DARKTRACE

How to Secure AI in the Enterprise

A guide to securing models, data, and agents

© 2025 Darktrace Holdings Limited. All rights reserved.

Contents

Preface

We've heard from our customers and observed firsthand that security leaders lack a clear, concise definition of what "securing AI" encompasses. While there are valuable resources out there, we have yet to find one that cohesively aligns risk with security functions. Broader frameworks such as ISO/IEC 42001 provide important governance principles but do not offer the practical guidance required to operationalize AI security within an enterprise.

This paper aims to fill that gap. It provides a vendor-neutral guide to understanding the cybersecurity scope of AI, outlining the key categories, risks, and security functions that organizations should consider as they adopt and integrate AI technologies across their business.

About Darktrace

Darktrace has been at the forefront of applying artificial intelligence to cybersecurity for over a decade, pioneering innovations that have reshaped the industry. In 2013, we launched Network Detection and Response (NDR) powered entirely by AI-driven detection methods, setting a new standard for autonomous threat identification.

By 2019, we introduced an autonomous AI investigation agent that mirrors human investigation and analysis at machine speed and scale. Our early adoption of advanced language models, including transformer-based approaches in 2021 and Domain-Specific Language Models (DSLMs) in 2023, further strengthened our ability to interpret complex cyber threats.

Today, Darktrace AI operates continuously across approximately 10,000 global organizations, delivering proven resilience at scale. In mid-2025, we became one of the first cybersecurity companies certified to ISO 42001, underscoring our commitment to responsible AI governance. This reflects not only our leadership in AI innovation but also our unwavering dedication to securing AI itself ensuring trust, transparency, and security remain at the core of every advancement.

10+ Years of applying AI to cybersecurity use cases

One of the first cybersecurity companies to be ISO 42001 accredited

AI Excellence Awards winner for Cybersecurity AI (2025)

Approx 10,000 customers and counting

Securing AI

Understanding what's at risk

Artificial intelligence is entering businesses at a pace that few governance and management systems can match.

Traditional frameworks are struggling to keep up as AI introduces new complexities and expands the attack surface. Every model, dataset and agent, along with the complex web of APIs, third-party integrations, and human interactions that power them presents new risks and creates fresh pathways for compromise.

AI promises gains in productivity, speed, and efficiency for some, but it also brings its own set of uncertainties and a growing need for oversight particularly within the scope of Generative AI and GenAI-based agents. Generative AI, which can produce unreliable outputs or hallucinations, lack transparency in decision-making, and operate autonomously in ways that traditional security tools were never designed to monitor expands the digital ecosystem where ownership, control, and accountability often blur.

For security leaders, the first challenge in securing a business through transformative times is about identifying what needs to be secured. AI is entering organizations through multiple channels. Employees now use AI-enabled SaaS and productivity tools. Cloud and on-prem providers are extending their infrastructure and services with embedded AI capabilities. Vendors and acquired software often introduce AI features, expanding exposure without clear oversight. Internally, developers can build or adapt home-grown models and agents that integrate AI directly into the business.

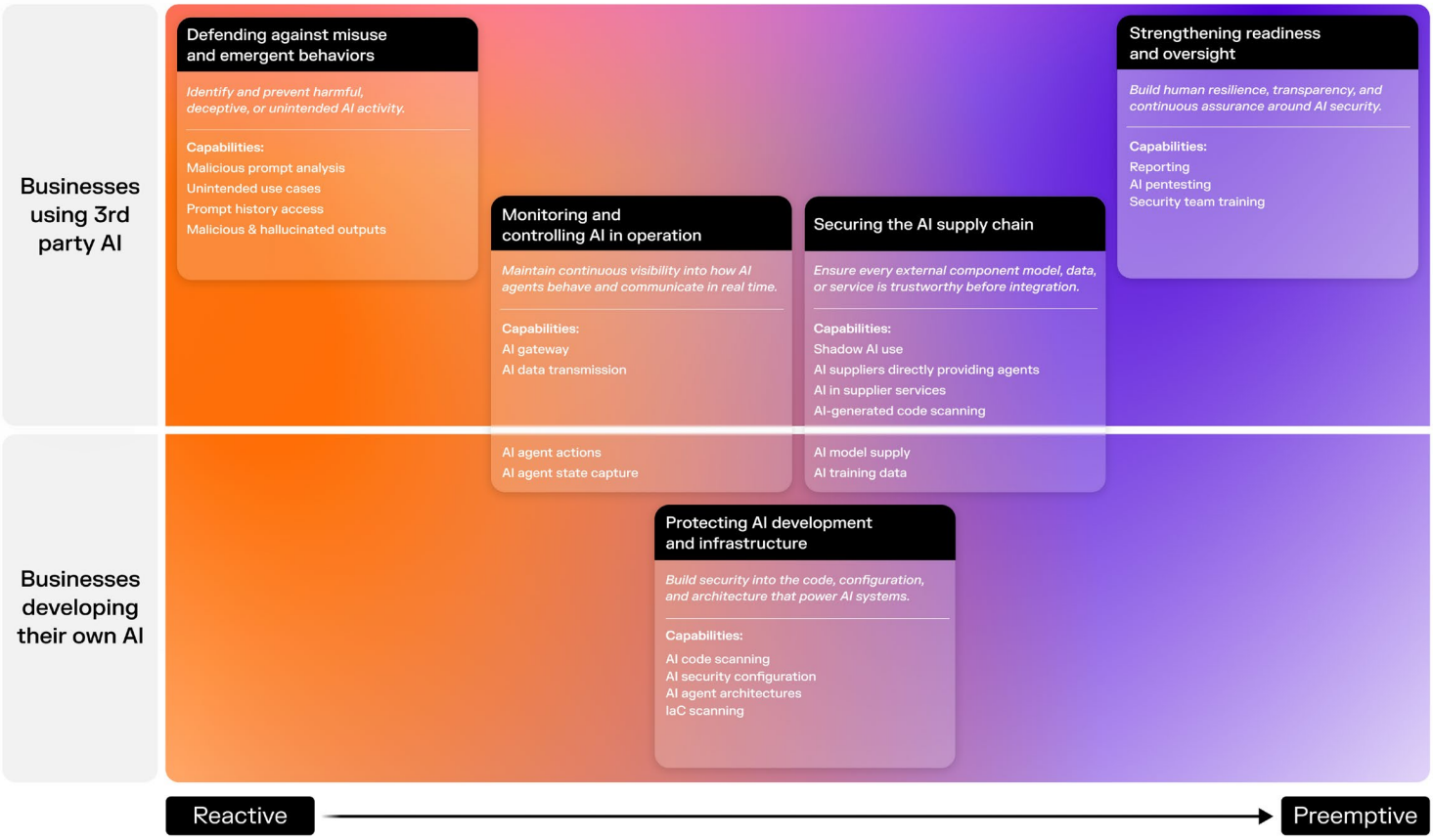
Across all these entry points, protecting an AI ecosystem requires visibility into every layer: from development and training to deployment and real-time operation. Each stage presents distinct risks, from data poisoning and dependency flaws to malicious agent behavior or unintended use of model outputs.

5 Categories of AI risk and how to address them

Each of the five categories described in this framework represents a distinct layer of AI security, from external suppliers and development pipelines to live operations and organizational governance.

Within each, specific capabilities define where risk can emerge, and where security functions can be applied to reduce exposure. The following section explores these capabilities in detail, pairing each with its associated risks and the security functions designed to address them. We've identified five critical categories of AI use. These categories mirror the ways AI enters and functions within an organization: through suppliers, development pipelines, operational systems, user interactions, and governance processes.

In summary the five categories are:



The table above provides a high-level summary of these areas, outlining how risks map to key security functions across the AI lifecycle. It offers a concise view of where defenses should be focused before the following pages explore each capability in greater depth.

Defending against misuse and emergent behaviors

Capabilities

Malicious prompt analysis

The detection and analysis of harmful or manipulative prompts that attempt to coerce AI systems or agents into unwanted or unsafe actions.

Risk description: AI agents can be told or tricked to take unwanted actions using their permissions and data access. This can be difficult to monitor or block, especially for external AI agents. Given enough access, a threat actor can explore and extract data from an AI model that was meant to remain hidden inside it.

Security function: Detect and where possible block prompts with malicious intent, whether internal or external, direct or indirect. Monitor for attempts at model data extraction or manipulation to prevent misuse and protect sensitive information within AI interactions.

Unintended use cases

The misuse or repurposing of AI systems in ways not originally intended or authorized, often resulting in unanticipated behaviors or risks.

Risk description: AI agents and systems that are misused can have a wide range of unwanted effects. The risks are heightened by how difficult it is to recognize misuse by authorized parties, if guardrails are not effective.

Security function: Identify and block unauthorized or non-approved uses of AI systems. Monitor for unintentional data exposure or misuse to ensure AI models operate only within approved and compliant parameters.

Prompt history access

The storage and management of AI prompt histories, which may contain sensitive data, contextual information, or behavioral patterns.

Risk description: Unapproved access to a prompt history can reveal confidential information, even when the service the prompts were submitted to is secure and does not retain the information.

Security function: Restrict access and monitor for compromise, since a threat actor with account control could exploit this information for further attacks.

Malicious and hallucinated output effects

The generation of AI outputs that are inaccurate, misleading, or intentionally harmful, potentially influencing human actions.

Risk description: Separate from risks relating to automation using hallucinated or malicious outputs (such as package-squatting and unexpected AI agent actions), these outputs can also cause humans to take unwanted actions. The scope of these potential actions is much wider than automated actions, for example if they are instructed to transfer money or to apply malicious settings to a critical IT system when these are deliberately not accessible to AI agents.

Security function: Monitor for harmful or fabricated AI outputs that could mislead users or trigger unintended actions. Implement hallucination detection and output validation to prevent misinformation or unsafe behavior influenced by AI-generated content.

Monitoring and controlling AI in operation

Capabilities

AI agent actions and connections

For businesses developing AI tools

The monitoring of AI agents' real-time behavior, including outbound connections, data exchanges, and interactions with other agents or services.

Risk description: AI agents may use their permissions improperly, or access data or actions that were not intended. They can also leak data or cause unwanted actions when calling other AI agents, or services (including MCP servers). AI agents have none of the caution of human staff, if guardrails do not anticipate possible problems or if they fail.

Security function: Monitor AI agents in real time for outbound connections, requests, and prompts to external services (including MCP servers), or other agents. Track credential use and enforce blocking where necessary to prevent data leaks or unauthorized actions.

AI gateway

A centralized access point or control layer that manages AI system interactions, including prompt submission, model routing, cost control, and data privacy enforcement.

Risk description: AI systems might not be accessed via gateways, or a gateway may be configured incorrectly or fail. This can result in lacking prompt security, incorrect model routing, uncontrolled costs or insufficient monitoring, and real-time control.

Security function: Deploy, centrally manage and enforce the use of AI gateways wherever possible. Include logs, alerts, and metrics in systems and security monitoring.

Where AI systems are deliberately deployed to keep data local (for example on-premises), security capabilities should follow and match.

AI agent state capture

For businesses developing AI tools

The process of preserving the operational state and information of potentially transient or ephemeral AI components, particularly in cloud environments, to enable post-incident investigation and analysis.

Risk description: If an ephemeral AI agent component (common in Cloud-hosted systems) does something that needs investigating or preserving, the component may vanish before security processes move to capture it. The security team might not have the capability to capture and preserve some components at all.

Security function: Automatically capture and preserve the full state of a suspect AI agent, including transient cloud instances. This enables investigation and response and ensures evidence and context are retained for analysis.

Highly regulated industries should pay special attention to this risk

AI data transmission

The transfer of data between AI components, systems, or external endpoints, including both automated and human-initiated exchanges.

Risk description: Data sent to AI systems may infringe data sovereignty or legal requirements (such as GDPR). Data can be sent by humans, by automated systems, or by AI agent decisions. Data transmissions to or by AI may use weak cryptography and need to be accounted for in PQC transition plans.

Security function: Secure data in transit between AI components using encryption methods, including post-quantum cryptography where applicable. Ensure all data transfers comply with relevant movement and privacy regulations such as GDPR.

Protecting AI development and infrastructure

Capabilities

AI security configuration

For businesses developing AI tools

The setup and ongoing management of security parameters, guardrails, and access controls that govern how AI components operate.

Risk description: Configurations given to AI agents and related and supporting individual components may be flawed and allow improper access or have weak security against attacks. Misconfiguration can also remove intended guardrails.

Security function: Ensure secure configurations across all AI components, including guardrails, policies, and permissions. Automated monitoring and analysis plus regular reviews of policies and enforcement help maintain a strong and compliant AI security posture.

AI agent architectures

For businesses developing AI tools

The design and structure of AI agents and their interconnected components, which determine how they interact, share data, and enforce security boundaries.

Risk description: The multi-component architecture of an AI agent in use, or AI agents that work together in a design, may be flawed as a whole and permit improper access or lack guardrails against unintended use, separate from issues with individual components.

Security function: Perform static analysis of AI agent architectures to identify security gaps. Assess localized network exposure, service credentials, and permissions to prevent unauthorized access or lateral movement. Proper segmentation and credential management are key to reducing architectural risk.

Development teams may make use of AI agent-building support services. Rationalize as far as possible to one service that offers acceptable security capabilities.

IaC scanning

For businesses developing AI tools

The automated analysis of Infrastructure-as-Code templates to detect misconfigurations or security weaknesses before deployment.

Risk description: When Infrastructure-as-Code is used, AI security configuration and AI agent architecture issues may be introduced due to improper access to the code, the introduction of malicious code internally or from an external source, or mistakes.

Security function: Use specialized analysis tools to scan Infrastructure-as-Code (IaC) for misconfigurations or embedded risks. Detecting security issues early in deployment scripts helps prevent vulnerabilities from propagating through AI environments.

AI-generated code scanning

For businesses developing AI tools

The review of code produced by AI systems (coding copilots) to detect security flaws, malicious inclusions, or unreliable dependencies.

Risk description: Generated code can have security flaws, whether accidental or due to a malicious effect of the generating model. This can include adding malicious dependencies, for example threat actors will create packages using names that are hallucinated. Generated code can be significantly harder to understand and debug than code written manually. Issues are more likely when the scope and complexity of the generated code increases.

Security function: Scan AI-generated code for flaws, insecure logic, or malicious inclusions. Detect issues such as hallucinated dependencies or package squatting to prevent compromised or unsafe code from entering production environments.

Securing the AI supply chain

Capabilities

Shadow AI use

The use of unapproved or unsanctioned AI tools by employees or departments, often outside official security and compliance controls. These services typically operate externally and may collect or train on sensitive business data without oversight.

Risk description: Data is leaked to unapproved external parties, which might include protected information, financial information, customer information, or confidential IP. Unapproved services are likely to use any submitted data for training.

Security function: Detect and block the use of non-approved AI services, such as unapproved generative AI tools or use of untrusted external MCP servers by approved tools. Unmonitored use can expose sensitive data and create blind spots in security oversight.

AI suppliers directly providing agents

External AI agents or GenAI services that are integrated into business systems and granted internal access or permissions to perform automated tasks on behalf of users or applications.

Risk description: The activities of these agents may not be visible or controllable once initiated, at either the prompt or connectivity level.

Security function: Monitor external AI agents granted internal access, as they often operate with broad privileges and limited visibility. Enforce strict access controls, logging, and oversight to prevent misuse or unintended actions within internal environments.

AI in supplier services

The use of external suppliers or third-party services that apply AI to process or analyze business data. This includes both declared AI offerings and situations where suppliers integrate AI into their products or workflows without explicit disclosure to customers.

Risk description: Suppliers providing a known AI service may not secure business data appropriately, either contractually or through operational failures. Suppliers may use AI internally and invisibly, avoiding oversight from the business giving them confidential data.

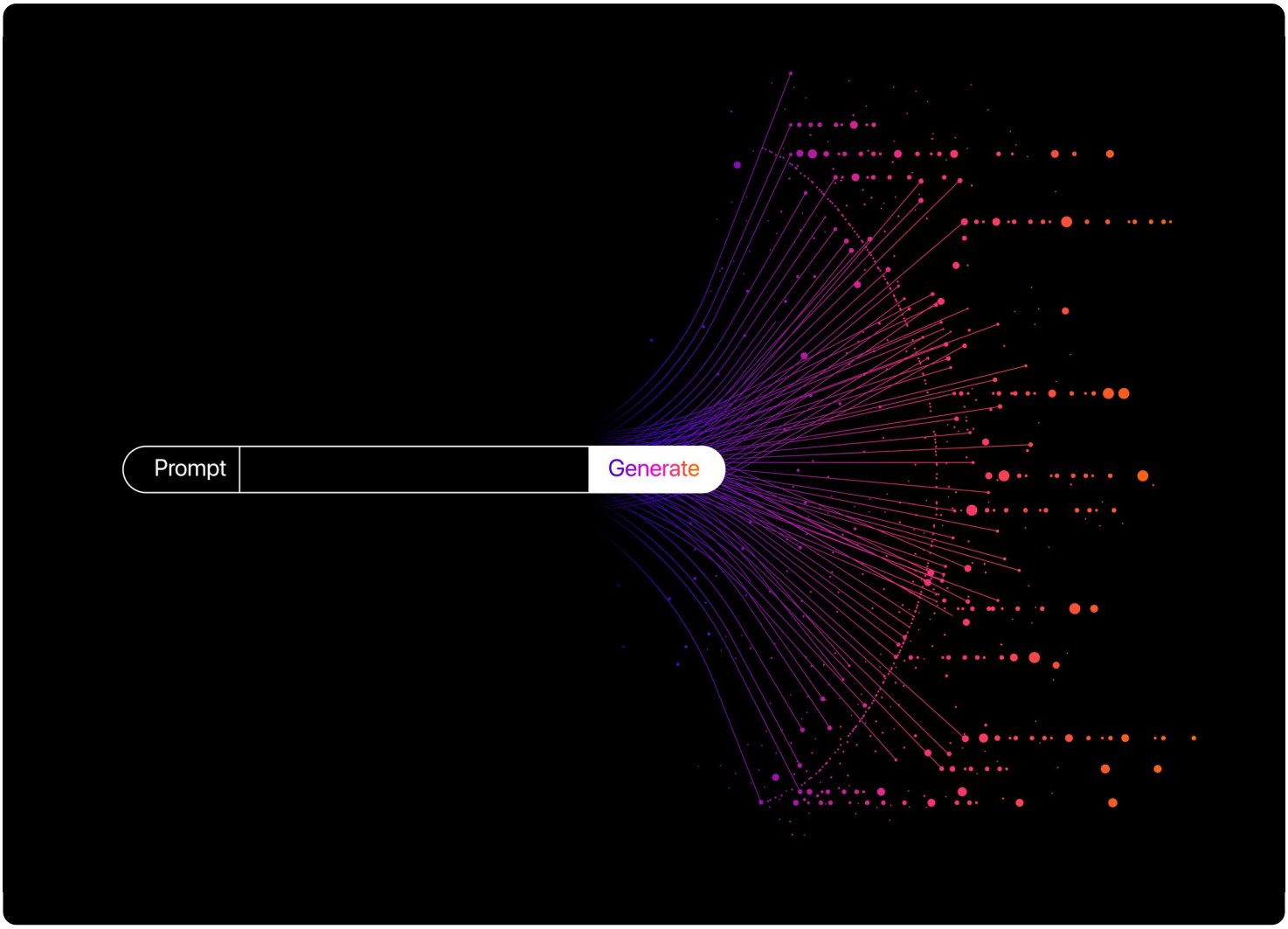
Security function: Assess suppliers that use AI on your business data (whether disclosed or hidden). Verify how external AI systems process, store, and protect that data to prevent unauthorized use or unintended exposure through third-party services.

AI agent code and dependency scanning

The process of examining AI agent source code and its supporting software dependencies to identify vulnerabilities, licensing issues, or malicious inclusions. This ensures that the components used to build or extend AI systems are secure, compliant, and functioning as intended.

Risk description: Software dependencies may include unsuitable licenses or have been compromised to be malicious whether open source or commercial. Code developed for AI agents may have security flaws reducing or voiding the intended security goals.

Security function: Conduct static security analysis of internal AI agent code to identify weaknesses and insecure configurations. Review all software dependencies to detect known vulnerabilities or malicious inclusions before deployment. Regular scanning helps maintain code integrity and reduces risk in the AI development pipeline.



AI model supply

For businesses developing AI tools

The process of sourcing or integrating external AI models, whether open source, third-party, or vendor-provided, into an organization's systems or development pipelines.

Risk description: Models brought in from external sources may include malicious or incorrect content and instructions creating outputs that cause AI agents or users to take unwanted actions. They could include levels of bias, accuracy, drift, and other quantities that are unacceptable to the business or their specific use cases. They might not have licenses suitable for their use within the business.

Security function: Securing the model supply chain means validating both origin and content. Each external model should be scanned for embedded risks, confirmed for proper licensing, and stored securely with version control.

AI training data

For businesses developing AI tools

The collection and preparation of datasets used to train or fine-tune AI models or in TEVV (testing, evaluation, validation and verification). These datasets may come from internal sources, public repositories, or third-party providers, and directly influence model behavior, accuracy, and compliance.

Risk description: Training data might include malicious data that causes a trained model to have unwanted effects. It could include unlawful protected data, data that is not licensed properly, or a dataset as a whole may not be licensed properly for its uses within the business.

Security function: Securing training data begins with knowing exactly where it comes from and what it contains. Verify data origin, licensing, and integrity before use. Protect against internal or external data poisoning. Recognize and manage risks around data licensing, especially in foundational models, and the potential for training data to be reproduced directly in model outputs.

Strengthening readiness and oversight

Capabilities

Reporting

The process of capturing and communicating AI-related risks, metrics, and trends to support informed decision-making and governance.

Risk description: Objectives and metrics are vital to effective management and continuous improvement. Critical issues or trends may be missed without effective AI management reporting. The wider business may make unwanted decisions about AI without effective reporting to their interests.

Security function: Generate clear summaries of AI-related security events, risk metrics, and system performance. Track adoption of intended AI use cases and flag any unintended or unauthorized uses to maintain oversight and governance.

AI pentesting, red and purple teaming

The testing and validation of live AI systems through simulated attacks and adversarial exercises to identify vulnerabilities and detection gaps.

Risk description: Live systems often have latent issues that are not detectable in configuration or code. Without penetration tests, threat actors may discover specific exploits that security teams are unaware of. Without red or purple team testing, wider weaknesses in exposure and detection capabilities may be present but unknown.

Security function: Conduct penetration tests and red or purple team exercises to simulate real-world attacks against live AI systems. Test for exploitable weaknesses, especially in external-facing environments, to anticipate and mitigate adversary techniques before they occur.

Security team training and exercises

Ongoing education and simulation exercises designed to build AI-specific detection, investigation, and incident response skills within security teams.

Risk description: Insufficient individual training or team incident response practice can result in missed detections, poor investigations, or an inability to stop or remove a cybersecurity issue or threat actor.

Security function: Develop AI-specific skills and readiness within security teams. Regular training and simulations strengthen competence in detecting, responding to, and recovering from AI-related incidents.

Reframing AI security for the boardroom

The same intelligence that accelerates business decisions can, if left unsecure, amplify risk at unprecedented speed and scale.

For security leaders, the mission is not only to safeguard AI systems, but to translate their complexity into clear, strategic guidance for executive leadership.

At the board level, this can begin with framing AI security in terms of trust, accountability, and resilience:

Trust

Boards want assurance that AI-driven decisions are accurate, interpretable, privacy-preserving, and free from tampering. CISOs should highlight the controls that verify data integrity, model behavior, and the provenance of every component in the AI supply chain.

Accountability

Clarify who owns each layer of AI risk: data scientists, cloud providers, vendors, and security teams each have defined responsibilities. Boards respond to structure and shared accountability, not ambiguity.

Reinforce accountability by mapping AI security principles to established governance frameworks such as NIST AI Risk Management Framework (RMF), ISO/IEC 42001, and guidance from CISA and NCSC. This linkage demonstrates maturity, ensures consistency across compliance obligations, and provides a clear benchmark for oversight.

Resilience

Emphasize that AI is now part of the organization's critical infrastructure. Securing it means ensuring continuity under attack, auditability under scrutiny, and agility under regulatory change.

As AI becomes embedded across business operations, linking its security to outcomes like intellectual property protection, operational reliability, and stakeholder trust will be essential to enterprise resilience. Organizations with management systems in place that actively enable while still securing the expansion of AI use cases will be best positioned to innovate securely, adapt quickly, and maintain the confidence of customers, regulators, and investors.

Conclusion

The same forces that make AI transformative also make it difficult to secure when woven through SaaS tools, cloud infrastructure, supplier systems, and homegrown applications. In this environment, visibility, ownership, and accountability can quickly blur.

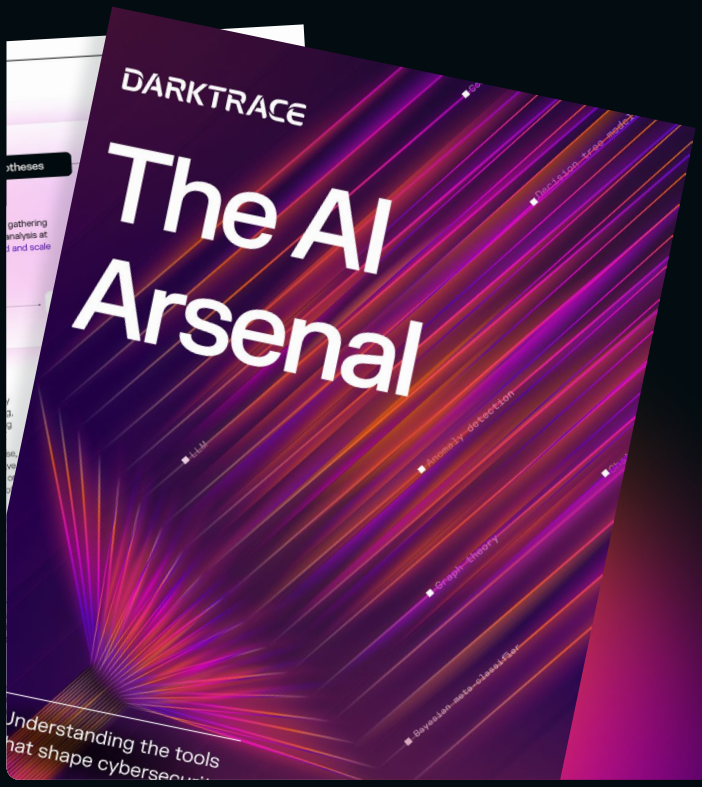
Securing AI means restoring that clarity. It requires understanding not only how AI is built, but where it operates, who interacts with it, and how its decisions shape the business. By viewing AI security risks through this broader lens spanning supply chains, development, real-time operation, and governance organizations can begin to turn uncertainty into assurance.

The journey to secure AI is ongoing, but it begins with awareness: knowing where AI exists, how it behaves, and how to govern it responsibly.

Building AI securely means building it responsibly.

Explore how to align innovation with governance in our paper, Towards Responsible AI in Cybersecurity

Learn more ↗



Discover the spectrum of AI types in cybersecurity

Understand the tools behind modern resilience, from supervised machine learning to NLP, and how they work together to stop emerging threats.

Learn more ↗

About Darktrace

Darktrace is a global leader in AI cybersecurity that keeps organizations ahead of the changing threat landscape every day. Founded in 2013 in Cambridge, UK, Darktrace provides the essential cybersecurity platform to protect organizations from unknown threats using AI that learns from each business in real-time. Darktrace's platform and services are supported by 2,700+ employees who protect nearly 10,000 customers globally. To learn more, visit www.darktrace.com.

North America: +1 (415) 229 9100 | Europe: +44 (0) 1223 394 100 | Asia-Pacific: +65 6804 5010 | Latin America: +55 11 4949 7696

darktrace.com | info@darktrace.com

© 2025 Darktrace Holdings Limited. All rights reserved.