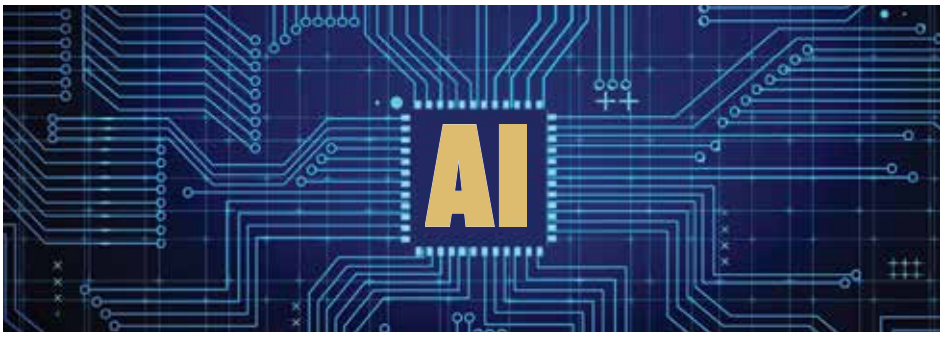
ENTERPRISE AI SECURITY

HANDBOOK

SPECIAL REPRINT EDITION



AN INTERVIEW WITH
FEATURED VENDOR

BLACKBIRD.AI

A ROADMAP FOR SECURING AI

AN AI SECURITY POLICY

THE TAG ANALYSTS

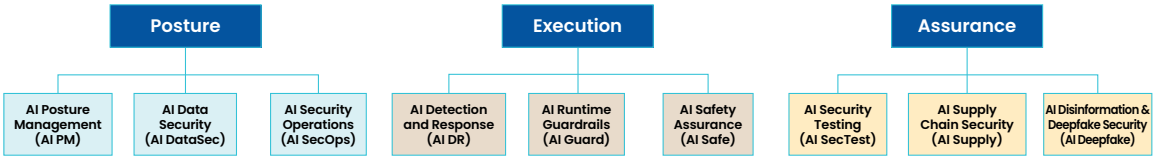
LEAD ANALYST DR. EDWARD AMOROSO
CEO, TAG INFOSPHERE | RESEARCH PROFESSOR, NYU

TAG

ENTERPRISE AI SECURITY HANDBOOK

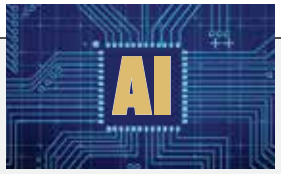
SPECIAL REPRINT EDITION

TAG'S TWO-TIERED ENTERPRISE AI SECURITY TAXONOMY



AI DISINFORMATION AND DEEPFAKE SECURITY

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. One vendor that we view as very strong is **Blackbird.AI**. We interviewed them for this volume.



INTERVIEW:

WASIM KHALED , CEO AND CO-FOUNDER, BLACKBIRD.AI

3

ROADMAP FOR SECURING AI
4

AN AI SECURITY POLICY
17

TAG

©2026 TAG Infosphere, Inc.

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?

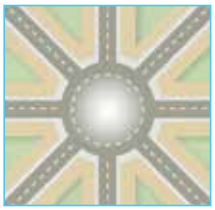
KHALED: The scale and speed of synthetic content generation make manual detection unrealistic. Human review simply cannot keep pace with the volume, variety, and sophistication of AI-generated media. Effective defense requires actionable AI-driven Narrative Intelligence detection and analysis that can 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.

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).

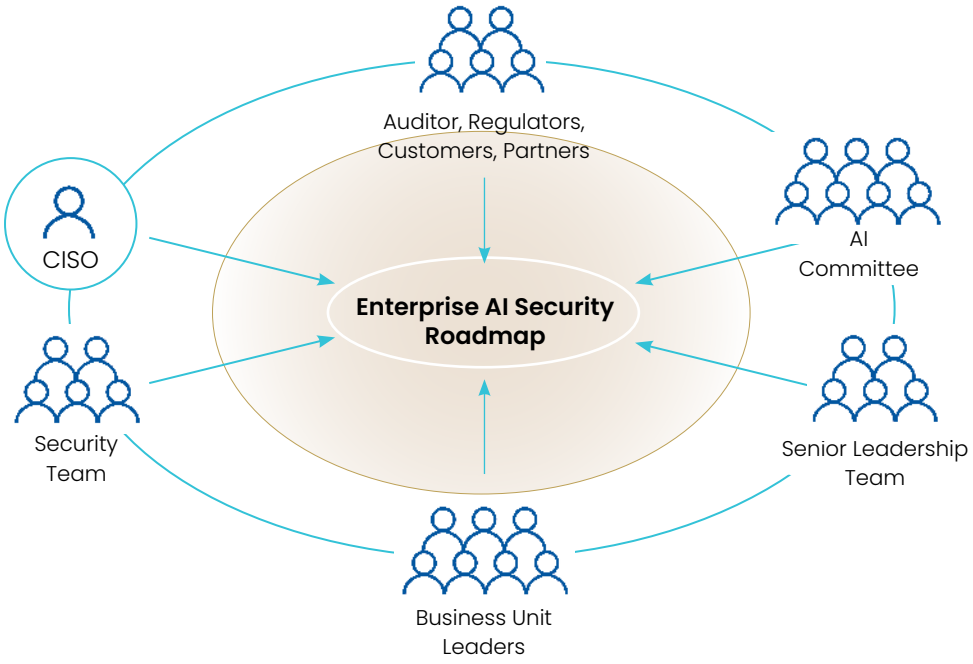


Figure 2-1. Management Ecosystem for AI Security Roadmap

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).

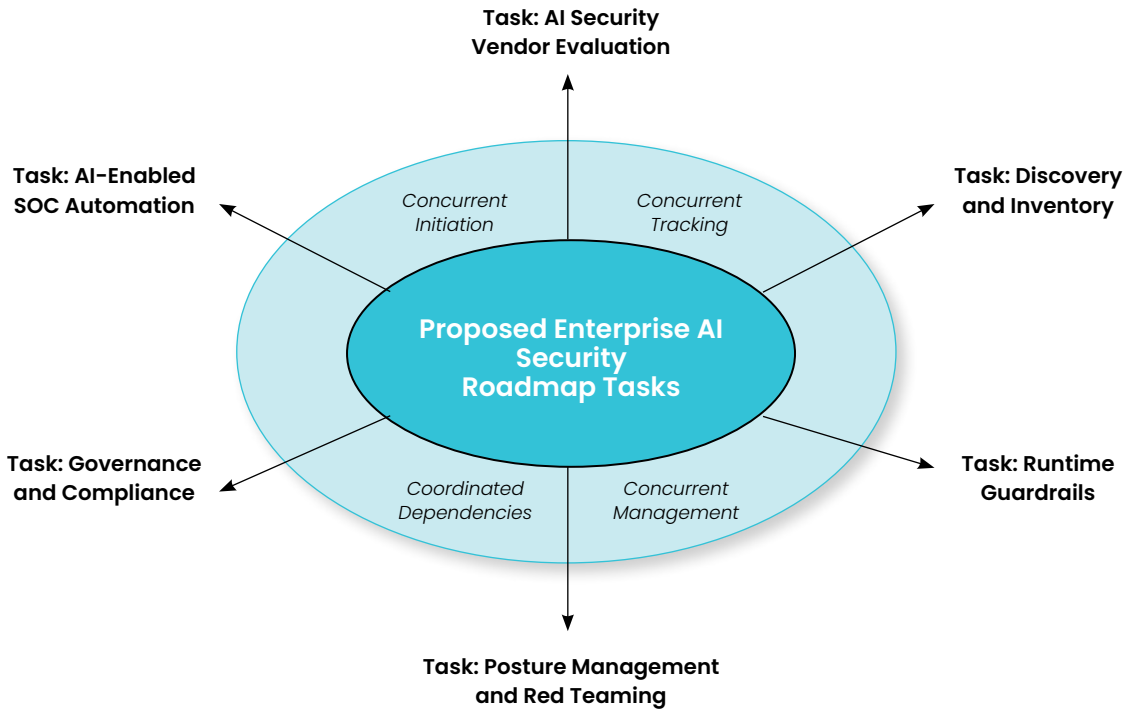


Figure 2-2. Enterprise AI Security Roadmap Tasks

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.

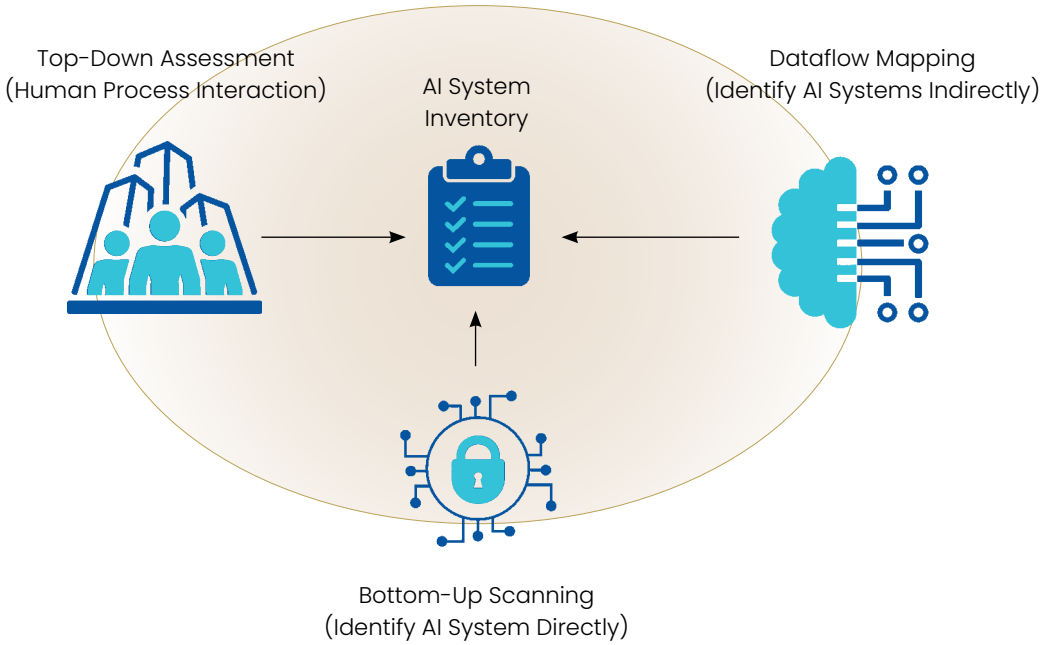


Figure 2-5. Three Strategies to Develop an AI Inventory

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.

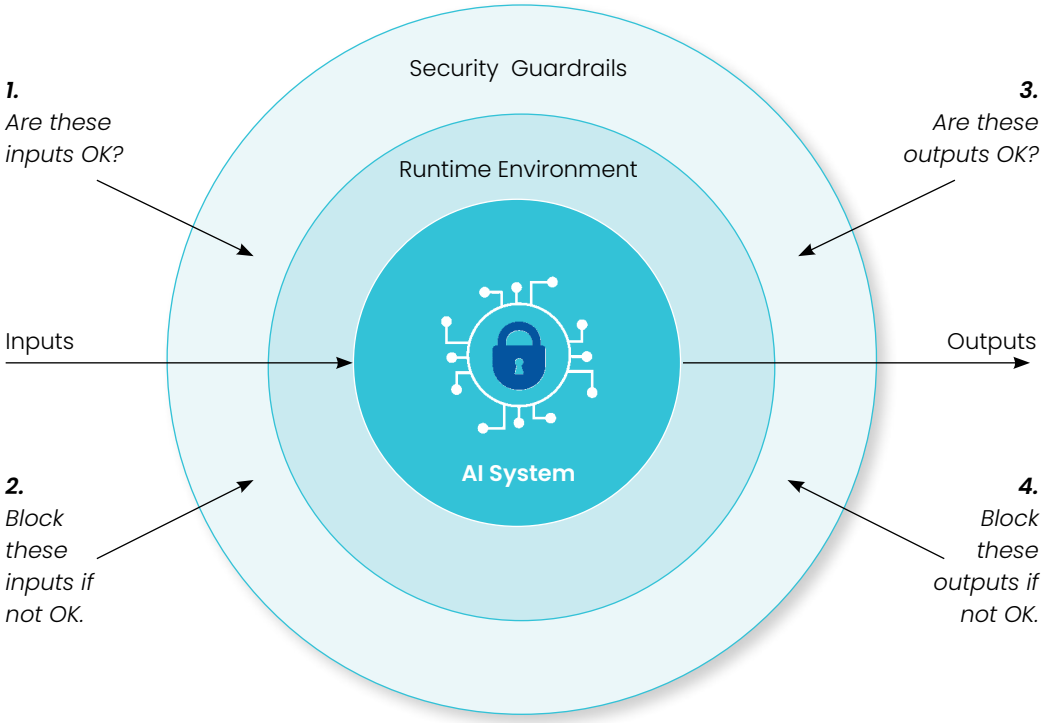


Figure 2-6. Runtime Mediation Schema for AI Security

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.

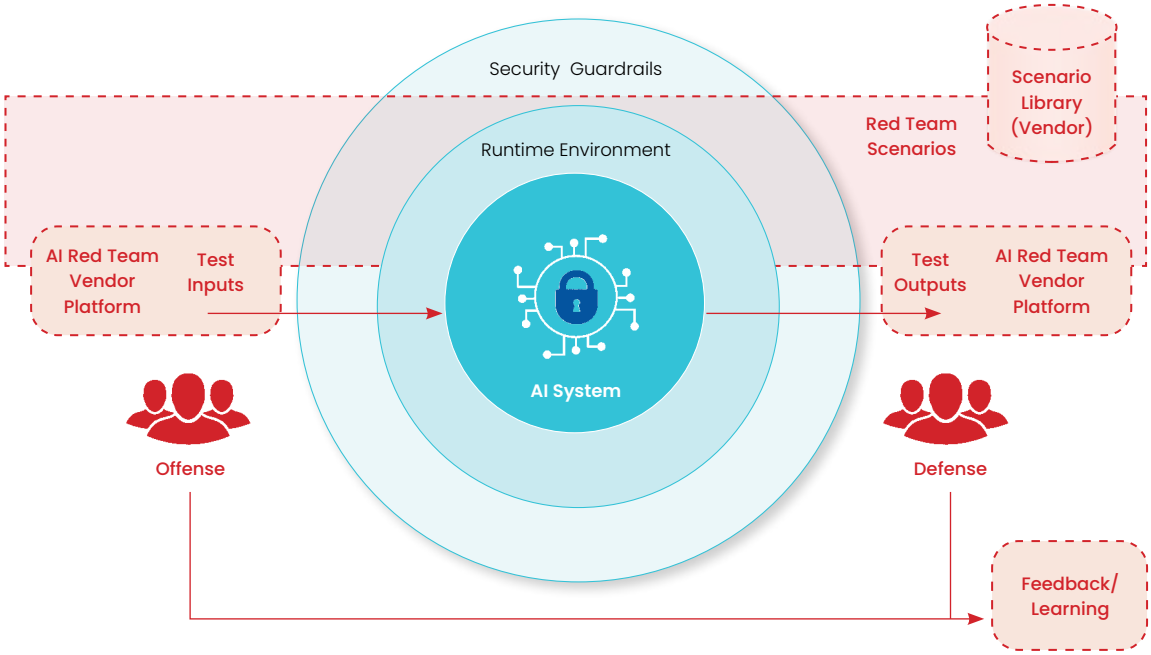


Figure 2-7. AI Security Red Team Focus Areas

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.

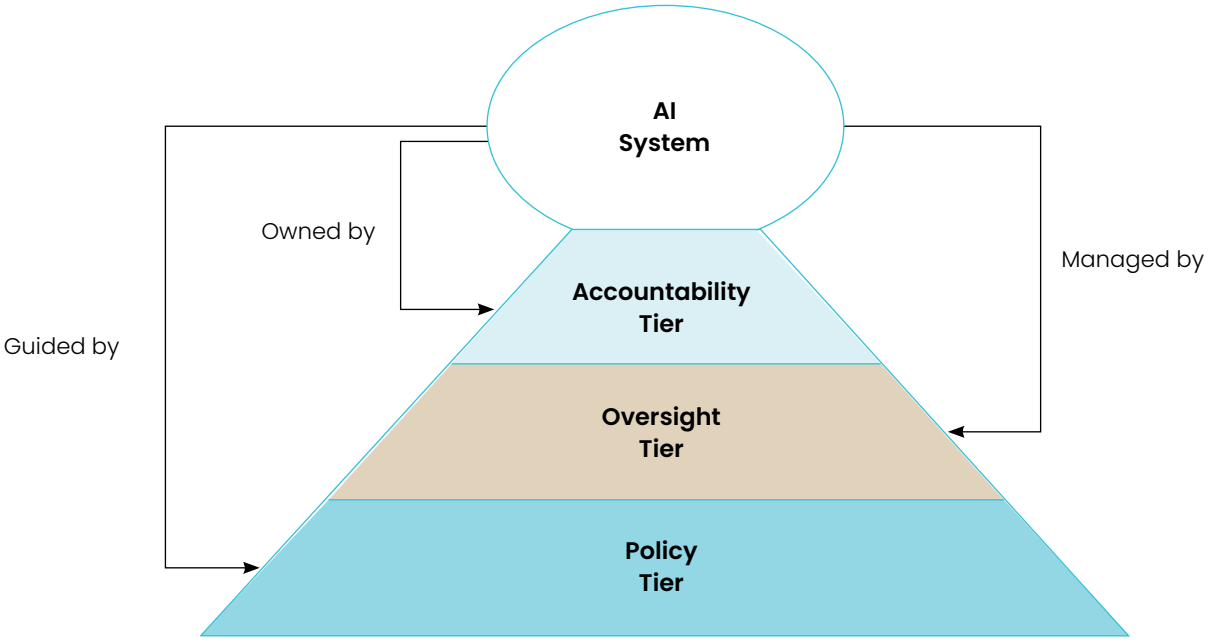


Figure 2-8. Governance Model for AI Security

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.

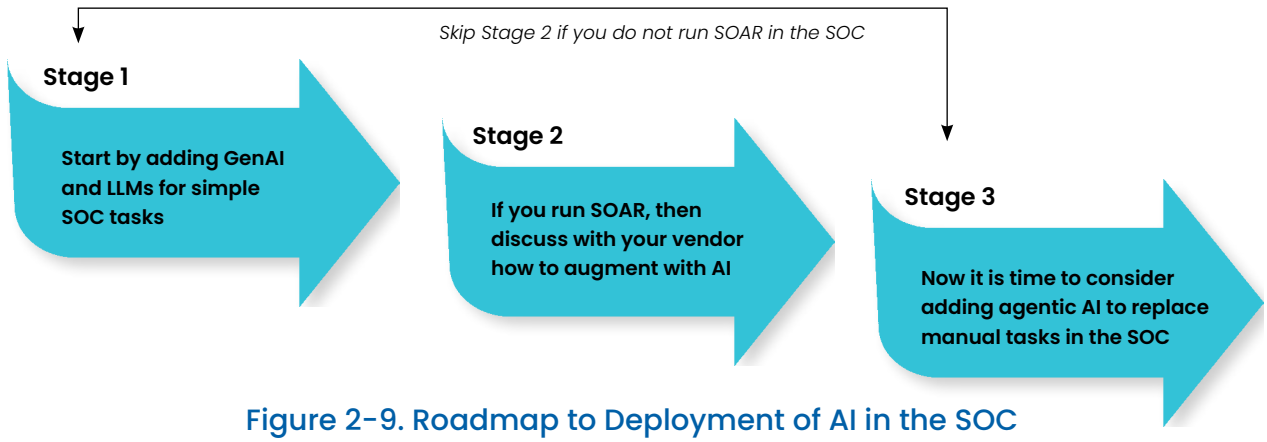


Figure 2-9. Roadmap to Deployment of AI in the SOC

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.

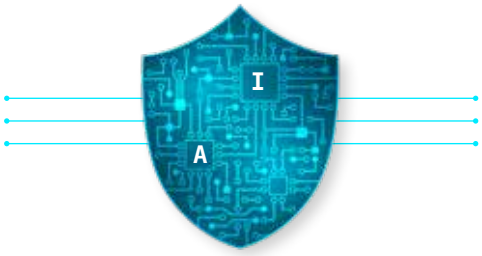
Beyond efficiency, AI-enabled SOC automation enhances resilience. During major incidents, when human analysts are saturated, AI agents can continue to process telemetry, detect anomalies, and preserve situational awareness. Over time, this automation layer becomes a force multiplier, allowing organizations 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.



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).

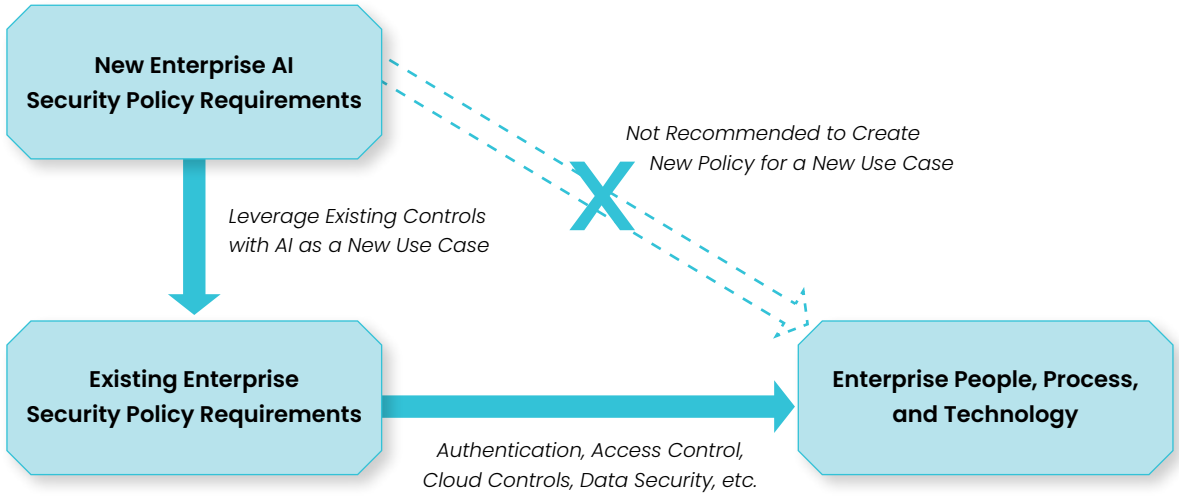


Figure 3-1. Recommended Security Policy Approach for AI

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.

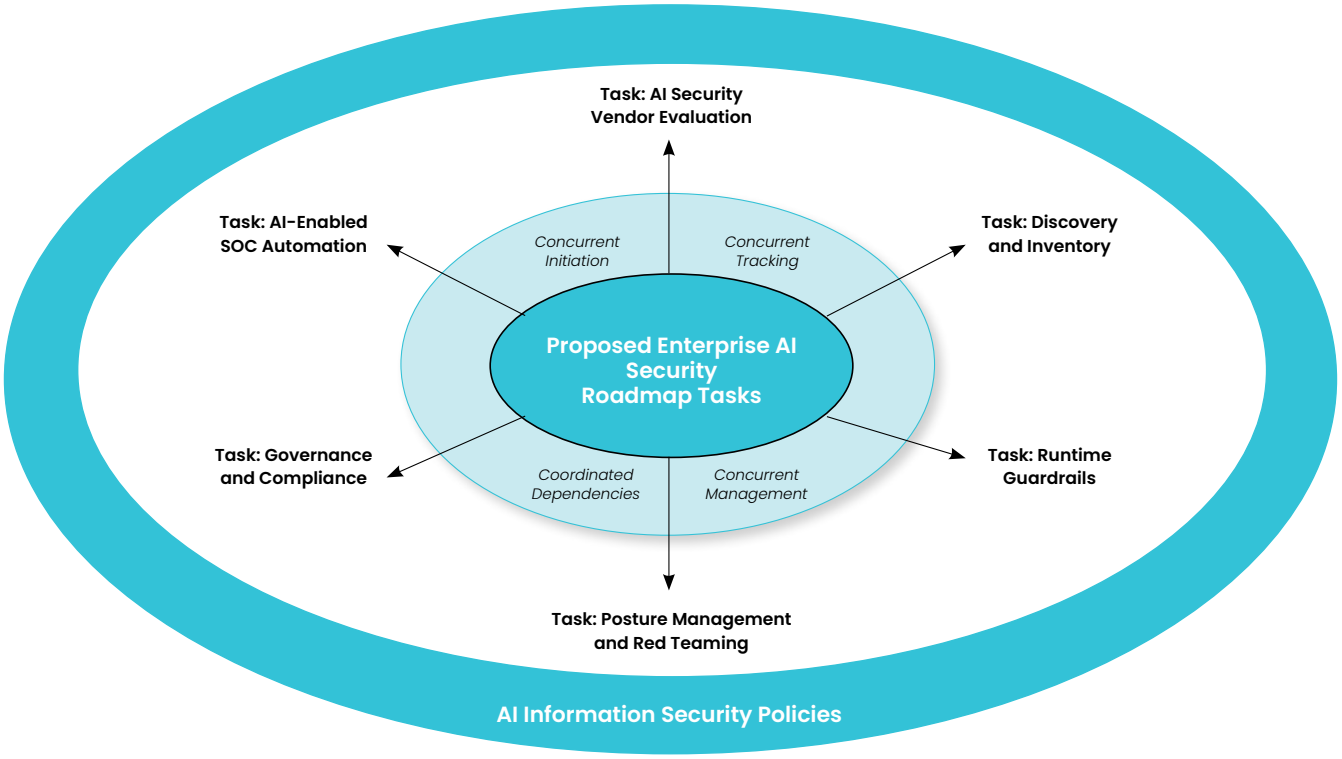


Figure 3-2. Foundational Support from AI Security Policies

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.

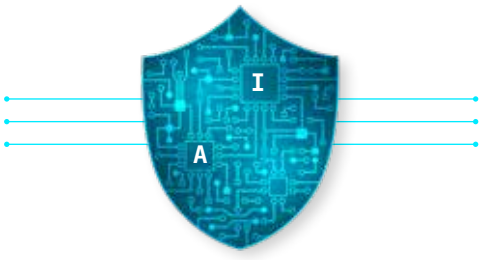
We should start by saying that most of the security frameworks such as MITRE ATT&CK, NIST CSF 2.0, and others are not well-suited, in our opinion, to developing an AI-ISP. MITRE ATT&CK, for instance, lays out a logical collection of attack methods, which we all find useful as a guide to offensive tactics. 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

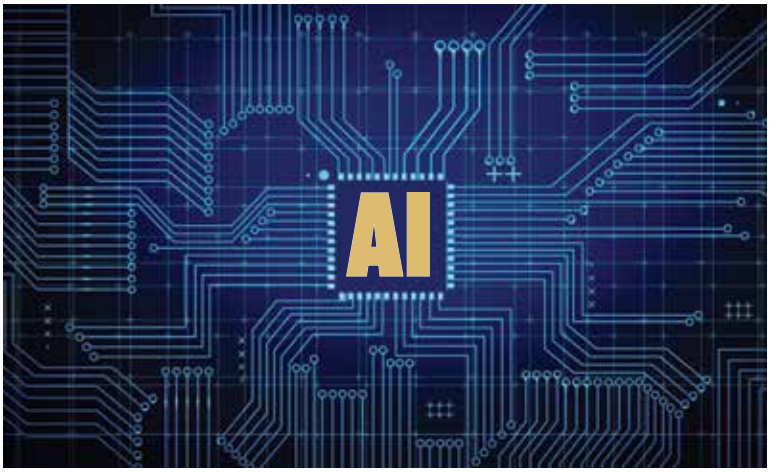
Figure 3-3. TAG Cybersecurity Taxonomy



TAG

ENTERPRISE AI SECURITY

HANDBOOK



THE TAG ANALYSTS

SPECIAL REPRINT EDITION

BLACKBIRD.AI

blackbird.ai

©2026, Tag Infosphere, Inc.