

Was KI nicht kann

Warum Resilienz mehr braucht als Algorithmen



Künstliche Intelligenz: Hoffnungsträger mit Grenzen

Künstliche Intelligenz (KI) gilt als Game-Changer in der IT-Security. Sie erkennt Muster, analysiert riesige Datenmengen in Echtzeit und reagiert schneller als jeder Mensch. Doch trotz aller Fortschritte zeigt sich in der Praxis: KI allein macht Unternehmen nicht resilient. Denn Resilienz, verstanden als die Fähigkeit, Cyberangriffe zu überstehen und sich schnell zu erholen, ist mehrdimensional. Sie entsteht im Zusammenspiel von Technologie, Mensch und Prozess.

Genau hier setzt das Beratungsverständnis von Adlon an. Als Experten für die Konvergenz von IT und OT helfen wir Unternehmen, diese drei Säulen in Einklang zu bringen und eine tragfähige Resilienz-Strategie zu entwickeln, die über den KI-Hype hinausgeht.

Technische Schulden

Der blinde Fleck der Automatisierung: In vielen Unternehmen – so auch in der Industrie – hat sich über Jahre hinweg ein Flickenteppich aus Altsystemen, Workarounds und nicht dokumentierten Schnittstellen gebildet. Diese technischen Schulden sind ein idealer Nährboden für Sicherheitslücken und ein Bereich, in dem KI oft ins Leere läuft.

Denn Algorithmen können nur das analysieren, was sie „sehen“. KI-Systeme sind auf Daten angewiesen, die ihnen zugänglich gemacht werden. Fehlen Protokolle, Logging oder Schnittstellen für Scans, bleibt ein System, dass für die KI unsichtbar ist. Veralterte Geräte ohne Monitoring, proprietäre Protokolle oder nicht inventarisierte Assets entziehen sich so jeder automatisierten Analyse.

Ein Beispiel

Ein Bericht von Adlon aus dem Jahr 2024 beschreibt ein Projekt bei einem mittelständischen Maschinenbauer: Dort entdeckte ein IT-Securityberater über 30 nicht dokumentierte Systeme im Produktionsnetzwerk, darunter einen Windows-XP-Rechner mit direkter Internetverbindung. Ohne menschliche

Intervention hätte kein KI-System diese Schwachstelle erkannt, weil schlicht keine Daten vorhanden waren, die auf das Gerät hingewiesen hätten. Für Unternehmen in der Produktionsautomatisierung besteht hier das höchste Risiko, da die Sichtbarkeit im OT-Netzwerk oft geringer ist als in der klassischen IT. Die kritische Schwachstelle liegt in der fehlenden Datenbasis, nicht im Algorithmus selbst.

Schatten-IT

Wenn der Mensch die Regeln umgeht: Ein weiteres Problemfeld in der Schatten-IT sind Anwendungen oder Cloud-Dienste, die Mitarbeiter ohne Wissen der IT-Abteilung nutzen. Laut Gartner greifen bis zu 40 % der Beschäftigten regelmäßig auf nicht genehmigte Tools zurück. Da KI auf zentrale Logs und Policies angewiesen ist, bleiben solche Umgehungen oft unerkannt oder werden falsch interpretiert. Nachhaltige Lösungen erfordern daher mehr als Technologie: Awareness-Maßnahmen, klare Governance und regelmäßige Audits. KI kann hier unterstützen, führen muss der Mensch.

Unternehmen können mit dem Prinzip „Security by Design“ Sicherheitsanforderungen von Anfang an in Prozesse und Systeme integrieren – und nicht erst nachträglich ergänzen. Standards wie ISO/IEC 27001, dem BSI IT-Grundschutz und branchenspezifischen Normen wie IEC 62443 helfen bei der Umsetzung. Um die Lücke zu schließen, müssen Unternehmen Governance-Frameworks einführen, die sowohl IT als auch OT umfassen und Mitarbeiter aktiv in die Cybersicherheitsstrategie einbinden. Tools zur Netzwerkerkennung können Schatten-Assets zwar auch mithilfe von KI identifizieren, doch die Entscheidung über deren Stilllegung oder Integration bleibt eine Governance-Aufgabe des IT-Leiters im Zusammenspiel mit der Geschäftsführung.

Mensch und Maschine

Ein resilientes Duo: Die Zukunft der IT-Security liegt nicht in der Ablösung des Menschen durch KI,

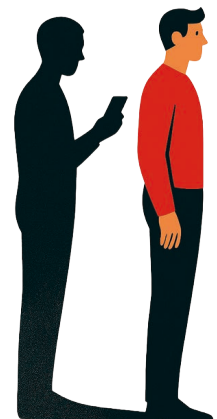
sondern in der Zusammenarbeit. Bevor eine KI überhaupt Angriffe erkennen kann, müssen die relevanten Datenquellen zuverlässig angebunden und korrekt aufbereitet werden. Das bedeutet: Logdaten müssen vollständig, strukturiert und kontextbezogen im SIEM-System (= Security Information and Event Management) ankommen. Eine Aufgabe, die technisches Verständnis, Prozesskenntnis und menschliche Sorgfalt erfordert.

Erst wenn diese Grundlage geschaffen ist, kann die KI „loslegen“: Muster erkennen, Anomalien identifizieren und automatisiert reagieren. Doch auch dann bleibt der Mensch unverzichtbar. Etwa bei der Interpretation komplexer Zusammenhänge, der Eskalation kritischer Vorfälle oder der strategischen Ableitung von Maßnahmen.

Beispiel: Incident-Response-Fall

In einem Incident-Response-Fall bei einem Zulieferer der Automobilindustrie konnte ein KI-System ungewöhnliche Aktivitäten in einem Produktionsnetzwerk identifizieren.

Die finale Bewertung und Einstufung als gezielter Angriff erfolgte durch ein interdisziplinäres Team aus IT, OT und externen Forensikern. Dieser interdisziplinäre Ansatz ist gerade in der Industrie 4.0 entscheidend, da ein KI-Alarm in der IT möglicherweise einen harmlosen Prozess in der OT signalisiert. Nur der Mensch mit Domänenwissen kann diese Kontexte korrekt bewerten und die unmittelbare Gefahr für die Produktion einschätzen.



Autor:
Tizian Kohler
Head of Security
Adlon
www.adlon.de

Agentic AI

Vom Mustererkenner zum aktiven Mitspieler: Neben den hervorragenden Analysemöglichkeiten durch KI kommen immer weitere Anwendungsfälle hinzu, denn die Entwicklung bleibt nicht stehen. Mit dem Aufkommen sogenannter „Agentic AI“ verschiebt sich die Rolle der KI von der reinen Unterstützung hin zum aktiven, eigenständig handelnden Mitspieler im Security-Team.

Was bedeutet das konkret?

Agentic AI bezeichnet KI-Systeme, die nicht nur auf Anweisungen reagieren, sondern eigenständig Ziele verfolgen, Aufgaben planen und mehrere Schritte selbstständig ausführen können. Im Security-Kontext heißt das: KI-Agenten können beispielsweise eigenständig Bedrohungsdaten zusammentragen, Playbooks abarbeiten oder sogar erste Gegenmaßnahmen einleiten – immer im Rahmen klar definierter Regeln und Workflows.

Der Mensch bleibt Dirigent:

Je autonomer KI-Systeme werden, desto wichtiger wird die Führungsrolle des Menschen. In der Praxis etabliert sich zwischenzeitlich ein Stufenmodell:

- **Human-in-the-Loop (HITL):** Der Mensch gibt jeden Schritt frei, was den Prozess zwar teilautomatisiert, aber dennoch langsam macht.
- **Human-on-the-Loop (HOTL):** Die KI agiert selbstständig, der Mensch überwacht und greift bei Bedarf ein.
- **Human-out-of-the-Loop (HOOT):** KI arbeitet vollautonom. Menschen greifen hier nicht in Echtzeit ein und werden auch gar nicht mehr operativ eingebunden.
- **Human-in-Command (HIC):** Der Mensch setzt Ziele, definiert Grenzen und bleibt für die Gesamtstrategie verantwortlich. Die KI dient als Berater und ausführendes Tool.

Je autonomer der KI-Agent ist, desto schneller kann er agieren und Aktionen einleiten, was im Security-Umfeld besonders sinnvoll ist. Hier drängt sich dann aber auch immer mehr die Frage der Richtigkeit der



Entscheidung der KI-Agenten und damit verbunden der Verantwortungsübernahme auf (siehe nachfolgender Abschnitt „Nicht-Determinismus: Warum KI nicht immer gleich reagiert“).

Ein Praxisbeispiel

Automatisierte Phishing-Response mit HITL-Gate: Ein Security-Agent erkennt eine verdächtige E-Mail und startet automatisch ein Playbook. Er sammelt Header-Informationen, prüft Links auf bekannte Phishing-Domains und analysiert Anhänge in einer Sandbox. Erkennt die KI ein hohes Risiko, schlägt sie vor, den betroffenen Account temporär zu sperren und alle ähnlichen Mails im Unternehmen zu isolieren. Doch bevor diese kritische Maßnahme umgesetzt wird, muss ein Analyst den Vorschlag prüfen und freigeben. Das sogenannte HITL-Gate. So bleibt die Kontrolle beim Menschen, während Routineaufgaben automatisiert ablaufen. Diese Entwicklung eröffnet neue Möglichkeiten: etwa schnellere Incident Response oder adaptive Verteidigung, birgt aber auch neue Anforderungen an Governance, Nachvollziehbarkeit und Kontrolle.

Kontroverse um den Anthropic-Vorfall

Wie autonom war der KI-Angreifer? Ein weiteres, sehr aktuelles Beispiel zeigt, welches Potenzial und welche Risiken in agentischen KI-Systemen stecken können. Anthropic (das Unternehmen hinter dem KI-Chatbot Claude) veröffentlichte im Herbst 2025 einen Bericht über eine vermeintliche Spionagekampagne einer staatlich unterstützten chinesischen Gruppe („GTG-1002“). Laut Anthropic sollen dabei 80 - 90 % der Angriffsschritte von Claude Code automatisiert ausgeführt worden sein, vom Aufklärungsscan über das Identifizieren von verwundbaren Systemen

bis hin zur teilweisen Datenexfiltration. Für viele Experten galt dies zunächst als erster dokumentierter Fall eines nahezu autonom agierenden KI-Angreifers.

Doch genau an diesem Punkt beginnt die Kontroverse. Mehrere Sicherheitsexperten und unabhängige Forscher zweifelten die Darstellung an. Hauptkritikpunkte waren unter anderem:

- **Fehlende technische Transparenz:** Es wurden kaum Details zu den tatsächlichen Angriffsschritten oder Indikatoren veröffentlicht.
- **Unklare Definition von „autonom“:** Viele der beschriebenen Prozesse könnten in Wahrheit automatisierte, aber nicht eigenständig initiierte Tasks gewesen sein.
- **Geringe Erfolgsquote der Attacke:** Laut Forschern deutet dies darauf hin, dass menschliche Operatoren weiterhin intensiv eingebunden waren. Dies deutet, um im obigen Stufenmodell zu bleiben, auf eine Mischung aus „Human-in-the-loop“ oder „Human-on-the-loop“ hin.
- **Marketinggetriebene Interpretation:** Einige Kritiker sehen den Bericht als Versuch, die Rolle von Agentic AI dramatischer darzustellen, als sie technisch belegbar ist.

Trotz dieser Unklarheiten zeigt der Vorfall sehr gut, wohin die Entwicklung gehen kann und wie entscheidend ein korrektes Verständnis des Automatisierungsgrads ist. Die strategische Herausforderung für IT- und Produktionsleiter liegt darin, die Grenzen zwischen Automatisierung (definiertes Skript) und Agentic AI (selbstständige Entscheidungsfindung) präzise zu verstehen, um die richtigen Sicherheitsarchitekturen und Überwachungsprotokolle zu implementieren.

Nicht-Determinismus

Warum KI nicht immer gleich reagiert: Dass der Mensch sinnvollerweise weiterhin in Prozesse der IT-Sicherheit mit KI einbezogen werden sollte, zeigt sich an dieser Erkenntnis: KI ist alles andere als ein Uhrwerk. Selbst bei identischen Eingaben liefert ein Sprachmodell wie GPT nicht immer dieselbe Antwort. Der Grund: Moderne KI arbeitet probabilistisch, nicht deterministisch.

Was steckt dahinter?

Große Sprachmodelle (LLMs) wie ChatGPT oder Copilot generieren Texte auf Basis von Wahrscheinlichkeiten. Schon kleine Änderungen in den Parametern führen zu unterschiedlichen Ergebnissen. Zum Beispiel der sogenannten „Temperature“ (= steuert die Zufälligkeit – je höher, desto kreativer und weniger vorhersehbar) oder der Auswahlstrategie für die nächsten Worte („Top-p“ = wählt Wörter aus dem kleinsten Wahrscheinlichkeitsbereich, „Top-k“ = beschränkt die Auswahl auf die k wahrscheinlichsten Wörter). Hinzu kommen Effekte aus der Cloud-Infrastruktur: parallele Anfragen, Updates

Wer schreibt:

Tizian Kohler ist seit Mai 2025 Head of Security bei ADLON Intelligent Solutions GmbH. Zuvor war er bei der Kriminalpolizei als Referent für Cybercrime und Digitale Spuren tätig. Seine Expertise umfasst Netzwerksicherheit, Cloud-Security, Incident Response und digitale Forensik.

Mit dieser einzigartigen Kombination aus polizeilicher Cybercrime-Erfahrung und Unterneh-



mensberatung bringt er praxisnahe und compliance-orientierte Perspektiven in die Sicherheitsstrategie von Unternehmen.

am Modell oder gar unsichtbare Änderungen im Backend können das Resultat beeinflussen. Selbst bei fest eingestellten Parametern zeigen Studien, dass identische Prompts keine identischen Antworten liefern.

Warum ist das ein Problem für Security?

In der IT-Sicherheit zählt Nachvollziehbarkeit. Ob bei der forensischen Analyse eines Sicherheitsvorfalls der Dokumentation von Maßnahmen oder im Audit: Es muss klar sein, wie eine Entscheidung zustande kam – und dass sie wiederholbar ist. Wenn ein KI-System aber heute eine E-Mail als harmlos einstuft und morgen als Phishing, wird es kritisch.

Praxisregeln für den sicheren Einsatz

So schaffen Sie Nachvollziehbarkeit: Unternehmen müssen klare Regeln für den Einsatz von KI schaffen, insbesondere wenn diese in kritischen Pfaden der Security-Architektur eingesetzt wird:

- **Parameter festlegen:** Temperatur möglichst niedrig ansetzen, dass die Vorhersagbarkeit besser wird.
- **Versionen und Kontext loggen:** Jede Modell- und Prompt-Version, alle verwendeten Tools und den genauen Kontext dokumentieren. Dieses umfassende Logging sollte in das zentrale SIEM-System integriert werden, um die gesamte Entscheidungsstrecke des KI-Agenten im Falle eines Audits oder Incident-Response-Falls lückenlos nachvollziehen zu können.

- **Deterministische Pipelines bevorzugen:** Wo möglich, stabile Seeds und batch-invariante Abläufe nutzen.

- **Menschliche Kontrollpunkte einbauen:** Bei kritischen Aktionen wie etwa dem Sperren von Konten oder dem Entziehen von App-Berechtigungen sollte möglichst ein Mensch die finale Entscheidung treffen. Agenten dürfen nie ihre eigenen Grenzen erweitern. Diese Kontrollpunkte sind als „Golden Rules“ in den IT- und OT-Sicherheitsrichtlinien zu verankern.

Verantwortung und Haftung

Die letzte Instanz bleibt der Mensch: Die Frage nach der Haftung bei Fehlentscheidungen eines autonomen KI-Agenten ist juristisch noch weitgehend ungeklärt. Wer haftet, wenn ein HOOT-System (Human-out-of-the-Loop) einen Produktionsprozess aufgrund eines Fehlalarms stoppt und einen Millionenschaden verursacht? Nach aktuellem Stand trägt die Verantwortung immer der Betreiber oder der Entwickler des Systems, nicht die KI selbst.



der Industrie ist kein reines Technologieprojekt, sondern ein ganzheitlicher Governance-Ansatz, der technische Schulden systematisch abbaut, Schatten-IT bekämpft und den Menschen als letzte Kontrollinstanz in den Mittelpunkt stellt.

Zukunftsperspektive

Der resiliente Weg zum autonomen Security-Agenten: Der Weg zum vollautonomen Security-Agenten ist noch lang und führt über eine sorgfältige Governance. Unternehmen, die heute die Basis schaffen – durch saubere Datenpipelines, das Schließen von Transparenzlücken in IT- & OT-Bereichen und die Definition von klaren HITL- und HOTL-Prozessen – werden als Erste von der nächsten Generation der Agentic AI profitieren. Sie verwandeln die KI von einem bloßen Mustererkennungstool in einen verlässlichen, wenn auch kontrollierten, aktiven Mitspieler in der Cyberabwehr.

Möchten Sie wissen, welche Governance-Stufe (HITL, HOTL, HIC) für Ihre kritischen Produktionsprozesse die richtige ist und wie Sie Ihre Datenbasis KI-fähig machen? Sprechen Sie mit den Experten von Adlon, um Ihre individuelle Roadmap zur Cyberresilienz zu entwickeln.

Über Adlon

ADLON Intelligent Solutions GmbH ist Spezialist für den digitalen Arbeitsplatz und IT-Sicherheit. Das Unternehmen unterstützt seine Kunden seit über 35 Jahren bei der Entwicklung und Umsetzung ganzheitlicher Digital- und Sicherheitsstrategien – von Beratung über Implementierung bis zum Betrieb.

- Familienunternehmen, seit 1988
- 60 Mitarbeiter an drei Standorten in Deutschland: Ravensburg, Ulm, Friedrichshafen
- Zertifiziert nach ISO 9001, ISO 14001 und ISO 27001
- Managed ECO-Partner-Netzwerk
- IT-Beratungsunternehmen für den digitalen Arbeitsplatz mit IT-Security Fokus

Quellen:

Gartner „Market Guide for Shadow IT“ 2024

ADLON interne Fallstudien und Beratungserfahrungen ◀

Events und Downloads:

Von der Sicherheitsillusion zur echten IT-Sicherheit

On-Demand Vortrag, kostenlos

<https://adlon.de/event-von-sicherheitsillusion-zu-it-sicherheit>

NIS-2 Richtlinie pragmatisch begegnen

On-Demand Event, kostenlos

<https://adlon.de/event-nis-2-begegnen/>

Die Evolution des Security Operations Center

On-Demand Event, kostenlos

<https://adlon.de/event-evolution-soc/>

Vulnerability Management mit MXDR

Whitepaper zum kostenlosen Download

<https://adlon.de/vulnerability-management-mit-mxdr/>

