DARKTRACE

CISO's Guide To Buying AI

Content

Introduction

AI has rapidly become embedded in security products and will continue to play a large role in how these products detect, analyze, and respond to threats. Saving time and resources, mining intelligence and insights from valuable data, and performing deeper and more complex analysis than a traditional security solution, AI enhances an organization's capabilities and security posture.

With this acceleration in AI adoption, accompanied by the recent boom in agentic AI and autonomous agents, CISOs must look beneath the surface of these capabilities to understand how different models are built, trained, and validated, and how they behave when confronted with new or unfamiliar data.

Vendors use a wide range of approaches, and each design choice influences accuracy, reliability, and the likelihood of introducing blind spots. Gaining clarity on how an AI feature works, and whether it is truly suited to the task it claims to solve, allows security leaders to ask more precise questions and ensure the technology is being applied in the right way without adding unnecessary risk.

Equally critical is understanding whether the AI is deployed safely and governed responsibly. Poor visibility, weak controls, and inconsistent maintenance can lead to performance degradation, model drift, bias, and exploitable vulnerabilities.

Well-designed solutions include strong safeguards, transparent reasoning paths, and clear controls around data handling and protection. CISOs should feel confident challenging vendors on their governance frameworks, testing and validation workflows, resilience against adversarial inputs, and safeguards that prevent misuse or data leakage. This level of due diligence is especially important given the recent surge in AI investment, development, and vendor claims.

While AI can unlock meaningful improvements in productivity, insight generation, and security efficacy, the market is also shaped by hype, misconceptions, and a lack of transparency.

This guide is designed to help CISOs cut through that noise, distinguishing proven capabilities from buzzwords, so they can confidently evaluate AI-enabled security tools and ensure any adoption is reliable, responsible, and aligned with their organization's risk and governance requirements.

What to consider when making an AI investment?

An AI Service Provider (AISP) is a vendor that builds and delivers AI-as-a-Service, handling everything from model choice, design, and training, to refinement, validation, and ongoing governance. An AISP's solutions can be formed of either a single model, or through a combination of models which are arranged together to produce a more complex and encompassing output.

Almost all security vendors are beginning to use AI as part of their solutions. In the context of cybersecurity, AI can be an extremely valuable. While traditional security systems rely on signature-based detection, AI systems are able to identify patterns and trends, generalizing to identify novel attacks.

What to know before you evaluate an AISP

AISPs can use multiple different models within a single solution

AI covers a wide range of technologies (including, but not limited to, generative AI)

AI isn't always correct, and needs to be optimised

All AI systems introduce their own risks, requiring governance

AI is constantly evolving, making it difficult to maintain a clear definition of "what is an AI system". The Organization for Economic Co-operation and Development (OECD) defines AI as a machine-based system which can infer outputs from inputs with the need for explicit programming.

The field of AI is broad, with a wide range of techniques falling under the AI umbrella, including clustering, classification, regression, computer vision, large language models (LLMs), domain-specific language models (DSLMs), and agentic AI.

LLMs and DSLMs

One field of AI which has seen a recent boom in popularity is large language models (LLMs). LLMs are general-purpose models designed to handle free-text input and are well suited to natural language tasks.

Similarly, **domain-specific language models (DSLMs)** are LLMs which have been designed for a specific task, or to operate in a specific domain, such as Darktrace's DEMIST-2 Language Model which has been specifically trained for security related tasks.

Generative AI

Generative AI refers to an AI system which can produce original output – including natural language text, images, and audio. These systems have become increasingly popular with the rise of cloud-based AI chatbot systems which are backed by LLMs. generative AI doesn't exclusively refer to LLMs – there are many other generative AI systems such as image and video generation models (like OpenAI's Dalle-2, or Google's Veo) which are built on the same technology as LLMs, along with other generative statistical techniques.

It's also important to note that LLMs don't always have to be used generatively. There are a wide range of other LLM-based approaches which can be better suited than generative AI for some tasks.

Agentic AI

Some AISPs more recently have put a large focus on producing agentic AI solutions. Agentic is a buzzword currently being used a lot in the AI industry, with a lot of definitions being passed around and many AISPs claiming their systems as agentic.

The definition of agentic AI varies heavily from provider to provider, ranging from any autonomous system, to any generative AI system capable of using tools, to even just a marketing term which can be applied to any AI system.

With this variation, it's necessary to ask any potential AISPs what they define as agentic AI, how agentic AI is implemented into their solution, and consider what risks their implementation may introduce.

Composite AI

Another key technique AISPs may use is composite AI – a combination of multiple different AI techniques into a single solution. This allows multiple systems to support each other, reducing blind spots, and extending the scope of the solution.

Understanding the limititations of AI in cybersecurity

While introducing AI systems into a solution can be beneficial, providing unique insights and efficiency boosts through autonomous actions, it's important to remember that all AI systems have their own disadvantages and potential risks.

For instance, while generative AI systems are increasingly popular due to their flexibility and ability to process and output unstructured natural language data, they perform better at some tasks than others.

When considering an investment in an AI solution, buyers should assess what advantages AI offers over other approaches and evaluate the variety of AI methodologies available, along with the unique benefits each one provides. In this paper, we'll cover some of the key questions you might want to ask your AISP to understand their approach to AI development, implementation, and governance.

Finally, while the adoption of AI systems is rapidly accelerating, much of the governance and risk management around them is still in its infancy. Additionally, the requirements around governance and risk management only increase as systems become more complex. Understanding the controls vendors have in place here can be a crucial component in differentiating between impressive technical demos and systems capable of trustworthy, robust enterprise use.

For generative AI in particular, Gartner suggests that it is:

Highly useful for content generation, conversational user interfaces, knowledge discovery

Somewhat useful for segmentation/ classification, recommendation systems, perception, intelligent automation, anomaly detection/monitoring, autonomous systems

Hardly useful for prediction/forecasting, planning, decision intelligence

In addition, some Gartner research has reached the conclusion that:

To achieve more accurate and robust AI solutions, AI leaders should move beyond using just one model or technique, embrace composite AI practices and adopt a holistic AI system perspective.

How to evaluate AI vendors

In initial research of AISPs, there are a few places buyers can look to narrow down their pool of candidates. This can be done by looking for proof points and customer testimonials.

These will provide evidence that the AI tool works, and that it can also work for your organization. Especially look for insight from customers in similar situations as your organization, whether that be industry, organization size, region, or other attributes.

Also, when evaluating an AISP it would be valuable to assess the provider for a culture of good governance. This means understanding how governance is embedded into how the organization builds, tests, and deploys AI.

Research from our Darktrace State of AI Cybersecurity report found that only 37% of organizations have an AI governance policy.

It's important to recognize that even if a company does have an AI governance policy or claims to align with a known AI governance standard, the important questions to ask are how that company implements those policies and processes. The stronger culture and adoption of AI governance a company has, the greater likelihood this will filter to their products.

Buyers should ask questions to potential AISPs about the governance and data security processes they implement, the technologies they use, and the workflows they follow for testing, evaluation, validation, and verification. By asking these questions, a buyer can gather more information to confirm that a solution meets their business requirements and policies.

In the following sections we outline questions buyers can ask around 5 key categories:

01 Governance, safety, and data controls

02 Data gathering and training

03 Model and technique choice

04 Performance and accuracy validation

05 Interpretability, adjustability, and transparency

Governance, safety, and data controls

AI governance is essential for ensuring that all AI systems and models within a solution are secure, accurate, and well maintained, as well as verifying that all data being passed into a system is used in a responsible, ethical, and legal manner. A strong governance system grants sufficient visibility over all systems internally, allowing an organization to ensure data safety, security, and integrity.

When it comes to more complex, composite AI systems featuring a large number of models, governance is critical. For solutions which feature multiple AI systems, it would be difficult to ensure that all models are kept up-to-date, secure, and are consistently refined for accuracy improvements without a sufficiently robust governance solution to track and document all the systems.

Inconsistent maintenance and lack of visibility over systems internally could then lead to vulnerabilities, weaknesses, and biases being introduced which would negatively impact performance and accuracy of the solution.

The following are a few examples of questions a buyer should ask potential AI Service Providers when gathering information about their governance policies and processes.

QUESTIONS FOR VENDORS

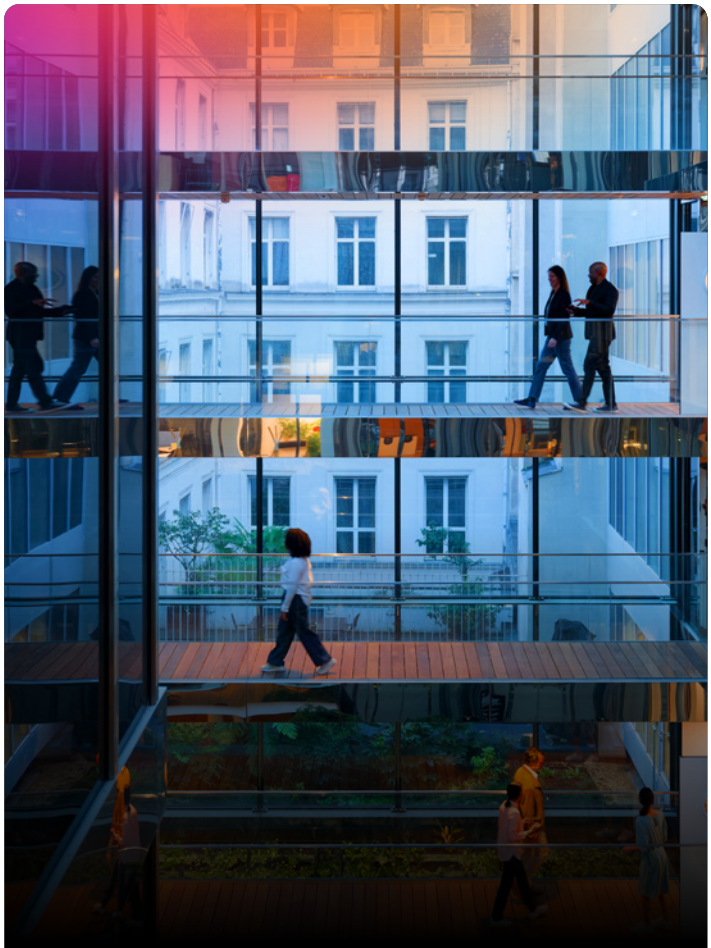
What AI governance policies and frameworks do you follow, and/or certifications do you currently maintain?

How do you ensure data security, privacy, and ring fencing of client information (IP, PII, etc...) within your AI solutions?

How do you protect your AI systems from malicious inputs, data poisoning, and other adversarial attacks?

Do you utilize any externally hosted 3rd party models or platforms to process or handle client or company data? And if so, how do you ensure these systems are safe, secure, and confidential?

When reviewing answers to questions about governance and data security, buyers should look for references to specific certifications, workflows, or frameworks their AISP uses to ensure strong governance.



Darktrace is certified to the ISO/IEC 42001 standard. ISO/IEC 42001 is the world's first AI management system (AIMS) standard, providing valuable guidance on the unique challenges AI poses. For organizations, it sets out a structured way to manage risks and opportunities associated with AI, supporting strong and enduring governance of all AI systems.

It's important to note that ISO/IEC 42001 isn't the only AI Governance framework available, with other frameworks like the NIST AI Risk Management Framework also being widely used. ISO/IEC 42001, however, provides an additional level of assurance through external auditing, rather than relying on self-claimed adherence.

In addition to AI governance policies, data security and privacy are critical for all systems, so identification of data security frameworks and certifications can also provide buyers with confidence that a system is sufficiently secure and is able to handle their data safely.

Data gathering and training

One of the most important steps of producing an AI system is the initial gathering of training data. If an AI system is trained on poor quality, inconsistent, inaccurate, or biased data, the resultant system will never perform optimally.

ISO/IEC 42001 and other governance frameworks provide controls which should be followed when gathering data for AI systems. This includes source identification and audit logs to ensure that training data is traceable and hasn't been tampered, performing sufficient de-identification or abstraction on collected user data based on the risk-context, and thoroughly checking for bias where appropriate.

The following is a list of questions which buyers can reference as a starting point for questions regarding data gathering.

QUESTIONS FOR VENDORS

What steps do you take to prevent bias in your AI models and training data?

How do you minimize the risks associated with training data (e.g. Regurgitation, data provenance)?

When reviewing responses, buyers should assess whether the provider has an AI Management System in place that governs training data sourcing and handling, and whether clear processes exist to identify and mitigate bias or systemic issues.

In cases where PII is present in AI training data (outside of organization-specific self-learning systems), AISPs should be able to reference specific controls and processes used to de-identify the data (remove PII) where appropriate, or any other anonymization processes.

Model and technique choice

Different AI models and techniques are suited to different tasks. Selecting the right model for a particular task is an important step in ensuring that a solution is accurate and efficient.

Models and techniques also require strong governance, and it's not just about compliance with standards and certification; it's also about adopting the most effective and efficient models and processes to achieve the intended objective. This involves keeping detailed logs of which models and techniques are being used and ensuring that these are continually analyzed for risks and vulnerabilities, including supply chain risks of 3rd party models and internal models.

To ensure AI Service Providers are following appropriate governance procedures when selecting both internally hosted and externally hosted models and services, here are some questions buyers should ask:

QUESTIONS FOR VENDORS

What type(s) of AI model do you utilize in your solution?

Do you use any generative AI (internally or externally hosted), and if so, what safeguarding and guardrails do you have in place for these?

How do you ensure your choice of models is optimal for each required task?

Do you use agentic AI? If so, how and where is this used?

Do you use foundation or general-purpose models? If so, how are these hosted?

While specific detailed information about custom systems used by AISPs is likely proprietary, buyers should expect vendors to be able to provide an overview of the broad techniques used. Transparency around the use of third-party or cloud-hosted foundation models, associated APIs, and relevant data residency remains an important consideration for buyers seeking assurance regarding data handling and integrity.

References should be made by AISPs to due diligence and thorough investigation and analysis of 3rd parties before implementing their external systems to ensure safety and security.

Performance and accuracy validation

AI systems should undergo detailed Testing, Evaluation, Verification, and Validation (TEVV) checks prior to general use. Evaluation of a system, followed by optimizations, adjustments, and refinements, is essential to ensuring that the systems being produced are calibrated appropriately, and securely, to deliver meaningful insights for the intended task.

There's more to TEVV than just accuracy validation. Many techniques need to be applied to confirm that a system is free from bias, doesn't over or underfit data, and performs efficiently as well as accurately. Continual evaluation after release is also critical to prevent unwanted data and model drift. Testing should also cover the physical performance of the model, ensuring the model can produce outputs in an appropriate timeframe, and that it operates with reasonable resource consumption.

ISO/IEC 42001 requires the creation of a workflow to assess a system's relevance and accuracy which must be completed for every AI system before it can be used in production.

Below is a list of questions buyers could ask their AI Service Provider to ensure they are conducting appropriate TEVV and that there is sufficient governance when it comes to performance and accuracy validation.

QUESTIONS FOR VENDORS

How do you audit, test, evaluate, verify, and validate your AI model outputs?

How do your models adapt to an ever-changing external threat landscape, and how do you reduce the risk of data drift?

Does your TEVV workflow include tests for physical performance such as latency, memory consumption, energy and resource requirements?

Answers to these questions should feature direct references to a TEVV workflow, whether part of an existing AIMS (or other governance system) or a separate framework.

Interpretability, adjustability, and transparency

An important part of any AI system is ensuring the data output is interpretable; the steps taken to find that output are understandable, and that end-users can trust the AI outputs are correct. While an AI system is often a black box, a user should still be able to understand why a system may have come to a given conclusion, even if the exact steps are hidden.

In the world of cybersecurity, trust in an AI system is essential. These systems need to perform real-time actions, autonomously acting to block connections, reject messages, and restrict network access. If the system can't be trusted to perform these actions accurately and correctly, end-users may disable the features or be inclined to revert actions, critically reducing the defensive capabilities of the system. The high trust requirements in cybersecurity make many of its use cases poorly suited for generative AI automation. Adjustability of AI systems is an important function.

Every organization has a different set of behaviors, structures, and cultures. The ability to fine-tune AI models will help achieve overall objectives and allow a solution to fit with a wider range of use cases.

The following questions are a starting point for questions buyers could ask their AISP to confirm their system will be sufficiently interpretable, adjustable, and promote a strong trust relationship between human analysts and AI tools.

QUESTIONS FOR VENDORS

How do you ensure your AI systems are interpretable?

How do you promote a trust relationship between human analysts and AI outputs?

How can a SOC team fine-tune your models to suit their unique policies and business requirements?

An AISP should be able to describe how their solution can be adapted to fit an organization's unique structure, policies, and needs whether through self-learning and autonomous adaptation or through manual adjustments available to end-users.

Also valuable is confirmation that transparency and interpretability are considered. An AISP should provide information about the policies, workflows, and methods with which they ensure complete transparency and promote trust in their solution.

Darktrace as an AI cybersecurity vendor

Darktrace has been building and applying AI in cybersecurity for over a decade, developing its capabilities alongside an increasingly complex and fast moving threat landscape.

This experience has resulted in a mature, multi-layered approach to AI, where Darktrace's Self-Learning AI continuously learns the normal patterns of each organization to understand behavior, interpret context, and identify meaningful deviations — without relying on predefined rules or known attack signatures. Over time, this has enabled a proven behavioral understanding that helps uncover subtle signals of risk that may otherwise be missed.

This approach has been developed and validated across dynamic, cross-domain environments, where activity spans networks, identities, cloud, email, and applications. By understanding behavior as it moves across systems, Darktrace provides a continuous, organization-wide view of activity, enabling detection, investigation, and response that remain adaptive as environments and threats evolve. This reflects a long-standing commitment to innovation, built on real-world application and refinement, alongside a focus on responsible and transparent AI development.

As pioneers in space, we have also ensured responsibility is a core tenant of our process. This commitment is reflected in our certification in ISO/IEC 42001, demonstrating Darktrace's ability to deliver industry-leading Self-Learning AI in the name of cybersecurity resilience.

Our stakeholders, customers and partners can be confident that Darktrace is responsibly, ethically, and safely developing its AI systems, and is managing the use of AI in our day-to-day operations in a compliant, secure, and ethical manner.

How Darktrace secures AI systems

Darktrace now brings these capabilities to monitor and respond to risk generated from AI systems across organizations with Darktrace / SECURE AI. This solution analyzes how prompts, agents, and systems are used within the context of each organization, bringing every AI interaction into a single view. This unique approach helps teams understand intent, assess risk, protect sensitive data, and enforce policy across both human and AI agent activity. It provides real-time visibility and control across generative AI, AI agents, development environments, and shadow AI, allowing teams to identify misuse, misconfiguration, and drift that rule-based approaches often miss.

By understanding context and intent across users and systems, it helps organizations maintain oversight as AI adoption expands across the enterprise.

Sign up for the Secure AI Readiness Program here: This gives you exclusive access to the latest news on the latest AI threats, updates on emerging approaches shaping AI security, and insights into the latest innovations, including Darktrace's ongoing work in this area.

Ready to talk with a Darktrace expert on securing AI? Register here to receive practical guidance on the AI risks that matter most to your business, paired with clarity on where to focus first across governance, visibility, risk reduction, and long-term readiness.

Darktrace's approach towards responsible AI in cybersecurity

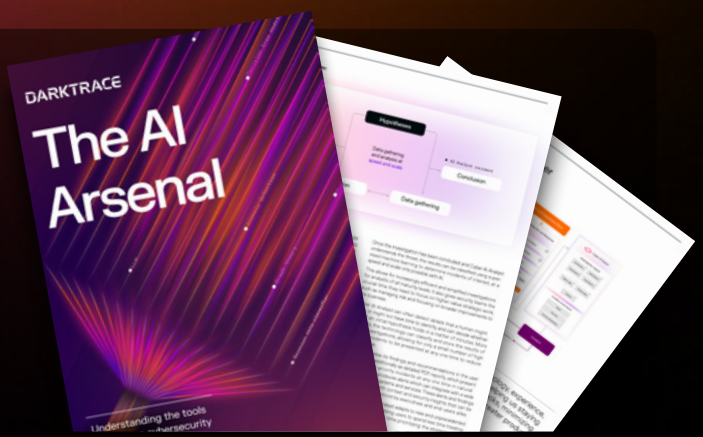
Explore the principles behind Darktrace's responsible AI approach, informed by collaboration with global experts in academia and governments, detailing how accountability, explainability, and continuous validation are built into its cybersecurity technology.

Read more here

AI Arsenal

Gain deeper knowledge on how supervised, and unsupervised, machine learning and LLMs can be applied to cybersecurity, and how Darktrace combines specific techniques into a multi-layered architecture to augment human security teams in the AI Arsenal White Paper

Read more here



Conclusion

As AI becomes further embedded in cybersecurity, the most important factor for security leaders is not whether a product uses AI, but how that AI is designed, governed, validated, and controlled. Buyers should expect clear answers around data sourcing and protection, model selection, testing and validation practices, resilience to adversarial inputs, and the transparency of AI-driven decisions.

Strong AI solutions are built using combinations of carefully chosen techniques, continuous evaluation, and governance frameworks that reduce risk over time, ensuring systems remain accurate, adaptable, and aligned with organizational policies as threats evolve.

When implemented responsibly, AI delivers meaningful advantages: improved detection of novel threats, faster and more precise investigations, reduced operational burden on security teams, and greater confidence in autonomous responses. Darktrace's multi-layered, self-learning approach demonstrates how diverse AI techniques, strong governance, and explainable outcomes can work together to deliver trusted cybersecurity at scale.

Contact Darktrace for a personalized demo

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