash it, and heaven knows what we can actually provide.

Amoroso: That is such a good point, because you can look at it from two sides of the coin. If you think about an autonomous AI-enabled robotic assistant or companion for grandma in the nursing home, you can think of that as a new capability. But it could also be replacing the human that was doing

that job. So it's sort of both. Not really sure at this point. I think you're right about infrastructure. And Bashar, I think you're right about evolution. We saw a big ramp when ChatGPT popped up, but since then I think there's been a leveling off of this advancement, as we try to digest—to Kristen's point.

Abouseido: We've seen that across all industries and in technology in particular. We've seen it from the time of horses versus vehicles, and the modes of transportation. We've seen it with the internet, where initially we thought it was just only for information, and then became more connectivity, and then commerce. And then we got to substantial benefits that we cannot live without. I feel we have to go through the same levels of maturity to learn about how to use this capability. And I like to focus on skilled intelligence. I think that's the key. I am not looking at it as pure cost savings. I think of it as elevating to the human brain the most critical and important thing in life, and removing the repetitive, boring tasks for which the human brain is really not being utilized at the level that it should be. Because if a machine can do it, we should be able to do something more productive with our abilities. And I think that's elevating the human factor to be a lot more impactful, without removing it. And there are many cases where I think the human brain can be leveraged in a substantially more effective way than the ways we do it today. But we should have a choice, and that's what we're going to learn—where we will make those choices. When do we just delegate to a machine, because we trust it— to Kris's point—and we think it's appropriate, it's economical, and it makes sense? And when do we think, "No, I need the human in the loop. I need the human to perform this."? And I think we'll arrive at some type of equilibrium between how much we want the machine to do, and how much of it we want humans to be fully in charge.

Amoroso: What do you both think about security now? We're all security experts. Is security the balancing component? Like, if we don't get the security right, then we can't really leverage all the nice things Bashar was just laying out? Kris, is that the right way to think about the cybersecurity component?

Lovejoy: Personally, I don't think AI can be adopted without security. And I think there's a positive and a negative. One of the reasons we've haven't seen a run toward the technology—there are some technical limitations from a scale perspective, but a lot of it is trust. However, our adversaries don't have those kind of ethical quandaries. And so we're seeing a lot more novel uses of the technology for offensive reasons, which I think is something that we have to be very aware of and concerned about. But yes, I do think that cyber is the centerpiece of everybody's strategy, or should be the centerpiece of everybody's strategy right now, vis a vis AI, for both offensive and defensive reasons.

Abouseido: I agree. Security is a fundamental component of the trust we need to maintain in the new capability and the systems of AI. Now I think AI will re-baseline the cyber security industry. I have no doubt in my mind it will re-baseline everything we do, because the reality is the bad guys, the bad actors, are all preparing big plans to leverage AI and to leverage machines. And you will only be able to stay ahead if machines are able to defend machines, and we're able to scale with AI and machines to defend against AI and machines.

Amoroso: That's music to my ears. I listen to these AI startups tell me that they're worried about the problems of prompt injection and model bias and so on. That sounds so trite to me. I know that those are initial issues and you've got to clean them up. But that just seems like housekeeping to me. I'm worried about autonomous attacks. That scares the life out of me. Do we envision a future where it's robot against robot? More autonomous offense—we know that's coming. And then Kristen, you would be sort of putting in place, and Bashar, you and I would certainly be coaching a much more autonomous defense. If machines are coming at you, I don't think you can go after that with people, right? You have to have machines. Is that where we're headed? Kris, what does that sound like to you?

"WHAT IF THE ANALYST HAS MADE A MISTAKE AND TRAINED THE AI ON THE WRONG INFORMATION? AT WHAT POINT DO I NEED TO REVERT THE MODEL? HOW DO I EXTRACT THE FEATURES?"

Kris Lovejoy

Lovejoy: Oh, absolutely. I think Anthropic's Claude-related event was pretty eye-opening. And something we all forecasted. And there it was. Now it's not fully autonomous, but it's just one step away. I'm looking at it more from a commercial perspective. I mean, we're running at building more AI capabilities into the SOC. It is forcing us to rethink our SOC architecture. What we've realized is in order for us to use AI at scale to enable us to protect our customers, you've got to integrate network monitoring as well as security monitoring, and so SNOCS (security network operations center) are going to be the next big step. And then after the SNOC [she laughs], you're going to see more of the fusion centers. Bashar. I'm sure you have a fusion center already, but most of the world is not as sophisticated as the tier one financials. So I think for them, at least, these concepts are pretty new. But this kind of integrated approach is going to be the way it is in the near future.

Abouseido: But I think AI will come in to democratize that scale and that level of intelligence for all levels. It's not going to be limited to the ones with all the resources at the highest end, with the team with the largest number of resources—human resources. It is going to be a machine versus machine war, and it will be those who have more sophisticated machines that are dynamic in the response. In my career, I've seen digital processes take about four weeks to get discovered on the internet. And then it gets attacked. Before I left, it was less than five minutes before a digital process is discovered, mapped, and automatically attacked. Now it's robotically, it's not an intelligent attack, but it is a fairly sophisticated level of attacks. But what we've seen with Anthropic is that you will have thousands of agents of AI constantly discovering and doing reconnaissance—discovering and identifying vulnerabilities, possibly having access to the code and identifying in real-time opportunities or vulnerabilities that they should attack. Building the script and actually exploiting it. So what we're going to see, instead of every two to three months a new massive Zero Day vulnerability is discovered, we're going to see them weekly, daily, and maybe hourly. That is a very, very scary scenario. If many of these bad actors are going to deploy thousands of agents to constantly discover Zero Days in all the technologies we use, and be able to exploit them almost instantaneously, then we need to have a machine-level intelligent defense that can also, at the second that it gets initiated, detect and respond to it. And have some type of enforcement. I still think for things not to get out of control, that we always need a human in the loop and strategy behind it, and we need to understand when to go autonomous versus not.

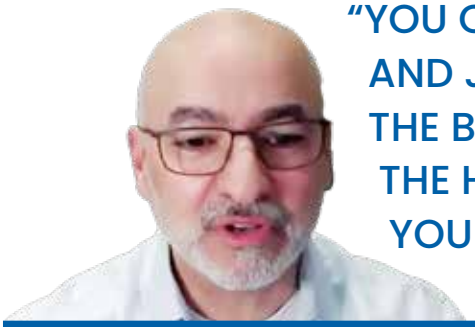
Amoroso: Such an interesting topic to debate. I'm on the side of what you guys are alluding to, machine versus machine. But a lot of people would say that it's a little more uncertain at this point. Like a typical SOC team right now—my observation is that the way they're responding to this is that the Tier 3 analysts—you guys have both had Tier 3s work for you. They're usually more experienced. They're good, and you rely on them. They give you your briefings, they give you the PowerPoints, they explain what happened. But then the Tier 1s usually are people you've just hired out of university. What I'm noticing is that those hires are slowing down now. That's the first visible thing that I see: that it's harder for a young person to come in through the SOC. And that you buy one of these SOC co-pilots,

or like next generation SOAR systems that automates the workflow. You lay it out, and all the things you would have been doing as a human Tier 1—it's all workflow and it's autonomous. It's pretty good. I watch the demos, and I go, "Not too much I hate about that." And it works all night and all day, and it doesn't take breaks. So that's what I see happening. Now, maybe that's your evolution. You were saying, Bashar, that we like don't jump immediately to machine versus machine. It evolves. But that's what I see. That has a real impact on human beings. It's fewer jobs. So you could argue that one sort of dystopian view here is that it becomes a little harder for human beings to develop skills and to develop the background. Maybe I'm wrong.

Abouseido: I have a different point of view, because I went through multiple transformations of the SOC. And I ended up in a very nice place. I still think more transformation is needed for a SOC, but the typical situation is they're understaffed for the volume they process. They're mainly process-based, and even with all the engineering you do, you end up focusing on 20% of all the things you need. And the Tier 1 analyst is just swimming in alerts—mostly false positives. They're not getting to the core risk that they need to address and respond to, and they end up missing a lot, and that shows up in the testing. They're constantly not seeing things that they should be seeing. They feel like they're not contributing appropriately, because they just simply cannot handle that level of volume. If everything is important, nothing is important. Now with AI, and just the initial part of the transformation, AI is able to take a look at every single alert, not the chosen 20% that we decide are more important. But if you continue to see what happens when you have humans in the loop, it's to continually teach AI how to understand your context, how to understand and work for you, specifically as a business. If everybody's using the standard models, then we all have the same defense that's not specific for your business. As you start building a loop between the analyst and AI, you're starting to teach AI context. You start to teach AI how your business is running, and you're able to elevate and escalate only what you consider to be a near certainty of what the threats are, where the risks are, and allow the human to start getting involved in where we need to make a decision and why. The second part is our approach in the SOC today is very static. You write the rules. It's very deterministic. It's black and white, either you look at it or you don't. What I think we need—because here's what the bad guys do: As soon as they figure out exactly how you respond, they always reverse-engineer. As soon as they know exactly where your limits are, they start going below to evade detection. But what we want is to give that control back to the defenders in the SOC. That means we need to be dynamic. That means I should be able to come in and say, "For this week we have a new acquisition. I am going to focus on the new acquisition. I'm releasing a new digital process. I am going to focus on the risk associated to that today, because that's the new launch." For this week, this month, whatever it is, the leaders within that particular team should be able to decide how they want to defend based on their business needs versus the static nature of what we've been doing the last 10 to 20 years. AI will help you do that, but it takes expertise and takes humans to be fully a part of that conversation.

Amoroso: That makes sense. Kris, I want to shift a little bit away from using AI for security and now talk about securing AI. Because I know that's a big part of the strategy work that you're doing now. So when a board member, or senior leader, or one of your peers says, "Hey, should we be putting guardrails up around AI?" what comes to mind? Is that a data leakage problem? Is that an access problem? Is it avoiding hallucination and bias? Is it something else? In your mind, when you hear "guardrail," what are we guarding against? What is it we're worried about when the business starts to use generative AI for writing reports and doing research?

Lovejoy: Where I'm more focused right now is around agentic AI, because that's where I think we're going. And Basher was just talking about the way in which these technologies are going to be used for a SOC. My early company, **BluVector**, which I think you probably remember I built around advanced threat detection back in like 2019. The way it works is you have two forms of learning. You can either do the automated learning, where AI teaches itself. Or it can be reinforcement learning. In



> **"YOU CANNOT JUST USE THE PUBLIC MODELS AND JUST BE HAPPY AND MAKE DECISIONS. THE BIGGER SCOPE OF DATA YOU HAVE, THE HIGHER THE PROBABILITY IT IS FOR YOU TO GET HALLUCINATION AND TO GET THINGS WRONG."**
>
> Bashar Abouseido

either case, what happens is the AI begins to gather more context. New features are integrated into the corpus. It's becoming more and more intelligent. The challenge is, What if it's being trained on the wrong thing? What if the analyst has made a mistake and trained the AI on the wrong information? At what point do I need to revert, or revert the model? How do I extract the features? Let me tell you, if anybody's tried to do feature extraction, it ain't easy. And without auditability, without clarity around exactly how AI learned, at what point it learned, why it learned, how it changed, it becomes very hard for you to untrain AI. And so when I think about security within this context. It's about observability. It's about auditability. It's about ensuring you understand, from a learning perspective, exactly how the model was updated, and what does one need to do to back that model off so that it can once again be effective. This is not something that we have fully invented yet, quite honestly.

Amoroso: With a neural network, that is a really hard thing.

Lovejoy: And so this is where I I'm spending a large part of my energy right now. The way we're doing it is we test agents within what is essentially a digital twin to see how it's evolving. We apply a policy, and then we monitor and enforce the policy based on the variables that we've established for how we believe an agent should be evolving. So what we're essentially doing is putting our point in time ahead of where the agent is, and saying it's acceptable for the agent to evolve to these parameters, but no farther. Once it gets to those parameters, then we're either going to back it off, we're going to refresh it, or we're going to allow it to continue—but with additional monitoring and changes to the policy. And then auditing, which we haven't necessarily figured out yet. This is how we're thinking about it. This isn't like a product you can buy. These are capabilities. There are lots of different capabilities, some of which we've had to invent on our own to make this happen. But these are all things that, going back to the question of scale, are the components of an effective digital trust system that I believe needs to be enabled. It's essentially allowing us to instantiate a policy based on a future perspective on how, again, that agent is going to evolve. And then ensure that we've got the right way to assess and deconstruct how that agent has learned in the event that they've learned something that we didn't know, or ultimately, we need to retrain it, or untrain it, to perform the actions that we expect.

Amoroso: Interesting. Bashar, you've been involved in both B2B and B2C type environments. Thinking B2C, I've been a customer of places that you've been in charge of the security piece pretty frequently. With human beings, they make mistakes, and they have bias, and they can do something really stupid. When we look at AI, we say, "Gee, it can make mistakes, and be biased, and do stupid things." Did we always have that problem and just see it differently, or is this a more intense problem? Because our whole industry is obsessed now with the AI making stupid mistakes, and I'm for fixing that. But I've seen human beings and I've been in telecom my life, and I think occasionally some mistakes have been made.

Abouseido: I think it's an unrealistic expectation to think we will be right 100% on anything we do. We deal with this. My learning with AI is that it took us quite a bit to understand how to get benefits from AI, how to establish true ROI because we just didn't understand the difference between the capabilities and what we're trying to accomplish. And we need to understand the difference between a generative AI that has the ability to have a large language model where you can speak to it and interact with it, which is a very probabilistic type of approach. It's how it allows for innovation, because it is going to take a lower level probability and consider it and iterate on it. Versus autonomous. When you deal with autonomous decision-making, you want to use AI not to consult, not to look for ideas, to do some writing, to build proposals, where you still have a human looking at it and applying the last touches, but you want some creativity. If you move to autonomous driving, you're dealing with critical decisions. You want a tighter set of data and stronger kind of focus around predictable AI or predictive AI, because you want a higher probability of what the outcome and the decision is going to be. And you have to differentiate. You cannot just use the public models and just be happy and make decisions. The bigger scope of data you have, the higher the probability it is for you to get hallucination and to get things wrong. And all of us, when we make decisions, we don't get 100% of everything right, and we don't even assume that we're 100% right, but we have a high likelihood. We have a set of data that is very focused, that says, If I wrap this level of predictable outcomes with very definitive, very predictable "if-this, do-that," then I should be able to make a decision. And I don't think we've reached that maturity yet, meaning there are places for creative interaction. And I think we need guardrails because the risks are clear, and we need to be able to understand what that is.

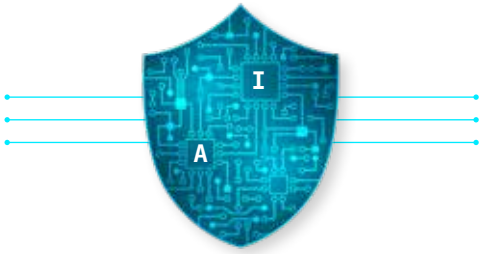
Amoroso: Super interesting. In the last few minutes that we have here, I'm going to ask you each to give a bit of sage advice to people following us here. I'll start by saying that for readers, I think that there are two ends of the spectrum here. There are people who are underestimating the impact of AI, rolling their eyes. I think they're wrong. I think this is a transformational technology. But at the other end of the spectrum, there are also people who see this as ending the human race. And I've always seen that as kind of silly. We always lose our minds a little bit with new sciences. Like math had to throw out numerology, and chemistry had to throw out alchemy. And maybe computing has to throw out some of the nutty stuff you hear about. You know: the end of the human race. But somewhere in the middle, I think you've got a very interesting thing happening. So that would be my advice. Kris, I'll ask you to maybe provide a little bit of advice, and then we'll give Bashar the final word

Lovejoy: I'll say something that I think you and Bashar would both echo. But as a security person, I think you know—and Bashar, you said this early on—our responsibility is to understand new technology, particularly that which can be disruptive, understand how it can be used, and then to assess the risk associated with it so that we can allow the business to make informed decisions and actually use it. If you are doubting the profound impact of AI, whether larger, generative AI or agentic AI, what I would say is, "Please lean into it, regardless of how concerned you are. Lean in. It is important for you to understand, from the positives and the negatives, how this technology can be used." I think for those who are on the policy side, the regulatory side, the academic side, I do think that we're going through a transitionary period right now. The jobs that are needed as we go into this transition period are going to be very different than the jobs we need once we get out of the transition period and into that kind of New Industrial Age of AI, if you will. I think we have to be very circumspect about what the impact is going to be to the workforce. I believe we all have an ethical responsibility to ensure that we've got the right job training programs in place to help us navigate this transition, as well as to make sure that once we get out of the transition, that we have a fully formed set of academic curricula that enable people to be successful. I think you and I have had this discussion before. It can go two ways here. I know you've been very optimistic. I tend to be more on the negative side. But I do think we all have a social responsibility, and we should be using our voices to ensure that people really understand what this means, and that we are really prepared to deal with the fallout, both positive and negative.

Amoroso: Such a realistic view. Thanks for participating. Kris. Bashar, the last bit of advice.

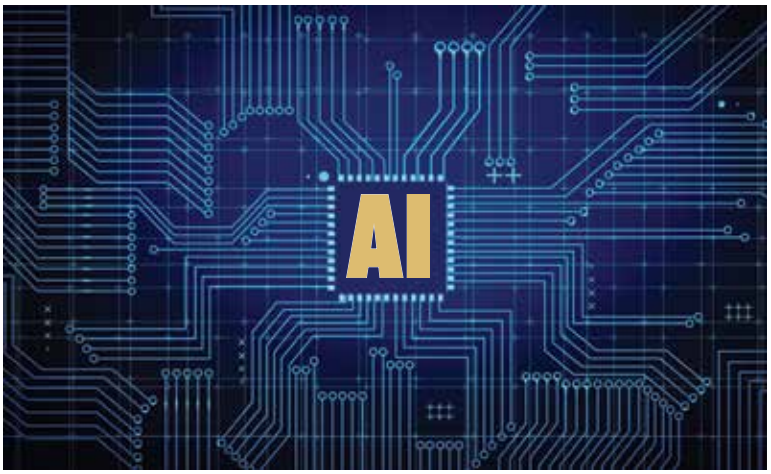
Abouseido: I agree with Kris as well. I think the potential for AI to help society across the board is phenomenal. It will improve our standards of living, make us more productive. From a cybersecurity perspective, I'd like us to think like an enabler and not to think about fear, uncertainty, and doubt (FUD). In the two years where I've seen various businesses use AI, I have not seen those risks materialize in a significant way. They're mainly conceptual so far. And it's fair to be prepared, but do not let that slow you down. Be a partner. Be part of the solution, not part of the problem. I would say AI will make us a lot more productive. It will create new jobs as well. It's been my experience that we required less resources in one place, but created new roles, new jobs that we didn't have before. And like anything else that is disruptive and innovative, it will impact sections of the economy, and it will create new opportunities. And I like to think, optimistically, that it will be a good balance. And we need to look at it from a social responsibility perspective. We should continue to be very transparent, very vocal, very accountable, and demand that across the board in order for this to be productive, to focus on good use of that technology, and not to be used only to maximize profit.

Amoroso: Let's hope people take your advice. I want to thank you guys. You guys are awesome. Thank you for the discussion.



ENTERPRISE AI SECURITY

HANDBOOK



THE TAG ANALYSTS

TAG