

DARKTRACE

Leitfaden für CISOs zum Kauf von KI

Erst prüfen, dann investieren:
So starten Sie Ihre KI-Einführung



Secure
What AI
Can Do

Inhalt

02	Einführung
03	Was sollten Sie bei einer KI-Investition beachten?
03	Was Sie wissen sollten, bevor Sie einen AISP evaluieren
04	Die Grenzen von KI in der Cybersicherheit verstehen
05	So bewerten Sie KI-Anbieter
06	Governance-, Sicherheits- und Datenkontrollen
07	Datenerfassung und Training
07	Modell- und Methodenauswahl
08	Leistungs- und Genauigkeitsvalidierung
08	Interpretierbarkeit, Anpassbarkeit und Transparenz
09	Darktrace als Anbieter von KI-gestützter Cybersicherheit
10	Fazit

Einführung

KI ist rasch zu einem festen Bestandteil von Sicherheitsprodukten geworden und wird weiterhin eine wichtige Rolle dabei spielen, wie diese Produkte Bedrohungen erkennen, analysieren und darauf reagieren. KI spart Zeit und Ressourcen, gewinnt Erkenntnisse und Einblicke aus wertvollen Daten und führt tiefgreifendere und komplexere Analysen durch als herkömmliche Sicherheitslösungen. Dadurch stärkt sie die Fähigkeiten und die Sicherheitslage eines Unternehmens.

Angesichts dieser beschleunigten KI-Einführung, die mit dem jüngsten Boom bei agentischer KI und autonomen Agenten einhergeht, müssen CISOs diese Funktionen genauer unter die Lupe nehmen. Sie müssen verstehen, wie verschiedene Modelle aufgebaut, trainiert und validiert werden und wie sie sich verhalten, wenn sie mit neuen oder unbekannten Daten konfrontiert werden.

Anbieter nutzen eine Vielzahl unterschiedlicher Ansätze, und jede Designentscheidung beeinflusst die Genauigkeit, Zuverlässigkeit und die Wahrscheinlichkeit, blinde Flecken zu verursachen. Klarheit darüber zu gewinnen, wie eine KI-Funktion arbeitet und ob sie tatsächlich für die Aufgabe geeignet ist, die sie zu lösen vorgibt, ermöglicht es Sicherheitsverantwortlichen, präzisere Fragen zu stellen und sicherzustellen, dass die Technologie richtig eingesetzt wird, ohne unnötige Risiken zu schaffen.

Ebenso entscheidend ist es zu verstehen, ob die KI sicher eingesetzt und verantwortungsvoll gesteuert wird. Mangelnde Transparenz, schwache Kontrollen und inkonsistente Wartung können zu Leistungseinbußen, Model Drift, Bias und ausnutzbaren Schwachstellen führen.

Gut konzipierte Lösungen umfassen wirksame Schutzmaßnahmen, transparente Entscheidungswege und klare Kontrollen für die Datenverarbeitung und den Datenschutz. CISOs sollten Anbieter selbstbewusst zu ihren Governance-Frameworks, Test- und Validierungs-Workflows, der Resilienz gegenüber feindlichen Eingaben und Schutzvorkehrungen gegen Missbrauch oder Datenabfluss befragen. Dieses Maß an Due Diligence ist angesichts des jüngsten Anstiegs bei KI-Investitionen, -Entwicklungen und -Anbietersversprechen besonders wichtig.

Während KI bedeutende Verbesserungen bei Produktivität, Erkenntnisgewinnung und Sicherheitseffizienz ermöglichen kann, ist der Markt auch von Hype, Missverständnissen und mangelnder Transparenz geprägt.

Dieser Leitfaden soll CISOs dabei unterstützen, dieses Hintergrundrauschen zu durchdringen und bewährte Funktionen von reinen Schlagwörtern zu unterscheiden, damit sie KI-gestützte Sicherheitstools souverän evaluieren und sicherstellen können, dass jede Einführung zuverlässig, verantwortungsvoll und auf die Risiko- und Governance-Anforderungen ihres Unternehmens abgestimmt ist.



Was sollten Sie bei einer KI-Investition beachten?

Ein AI Service Provider (AISP) ist ein Anbieter, der KI-as-a-Service entwickelt und bereitstellt und dabei alles von der Modellauswahl, dem Design und Training bis hin zur Verfeinerung, Validierung und fortlaufenden Governance übernimmt. Die Lösungen eines AISP können entweder aus einem einzelnen Modell bestehen oder aus einer Kombination von Modellen zusammengestellt sein, die zusammenwirken, um ein komplexeres und umfassenderes Ergebnis zu erzielen.

Fast alle Sicherheitsanbieter beginnen damit, KI als Teil ihrer Lösungen zu nutzen. Im Bereich der Cybersicherheit kann KI extrem wertvoll sein. Während herkömmliche Sicherheitssysteme auf signaturbasierte Erkennung setzen, sind KI-Systeme in der Lage, Muster und Trends zu erkennen und zu generalisieren, um neuartige Angriffe zu identifizieren.

Was Sie wissen sollten, bevor Sie einen AISP evaluieren



AISPs können mehrere unterschiedliche Modelle innerhalb einer einzigen Lösung nutzen



KI umfasst ein breites Spektrum an Technologien (einschließlich, aber nicht beschränkt auf generative KI)



KI ist nicht immer korrekt und muss optimiert werden



Alle KI-Systeme bringen ihre eigenen Risiken mit sich, die eine angemessene Governance erfordern

KI entwickelt sich ständig weiter, was es schwierig macht, eine klare Definition von „Was ist ein KI-System?“ aufrechtzuerhalten. Die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) definiert KI als ein maschinenbasiertes System, das ohne explizite Programmierung Ausgaben aus Eingaben ableiten kann.

Das Feld der KI ist breit gefächert, und unter den Oberbegriff KI fällt eine Vielzahl von Verfahren, darunter Clustering, Klassifikation, Regression, Computer Vision, Large Language Models (LLMs), domänenspezifische Sprachmodelle (DSLMS) und agentische KI.

LLMs und DSLMs

Ein Bereich der KI, der in jüngster Zeit einen enormen Boom erlebt hat, sind Large Language Models (LLMs). LLMs sind Allzweckmodelle, die für die Verarbeitung von Freitexteingaben entwickelt wurden und sich ideal für Aufgaben im Bereich der natürlichen Sprache eignen.

Ebenso handelt es sich bei domänenspezifischen Sprachmodellen (DSLMS) um LLMs, die für eine bestimmte Aufgabe oder für den Einsatz in einer bestimmten Domäne entwickelt wurden, wie beispielsweise das DEMIST-2-Sprachmodell von Darktrace, das speziell für sicherheitsrelevante Aufgaben trainiert wurde.

Generative KI

Generative KI bezeichnet ein KI-System, das originäre Ausgaben erzeugen kann – einschließlich Text in natürlicher Sprache, Bildern und Audio. Diese Systeme haben mit dem Aufkommen Cloud-basierter KI-Chatbot-Systeme, die auf LLMs basieren, zunehmend an Beliebtheit gewonnen. Generative KI bezieht sich nicht ausschließlich auf LLMs – es gibt viele andere generative KI-Systeme wie Modelle zur Bild- und Videogenerierung (wie Dalle-2 von OpenAI oder Veo von Google), die auf derselben Technologie wie LLMs sowie weiteren generativen statistischen Verfahren basieren.

Es ist auch wichtig zu beachten, dass LLMs nicht immer generativ eingesetzt werden müssen. Es gibt eine Vielzahl weiterer LLM-basierter Ansätze, die sich für manche Aufgaben besser eignen können als generative KI.

Agentische KI

Einige AISPs haben sich in jüngster Zeit stark auf die Entwicklung agentenbasierter KI-Lösungen konzentriert. „Agentisch“ ist derzeit ein viel genutztes Schlagwort in der KI-Branche, zu dem viele Definitionen im Umlauf sind und viele AISPs behaupten, ihre Systeme seien agentenbasiert.

Die Definition von agentischer KI variiert je nach Anbieter erheblich: Sie reicht von jedem autonomen System über jedes generative KI-System, das Tools nutzen kann, bis hin zu einem reinen Marketingbegriff, der auf jedes KI-System anwendbar ist.

Angesichts dieser Unterschiede ist es unerlässlich, potenzielle AISPs zu fragen, was sie unter agentischer KI verstehen, wie agentische KI in ihre Lösung implementiert ist und welche Risiken ihre Implementierung mit sich bringen könnte.

Composite AI

Eine weitere wichtige Technik, die AISPs einsetzen können, ist Composite-KI – die Kombination mehrerer unterschiedlicher KI-Techniken in einer einzigen Lösung. Dadurch können sich mehrere Systeme gegenseitig unterstützen, blinde Flecken reduzieren und den Anwendungsbereich der Lösung erweitern.

Die Grenzen von KI in der Cybersicherheit verstehen

Die Einführung von KI-Systemen in eine Lösung kann zwar von Vorteil sein und durch autonome Aktionen einzigartige Einblicke sowie Effizienzsteigerungen bieten, es ist jedoch wichtig zu bedenken, dass alle KI-Systeme ihre eigenen Nachteile und potenziellen Risiken mit sich bringen.

Während generative KI-Systeme beispielsweise aufgrund ihrer Flexibilität und ihrer Fähigkeit, unstrukturierte Daten in natürlicher Sprache zu verarbeiten und auszugeben, immer beliebter werden, schneiden sie bei manchen Aufgaben besser ab als bei anderen.

Wenn Käufer eine Investition in eine KI-Lösung in Betracht ziehen, sollten sie prüfen, welche Vorteile KI gegenüber anderen Ansätzen bietet, und die Vielfalt der verfügbaren KI-Methoden sowie die spezifischen Vorteile der jeweiligen Ansätze bewerten. In diesem Dokument behandeln wir einige der zentralen Fragen, die Sie Ihrem AISP stellen können, um dessen Ansatz für KI-Entwicklung, -Implementierung und -Governance zu verstehen.

Schließlich beschleunigt sich die Einführung von KI-Systemen zwar rasant, doch die damit verbundene Governance und das Risikomanagement stecken größtenteils noch in den Kinderschuhen. Darüber hinaus steigen die Anforderungen an Governance und Risikomanagement nur noch weiter an, je komplexer die Systeme werden. Zu verstehen, welche Kontrollmechanismen Anbieter hier implementiert haben, kann eine entscheidende Rolle dabei spielen, zwischen beeindruckenden technischen Demos und Systemen zu unterscheiden, die für einen vertrauenswürdigen, robusten Unternehmenseinsatz geeignet sind.

Insbesondere für generative KI legt Gartner nahe, dass sie:



Besonders nützlich für die Inhaltsgenerierung, dialogorientierte Benutzeroberflächen und die Wissenserschließung



Einigermaßen nützlich für Segmentierung/Klassifizierung, Empfehlungssysteme, Wahrnehmung, intelligente Automatisierung, Anomalieerkennung/Überwachung, autonome Systeme



Kaum nützlich für Vorhersagen/Prognosen, Planung, Decision Intelligence

Darüber hinaus sind einige Studien von Gartner zu folgendem Schluss gekommen: :

Um präzisere und robustere KI-Lösungen zu erzielen, sollten KI-Führungskräfte über den Einsatz eines einzelnen Modells oder einer einzelnen Technik hinausgehen, Composite-KI-Ansätze nutzen und eine ganzheitliche KI-Systemperspektive einnehmen.

So bewerten Sie KI-Anbieter

Bei der ersten Recherche zu AISP gibt es einige Anlaufstellen, an denen Käufer ansetzen können, um ihren Kandidatenkreis einzugrenzen. Dies kann durch die Suche nach Belegen und Kundenreferenzen erfolgen.

Diese liefern den Nachweis, dass das KI-Tool funktioniert und dass es auch für Ihre Organisation funktionieren kann. Achten Sie insbesondere auf Erkenntnisse von Kunden in einer ähnlichen Situation wie Ihre Organisation, sei es in Bezug auf Branche, Unternehmensgröße, Region oder andere Merkmale.

Auch bei der Bewertung einer AISP wäre es sinnvoll, den Anbieter hinsichtlich einer Kultur guter Governance zu prüfen. Das bedeutet zu verstehen, wie Governance in die Art und Weise eingebettet ist, wie das Unternehmen KI entwickelt, testet und bereitstellt.

Ergebnisse unseres Berichts **Darktrace State of AI Cybersecurity** haben ergeben, dass nur 37 % der Unternehmen über eine KI-Governance-Richtlinie verfügen.



Es ist wichtig zu erkennen, dass selbst dann, wenn ein Unternehmen über eine KI-Governance-Richtlinie verfügt oder angibt, sich an einem anerkannten KI-Governance-Standard auszurichten, entscheidend ist, wie dieses Unternehmen diese Richtlinien und Prozesse umsetzt. Je stärker die Kultur der KI-Governance in einem Unternehmen ausgeprägt ist und je stärker sie übernommen wird, desto wahrscheinlicher ist es, dass sich dies auf seine Produkte überträgt.

Käufer sollten potenziellen AISP Fragen zu den von ihnen implementierten Governance- und Datensicherheitsprozessen, den eingesetzten Technologien und den Workflows für Test, Bewertung, Validierung und Verifizierung stellen. Durch das Stellen dieser Fragen kann ein Käufer weitere Informationen einholen, um zu bestätigen, dass eine Lösung den geschäftlichen Anforderungen und Richtlinien entspricht.

In den folgenden Abschnitten stellen wir Fragen vor, die Käufer zu 5 Schlüsselkategorien stellen können:

- 01 Governance-, Sicherheits- und Datenkontrollen
- 02 Datenerfassung und Training
- 03 Modell- und Methodenauswahl
- 04 Leistungs- und Genauigkeitsvalidierung
- 05 Interpretierbarkeit, Anpassbarkeit und Transparenz



Governance-, Sicherheits- und Datenkontrollen

KI-Governance ist essenziell, um sicherzustellen, dass alle KI-Systeme und -Modelle innerhalb einer Lösung sicher, präzise und gut gepflegt sind, sowie um zu überprüfen, dass alle in ein System eingegebenen Daten auf verantwortungsvolle, ethische und rechtmäßige Weise verwendet werden. Ein starkes Governance-System bietet intern ausreichende Transparenz über alle Systeme hinweg und ermöglicht es einem Unternehmen, Datensicherheit, Schutz und Integrität zu gewährleisten.

Wenn es um komplexere, zusammengesetzte KI-Systeme mit einer großen Anzahl von Modellen geht, ist Governance entscheidend. Bei Lösungen, die mehrere KI-Systeme umfassen, wäre es ohne eine ausreichend robuste Governance-Lösung zur Nachverfolgung und Dokumentation aller Systeme schwierig sicherzustellen, dass alle Modelle auf dem neuesten Stand und sicher gehalten sowie zur Verbesserung der Genauigkeit kontinuierlich optimiert werden.

Eine inkonsistente Wartung und mangelnde Transparenz über interne Systeme könnten dann dazu führen, dass Schwachstellen, Mängel und Verzerrungen entstehen, was die Leistung und Genauigkeit der Lösung beeinträchtigen würde.

Im Folgenden finden Sie einige Beispielfragen, die Käufer potenziellen KI-Dienstleistern stellen sollten, wenn sie Informationen über deren Governance-Richtlinien und -Prozesse einholen.

FRAGEN AN ANBIETER

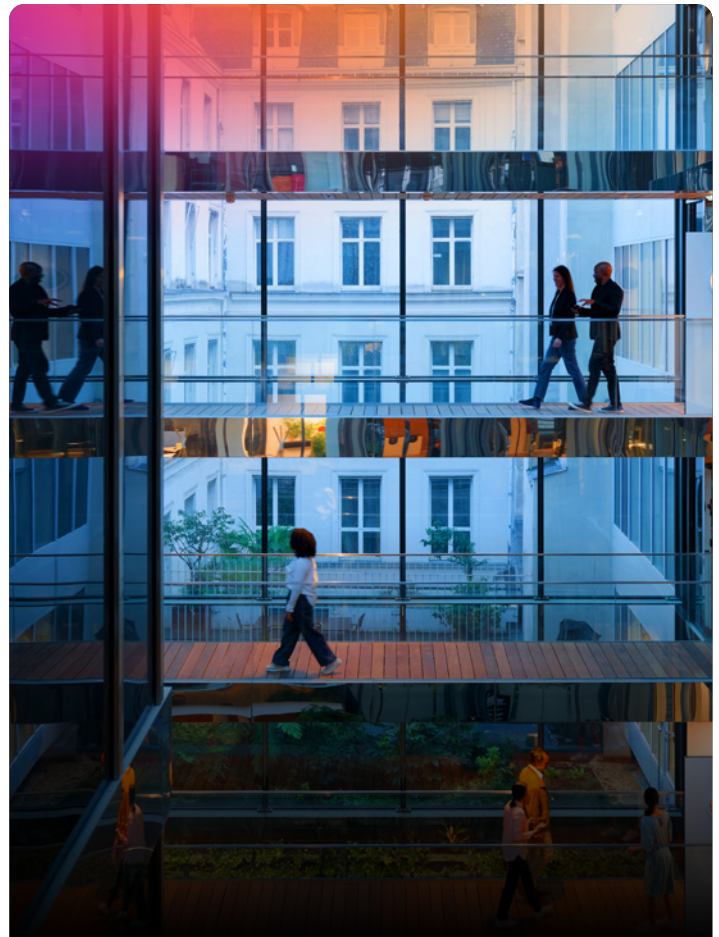
Welche KI-Governance-Richtlinien und -Frameworks befolgen Sie und/oder welche Zertifizierungen halten Sie derzeit?

Wie gewährleisten Sie Datensicherheit, Datenschutz und die Abschottung von Kundeninformationen (IP, PII etc.) in Ihren KI-Lösungen?

Wie schützen Sie Ihre KI-Systeme vor böswilligen Eingaben, Datenverfälschung und anderen adversariellen Angriffen?

Nutzen Sie extern gehostete Modelle oder Plattformen von Drittanbietern, um Kunden- oder Unternehmensdaten zu verarbeiten oder zu handhaben? Und wenn ja, wie stellen Sie sicher, dass diese Systeme sicher, geschützt und vertraulich sind?

Bei der Prüfung von Antworten auf Fragen zu Governance und Datensicherheit sollten Käufer auf Verweise auf spezifische Zertifizierungen, Workflows oder Frameworks achten, die ihr AISP nutzt, um eine starke Governance zu gewährleisten.



Darktrace ist nach dem Standard ISO/IEC 42001 zertifiziert. ISO/IEC 42001 ist der weltweit erste Standard für KI-Managementsysteme (AIMS) und bietet wertvolle Orientierung für die besonderen Herausforderungen, die KI mit sich bringt. Für Unternehmen legt er einen strukturierten Ansatz für das Management von Risiken und Chancen im Zusammenhang mit KI fest und unterstützt eine starke sowie nachhaltige Governance aller KI-Systeme.

Es ist wichtig zu beachten, dass ISO/IEC 42001 nicht das einzige verfügbare Governance-Framework für KI ist; auch andere Frameworks wie das NIST AI Risk Management Framework werden häufig genutzt. ISO/IEC 42001 bietet jedoch durch externe Audits ein zusätzliches Maß an Sicherheit, anstatt sich auf eine selbst erklärte Einhaltung zu stützen.

Neben KI-Governance-Richtlinien sind Datensicherheit und Datenschutz für alle Systeme von entscheidender Bedeutung. Daher kann die Identifizierung von Datensicherheits-Frameworks und -Zertifizierungen Käufern auch die Gewissheit geben, dass ein System ausreichend sicher ist und mit ihren Daten sicher umgehen kann.

Datenerfassung und Training

Einer der wichtigsten Schritte bei der Entwicklung eines KI-Systems ist die anfängliche Erfassung von Trainingsdaten. Wenn ein KI-System mit minderwertigen, inkonsistenten, ungenauen oder verzerrten Daten trainiert wird, wird das resultierende System niemals eine optimale Leistung erbringen.

ISO/IEC 42001 und andere Governance-Frameworks bieten Kontrollmaßnahmen, die bei der Erfassung von Daten für KI-Systeme befolgt werden sollten. Dazu gehören die Quellidentifizierung und Audit-Protokolle, um sicherzustellen, dass Trainingsdaten rückverfolgbar sind und nicht manipuliert wurden, die Durchführung einer ausreichenden De-Identifizierung oder Abstraktion der erfassten Benutzerdaten basierend auf dem Risikokontext sowie die gründliche Prüfung auf Verzerrungen (Bias), wo dies angebracht ist.

Im Folgenden finden Sie eine Liste von Fragen, die Käufer als Ausgangspunkt für Fragen zur Datenerfassung heranziehen können.

FRAGEN AN ANBIETER

Welche Schritte unternehmen Sie, um Verzerrungen in Ihren KI-Modellen und Trainingsdaten zu verhindern?

Wie minimieren Sie die Risiken im Zusammenhang mit Trainingsdaten (z. B. Regurgitation, Datenherkunft)?

Bei der Prüfung der Antworten sollten Käufer bewerten, ob der Anbieter über ein KI-Managementsystem verfügt, das die Beschaffung und Verarbeitung von Trainingsdaten regelt, und ob klare Prozesse zur Identifizierung und Minderung von Verzerrungen oder systemischen Problemen bestehen.

In Fällen, in denen personenbezogene Daten (PII) in KI-Trainingsdaten enthalten sind (außerhalb unternehmensspezifischer, selbstlernender Systeme), sollten AISPs in der Lage sein, auf spezifische Kontrollen und Prozesse zu verweisen, die gegebenenfalls zur De-Identifizierung der Daten (Entfernung von PII) oder für sonstige Anonymisierungsprozesse eingesetzt werden.

Modell- und Methodenauswahl

Verschiedene KI-Modelle und -Techniken eignen sich für unterschiedliche Aufgaben. Die Auswahl des richtigen Modells für eine bestimmte Aufgabe ist ein wichtiger Schritt, um sicherzustellen, dass eine Lösung präzise und effizient ist.

Modelle und Techniken erfordern außerdem eine solide Governance, und dabei geht es nicht nur um die Compliance mit Standards und Zertifizierungen; es geht auch darum, die effektivsten und effizientesten Modelle und Prozesse einzusetzen, um das angestrebte Ziel zu erreichen. Dazu gehört das Führen detaillierter Protokolle darüber, welche Modelle und Techniken verwendet werden, und sicherzustellen, dass diese kontinuierlich auf Risiken und Schwachstellen analysiert werden – einschließlich der Lieferkettenrisiken von Drittanbieter- und internen Modellen.

Um sicherzustellen, dass KI-Dienstleister bei der Auswahl sowohl intern als auch extern gehosteter Modelle und Services angemessene Governance-Verfahren befolgen, finden Sie hier einige Fragen, die Käufer stellen sollten:

FRAGEN AN ANBIETER

Welche Art(en) von KI-Modell(en) nutzen Sie in Ihrer Lösung?

Nutzen Sie generative KI (intern oder extern gehostet), und wenn ja, welche Schutzmaßnahmen und Guardrails haben Sie dafür eingerichtet?

Wie stellen Sie sicher, dass Ihre Modellauswahl für jede erforderliche Aufgabe optimal ist?

Nutzen Sie agentenbasierte KI? Wenn ja, wie und wo wird sie eingesetzt?

Verwenden Sie Foundation- oder Allzweckmodelle? Falls ja, wie werden diese gehostet?

Während spezifische, detaillierte Informationen über von AISPs eingesetzte maßgeschneiderte Systeme wahrscheinlich proprietär sind, sollten Käufer erwarten können, dass Anbieter einen Überblick über die grundlegend verwendeten Techniken bereitstellen. Transparenz hinsichtlich der Nutzung von Drittanbieter- oder Cloud-gehosteten Foundation-Modellen, zugehörigen APIs und relevanter Datenresidenz bleibt ein wichtiger Faktor für Käufer, die Gewissheit bezüglich des Umgangs mit Daten und deren Integrität suchen.

AISPs sollten auf Due-Diligence-Prüfungen sowie eine gründliche Untersuchung und Analyse von Drittanbietern verweisen, bevor sie deren externe Systeme implementieren, um Schutz und Sicherheit zu gewährleisten.

Leistungs- und Genauigkeitsvalidierung

KI-Systeme sollten vor dem allgemeinen Einsatz ausführlichen TEVV-Prüfungen (Testing, Evaluation, Verification, and Validation) unterzogen werden. Die Evaluierung eines Systems, gefolgt von Optimierungen, Anpassungen und Feinabstimmungen, ist unerlässlich, um sicherzustellen, dass die erstellten Systeme angemessen und sicher kalibriert sind, um aussagekräftige Erkenntnisse für die vorgesehene Aufgabe zu liefern.

Zu TEVV gehört mehr als nur die Validierung der Genauigkeit. Es müssen zahlreiche Verfahren angewendet werden, um sicherzustellen, dass ein System frei von Verzerrungen (Bias) ist, Daten weder über- noch unter- anpasst und sowohl effizient als auch präzise arbeitet. Eine kontinuierliche Evaluierung nach der Veröffentlichung ist ebenfalls entscheidend, um unerwünschten Daten- und Modelldrift zu verhindern. Tests sollten auch die physische Leistung des Modells abdecken und sicherstellen, dass das Modell Ergebnisse in einem angemessenen Zeitrahmen liefern kann und mit einem angemessenen Ressourcenverbrauch arbeitet.

ISO/IEC 42001 erfordert die Erstellung eines Workflows zur Bewertung der Relevanz und Genauigkeit eines Systems, der für jedes KI-System abgeschlossen werden muss, bevor es in der Produktion eingesetzt werden kann.

Nachfolgend finden Sie eine Liste von Fragen, die Käufer ihrem KI-Dienstleister stellen könnten, um sicherzustellen, dass dieser angemessene TEVV durchführt und eine ausreichende Governance für die Validierung von Leistung und Genauigkeit gewährleistet ist.

FRAGEN AN ANBIETER

Wie prüfen, testen, bewerten, verifizieren und validieren Sie die Ausgaben Ihres KI-Modells?

Wie passen sich Ihre Modelle an eine sich ständig verändernde externe Bedrohungslandschaft an, und wie reduzieren Sie das Risiko von Data Drift?

Umfasst Ihr TEVV-Workflow Tests zur physischen Leistung wie Latenz, Speicherverbrauch, Energie- und Ressourcenbedarf?

Antworten auf diese Fragen sollten direkte Verweise auf einen TEVV-Workflow enthalten, sei es als Teil eines bestehenden AIMS (oder eines anderen Governance-Systems) oder als separates Framework.

Interpretierbarkeit, Anpassbarkeit und Transparenz

Ein wichtiger Aspekt jedes KI-Systems besteht darin, sicherzustellen, dass die Datenausgabe interpretierbar ist, die Schritte zur Ermittlung dieser Ausgabe nachvollziehbar sind und Endbenutzer darauf vertrauen können, dass die KI-Ergebnisse korrekt sind. Auch wenn ein KI-System oft eine Black-Box ist, sollte ein Benutzer dennoch nachvollziehen können, warum ein System zu einer bestimmten Schlussfolgerung gelangt sein könnte, selbst wenn die genauen Schritte verborgen bleiben.

In der Welt der Cybersicherheit ist das Vertrauen in ein KI-System essenziell. Diese Systeme müssen Maßnahmen in Echtzeit ergreifen und autonom agieren, um Verbindungen zu blockieren, Nachrichten abzuweisen und den Netzwerkzugriff einzuschränken. Wenn dem System nicht vertraut werden kann, dass es diese Maßnahmen präzise und korrekt durchführt, deaktivieren Endbenutzer möglicherweise die Funktionen oder neigen dazu, Maßnahmen rückgängig zu machen, was die Abwehrfähigkeiten des Systems entscheidend beeinträchtigt. Die hohen Vertrauensanforderungen in der Cybersicherheit führen dazu, dass viele ihrer Anwendungsfälle für eine Automatisierung durch generative KI wenig geeignet sind. Die Anpassbarkeit von KI-Systemen ist eine wichtige Funktion.

Jedes Unternehmen hat unterschiedliche Verhaltensweisen, Strukturen und Kulturen. Die Möglichkeit, KI-Modelle fein abzustimmen, trägt dazu bei, übergeordnete Ziele zu erreichen, und ermöglicht es einer Lösung, ein breiteres Spektrum an Anwendungsfällen abzudecken.

Die folgenden Fragen bieten einen Ausgangspunkt für Fragen, die Käufer ihrem AISP stellen können, um sicherzustellen, dass ihr System ausreichend interpretierbar und anpassbar ist und eine starke Vertrauensbeziehung zwischen menschlichen Analysten und KI-Tools fördert.

FRAGEN AN ANBIETER

Wie stellen Sie sicher, dass Ihre KI-Systeme interpretierbar sind?

Wie fördern Sie eine Vertrauensbeziehung zwischen menschlichen Analysten und KI-Ergebnissen?

Wie kann ein SOC-Team Ihre Modelle feinabstimmen, um sie an seine individuellen Richtlinien und Geschäftsanforderungen anzupassen?

Ein AISP sollte beschreiben können, wie sich seine Lösung an die individuellen Strukturen, Richtlinien und Anforderungen eines Unternehmens anpassen lässt – sei es durch selbstlernende und autonome Anpassung oder durch manuelle Anpassungsmöglichkeiten für Endbenutzer.

Ebenfalls wertvoll ist die Bestätigung, dass Transparenz und Interpretierbarkeit berücksichtigt werden. Ein AISP sollte Informationen über die Richtlinien, Workflows und Methoden bereitstellen, mit denen er vollständige Transparenz gewährleistet und das Vertrauen in seine Lösung fördert.

Darktrace als Anbieter von KI-gestützter Cybersicherheit

Darktrace entwickelt und setzt seit über einem Jahrzehnt KI im Bereich der Cybersicherheit ein und baut seine Fähigkeiten parallel zu einer zunehmend komplexen und sich schnell verändernden Bedrohungslandschaft kontinuierlich aus.

Diese Erfahrung hat zu einem ausgereiften, mehrschichtigen Ansatz für KI geführt, bei dem die selbstlernende KI von Darktrace kontinuierlich die normalen Muster jeder Organisation lernt, um Verhalten zu verstehen, Kontext zu interpretieren und aussagekräftige Abweichungen zu erkennen – ohne auf vordefinierte Regeln oder bekannte Angriffssignaturen angewiesen zu sein. Im Laufe der Zeit hat dies ein bewährtes Verhaltensverständnis ermöglicht, das hilft, subtile Risikosignale aufzudecken, die andernfalls möglicherweise übersehen würden.

Dieser Ansatz wurde in dynamischen, domänenübergreifenden Umgebungen entwickelt und validiert, in denen sich Aktivitäten über Netzwerke, Identitäten, Cloud, E-Mail und Anwendungen erstrecken. Durch das Verständnis von Verhalten über Systeme hinweg bietet Darktrace eine kontinuierliche, unternehmensweite Übersicht über Aktivitäten und ermöglicht Erkennung, Untersuchung und Reaktion, die auch bei der Weiterentwicklung von Umgebungen und Bedrohungen adaptiv bleiben. Dies spiegelt ein langjähriges Engagement für Innovation wider, das auf praktischer Anwendung und kontinuierlicher Weiterentwicklung aufbaut – einhergehend mit dem Fokus auf eine verantwortungsvolle und transparente KI-Entwicklung. Als Pioniere in diesem Bereich haben wir zudem sichergestellt, dass Verantwortung ein zentraler Grundsatz unseres Prozesses ist. Dieses Engagement spiegelt sich in unserer Zertifizierung nach ISO/IEC 42001 wider und stellt die Fähigkeit von Darktrace unter Beweis, branchenführende selbstlernende KI für die Resilienz in der Cybersicherheit bereitzustellen.

Unsere Stakeholder, Kunden und Partner können darauf vertrauen, dass Darktrace seine KI-Systeme verantwortungsvoll, ethisch und sicher entwickelt und den Einsatz von KI in unserem täglichen Geschäftsbetrieb auf konforme, sichere und ethische Weise handhabt.

Wie Darktrace KI-Systeme schützt

Darktrace stellt mit Darktrace / SECURE AI nun diese Funktionen bereit, um von KI-Systemen ausgehende Risiken in Unternehmen zu überwachen und darauf zu reagieren. Diese Lösung analysiert, wie Prompts, Agenten und Systeme im Kontext des jeweiligen Unternehmens genutzt werden, und führt jede KI-Interaktion in einer zentralen Ansicht zusammen. Dieser einzigartige Ansatz hilft Teams dabei, Absichten zu verstehen, Risiken zu bewerten, sensible Daten zu schützen und Richtlinien sowohl für menschliche Aktivitäten als auch für Aktivitäten von KI-Agenten durchzusetzen. Sie bietet Echtzeit-Transparenz und -Kontrolle über generative KI, KI-Agenten, Entwicklungsumgebungen und Schatten-KI hinweg und ermöglicht es Teams, Missbrauch, Fehlkonfigurationen und Abweichungen zu erkennen, die regelbasierte Ansätze oft übersehen.

Indem es Kontext und Absicht über Benutzer und Systeme hinweg versteht, hilft es Unternehmen dabei, die Übersicht zu behalten, während der Einsatz von KI im gesamten Unternehmen zunimmt.

Melden Sie sich hier für das Secure AI Readiness Program an: Damit erhalten Sie exklusiven Zugriff auf die neuesten Nachrichten zu aktuellen KI-Bedrohungen, Updates zu neuen Ansätzen für die KI-Sicherheit und Einblicke in die neuesten Innovationen, einschließlich der fortlaufenden Arbeit von Darktrace in diesem Bereich.

Möchten Sie mit einem Darktrace-Experten über die Absicherung von KI sprechen? Registrieren Sie sich hier, um praktische Orientierung zu den KI-Risiken zu erhalten, die für Ihr Unternehmen am wichtigsten sind, sowie Klarheit darüber, worauf Sie sich bei Governance, Transparenz, Risikoreduzierung und langfristiger Vorbereitung zuerst konzentrieren sollten.



Der Ansatz von Darktrace für verantwortungsvolle KI in der Cybersicherheit

Erfahren Sie mehr über die Prinzipien hinter dem verantwortungsvollen KI-Ansatz von Darktrace, die auf der Zusammenarbeit mit weltweiten Experten aus Wissenschaft und Behörden basieren, und wie Verantwortlichkeit, Erklärbarkeit und kontinuierliche Validierung in die Cybersicherheitstechnologie integriert sind.

Lesen Sie [hier](#) mehr

AI Arsenal

Erfahren Sie im Whitepaper „AI Arsenal“ mehr darüber, wie überwachtes und unüberwachtes maschinelles Lernen sowie LLMs in der Cybersicherheit eingesetzt werden können und wie Darktrace spezifische Techniken in einer mehrschichtigen Architektur kombiniert, um menschliche Sicherheitsteams zu unterstützen.

Lesen Sie [hier](#) mehr



Fazit

Da KI immer stärker in die Cybersicherheit integriert wird, ist für Sicherheitsverantwortliche nicht mehr entscheidend, ob ein Produkt KI nutzt, sondern wie diese KI konzipiert, gesteuert, validiert und kontrolliert wird. Käufer sollten klare Antworten zu Datenbeschaffung und Datenschutz, Modellauswahl, Test- und Validierungspraktiken, Resilienz gegenüber gegnerischen Eingaben (Adversarial Inputs) und zur Transparenz KI-gestützter Entscheidungen erwarten können.

Starke KI-Lösungen basieren auf einer Kombination sorgfältig ausgewählter Verfahren, kontinuierlicher Bewertung und Governance-Rahmenwerken, die Risiken im Laufe der Zeit reduzieren. So bleiben Systeme präzise, anpassungsfähig und an Unternehmensrichtlinien ausgerichtet, während sich Bedrohungen weiterentwickeln.

Bei verantwortungsvoller Implementierung bietet KI wesentliche Vorteile: eine verbesserte Erkennung neuer Bedrohungen, schnellere und präzisere Untersuchungen, eine geringere operative Belastung für Sicherheitsteams sowie größeres Vertrauen in autonome Reaktionen. Der mehrschichtige, selbstlernende Ansatz von Darktrace zeigt, wie verschiedene KI-Techniken, eine starke Governance und nachvollziehbare Ergebnisse zusammenwirken können, um vertrauenswürdige Cybersicherheit im großen Maßstab bereitzustellen.

Kontaktieren Sie Darktrace für eine personalisierte Demo

[Erfahren Sie mehr ↗](#)

■ **Acerca de Darktrace**

Darktrace es líder mundial en ciberseguridad basada en IA que permite a las organizaciones adelantarse cada día al panorama de amenazas en constante evolución. Fundada en 2013 en Cambridge (Reino Unido), Darktrace ofrece la plataforma de ciberseguridad esencial para proteger a las organizaciones frente a amenazas desconocidas mediante una IA que aprende de cada empresa en tiempo real. La plataforma y los servicios de Darktrace cuentan con el respaldo de más de 2.700 empleados que protegen a cerca de 10.000 clientes en todo el mundo. Para más información, visite www.darktrace.com.