TrendAI™

Vertrauen in KI schaffen

sichtbar, kontrollierbar, sicher

Özgür Isik - Senior Security Consultant - TrendAI

“… und auf einmal waren die Texte super formuliert und ohne erkannbare Fehler…”

- Gespräch auf der GTC

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
-
-
-
-
-

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
-
- Vertrauliche Daten
-
-
-

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
-
- Vertrauliche Daten
-
- Halluzination
-

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
- Governance und Compliance
- Vertrauliche Daten
-
- Halluzination
-

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
- Governance und Compliance
- Vertrauliche Daten
- KI mit zu vielen Rechten
- Halluzination
-

TrendAI™

Welche Risiken müssen wir verstehen?

- Schatten KI
- Governance und Compliance
- Vertrauliche Daten
- KI mit zu vielen Rechten
- Halluzination
- Neue Angriffswege

TrendAI™

Die fünf Risikodimensionen von AI im Unternehmen

TrendAI

1

von fünf Risikodimensionen

Ist meine KI-Anwendung sicher, bevor sie live geht?

Automatisiertes Red-Teaming simuliert Tausende realer Angriffsmuster gegen die API der Anwendung: Prompt-Injection-Payloads, Jailbreak-Versuche, Datenextraktion. Das Ergebnis ist ein Prüfbericht vor dem Produktivstart.

Aktuell das am schnellsten wachsende Nachfragesegment im gesamten AI-Security-Markt, noch vor dem Laufzeitschutz.

Einfach erklärt · Ein TÜV-Check für die KI-Anwendung, nur dass das System Tausende „Testfahrten" in kürzester Zeit durchführt.

2

von fünf Risikodimensionen

Ist meine KI-Anwendung im laufenden Betrieb geschützt?

Ein Wächter zwischen Nutzer und Modell prüft jede Anfrage und jede Antwort in Echtzeit: eingehend auf Prompt Injection, ausgehend auf die Preisgabe vertraulicher Daten, und blockiert bei Verdacht.

Einmaliges Testen reicht nicht, weil sich Angriffstechniken laufend weiterentwickeln. Laufzeitschutz ist ein dauerhafter Prozess, kein einmaliges Siegel.

Einfach erklärt · Der Wachdienst, der dauerhaft im Haus bleibt, nachdem der TÜV-Check bestanden wurde.

3

von fünf Risikodimensionen

Habe ich den Überblick über alle KI-Assets?

Kontinuierlicher Scan der gesamten Cloud- und IT-Umgebung: findet auch KI-Dienste, die einzelne Teams ohne zentrale Abstimmung aufgesetzt haben, und bewertet jede Konfiguration gegen die Sicherheitsrichtlinien.

Ohne dieses Lagebild bleiben Fehlkonfigurationen, etwa ein offen zugänglicher Speicher mit Trainingsdaten, oft bis zum Vorfall unentdeckt.

Einfach erklärt · Ein Kontrolltrupp, der ständig prüft, ob alle Türen verschlossen sind, auch die, von denen niemand wusste, dass sie existieren.

4

von fünf Risikodimensionen

Sind die Bausteine meiner KI-Anwendung selbst sicher?

KI-Anwendungen bestehen aus fremden Bausteinen: vortrainierte Modelle, Open-Source-Bibliotheken, Container-Images, externe Trainingsdaten. Jeder Baustein wird geprüft, idealerweise bereits in der CI/CD-Pipeline.

Ein kompromittierter Baustein wirkt sich auf jede Anwendung aus, die darauf aufbaut.

Einfach erklärt · Liefert der Fisch-Lieferant verseuchte Ware, ist das Gericht ungenießbar, egal wie sauber die eigene Küche arbeitet.

5

von fünf Risikodimensionen

Wer nutzt welche KI-Dienste, ohne dass die IT es weiß?

Mitarbeiter kopieren Verträge, Kundendaten oder Quellcode in öffentliche KI-Dienste wie ChatGPT oder Copilot. Die Daten verlassen dabei unbemerkt die Kontrolle des Unternehmens. Unautorisierte Browser-Erweiterungen mit KI-Funktion greifen zusätzlich im Hintergrund auf Unternehmensdaten zu.

Wichtig für den Pitch: die öffentlich sichtbarste Dimension, aber nur eine von fünf gleichrangigen.

Einfach erklärt · Vertrauliche Unterlagen, die ein Mitarbeiter mit nach Hause nimmt, nur dass die Daten in Sekundenbruchteilen an einen fremden Onlinedienst gehen.

Vision One: eine Plattform für alle fünf Dimensionen

TrendAI

Drei Schutzrichtungen, ein Ziel

Security of AI

Schutz der KI-Systeme selbst: Modelle, Trainingsdaten, APIs und die zugrundeliegende Infrastruktur gegen gezielte Angriffe wie Prompt Injection oder Model Poisoning.

Security for AI

Der übergeordnete Rahmen: beide Richtungen vereint in einer einzigen, integrierten Plattform, statt sie als getrennte Projekte zu betreiben.

Security by AI

Einsatz von KI, um klassische Cyberangriffe schneller zu erkennen und abzuwehren: Vorfilterung, Korrelation und Kontextualisierung riesiger Datenmengen statt manueller Log-Analyse.

Einfach erklärt · Es reicht nicht, nur die ChatGPT-Nutzung abzusichern. Es geht um weit mehr Angriffsflächen, und KI selbst wird zum wichtigsten Werkzeug, um all das zu bewältigen.

Vision One AI Security

Fünf Dimensionen, fünf native Antworten

Risikodimension	Vision One Baustein	Kernfunktion
AI Security Testing	**AI Scanner**	Automatisiertes Red-Teaming und Pentesting von KI-App-APIs vor der Produktivsetzung
Runtime Defense	**AI Guard**	Echtzeitschutz gegen Prompt Injection, Missbrauch und Datenleckage im laufenden Betrieb
AI Security Posture Management	**AI-SPM**	Kontinuierlicher Überblick über den Sicherheitszustand aller KI-Cloud-Assets
AI-Software-Lieferkette	**AI Security Blueprint, Code & Container Security**	Absicherung von Code-Repositories, CI/CD, Containern und Modellen über den gesamten Lifecycle
Schatten-AI / Nutzungskontrolle	**AI Secure Access (ZTSA)**	Zero-Trust-Zugriffskontrolle und Prompt-Inspektion für Drittanbieter-KI-Dienste

Zusammenfassung



KI-Nutzung identifizieren



Auswirkung und Risiko verstehen



Zielgerichtete Kontrollen implementieren



Interne Kommunikation und Schulung

TrendAI™

Danke!!!

Özgür Isik

Oezguer_Isik@trendmicro.com

