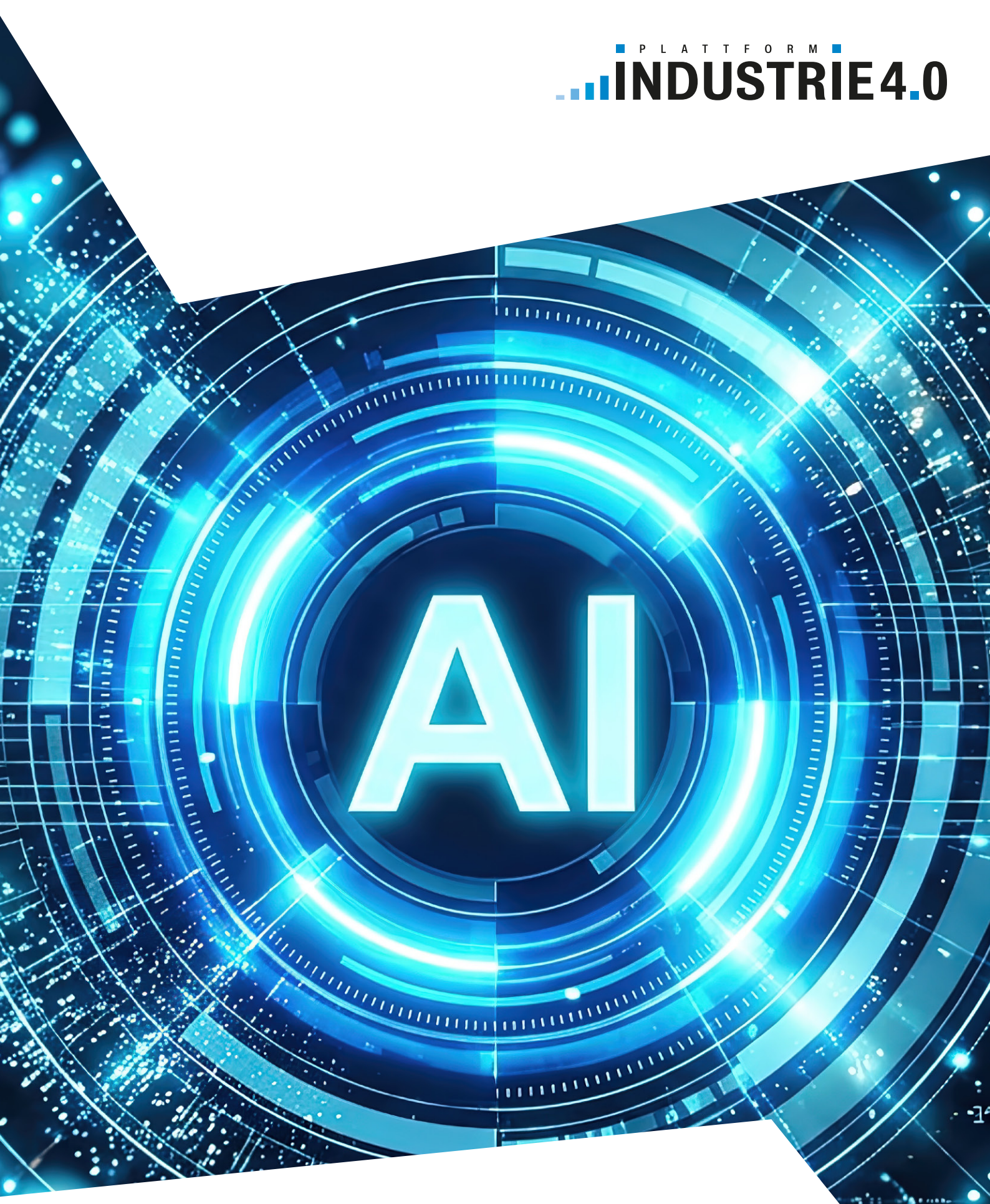
PLATTFORM INDUSTRIE 4.0



KI-gestützte Zero-Trust-Architektur für Industrie 4.0

Status, Proof of Concept und Handlungsempfehlungen für KMU

Inhalt

I. EINLEITUNG 4

II. FORTSCHRITTE UND ENTWICKLUNGEN 2024 BIS 2026 6
2.1 Kontinuierliche Verifikation und Analyse 6
2.2 Verbesserung von Threat Intelligence und Incident Response 7
2.3 Software, Hardware und Algorithmen 9
2.4 Erweiterung der Sichtweise: LLM-basierte Sicherheitsarchitekture 11

III. GRUNDVORAUSSETZUNGEN FÜR DIE NUTZUNG VON KI UND AKTUELLE GEOPOLITISCHE RANDBEDINGUNGEN 13

IV. INDUSTRIE 4.0-KONTEXTN 14
4.1 Einleitung: Der Industrie 4.0-Kontext als Arbeitsschwerpunkt 14
4.2 Zero Trust Architektur statt Perimeter 14
4.3 KI-Technologien als Assistenz 15
4.4 Kontrollstrukturen auf unterschiedlichen Ebenen 16
4.5 Weitere Aspekte 16

V. INFORMATIONEN ZUM AI-ACT DER EU 18
5.1 Risikoklassifizierung gemäß EU-AI-Act 18
5.2 Warnung vor Überklassifizierung 18

VI. DURCHFÜHRUNG UND ERGEBNISSE EINES PROOF OF CONCEPT FÜR AUSGEWÄHLTE USE CASES 19
6.1 Erfolgskriterien des PoCs 22
6.2 Untersuchte Use Cases 22
6.2.1 Experiment 1 – Echtzeit-Zugriffsanfragen auf eine (Industrie 4.0-) Entität, deren Regeln durch ein LLM geprüft und autark entschieden werden 22
6.2.2 Experiment 2 – LLM-unterstützte, asynchrone Regelgenerierung für Zugriffskontrolle, die in einer Zugriffs-Kontroll-Regelausführungsmaschine (Attribute Based Access Control/ABAC) innerhalb einer ZTA maschinell ausgeführt werden können 22
6.3 Ergebnisse des PoCs: Vorgehensweise für Experiment 2 23
6.4 Detaillierte Beschreibung von Experiment 2 26
6.4.1 Variante 1: Generierung von ABAC-Regeln für den Zugriff auf ein SAP-System 26
6.4.2 Variante 2: Generierung von ABAC-Regeln für den Zugriff auf einen Industrie 4.0-Data-/Collaboration Space 26
6.5 Zusätzliche, generelle Erkenntnisse aus dem PoC 27

VII. AUSBLICK UND HANDLUNGSEMPFEHLUNGEN 28
7.1 Zusammenfassung 28
7.2 Ausblick auf kommende Entwicklungen 29
7.3 Handlungsempfehlungen für KMU 29
7.4 Weitere Bemerkungen: 30

VIII. ANHANG 1: DETAILLIERTE UMSETZUNGS- UND EVALUIERUNGSSTRUKTUR (REFERENZ-REPOSITORY) 31
8.1 Compliance-Dokumentation ZTA+KI 31
8.1.1 Thema 31
8.1.2 Motivation 31
8.1.3 Hauptbotschaften 31
8.1.4 Dokumente im Ordner 31
8.2 Scope-Hinweis 31
8.3 Empfohlene Lesereihenfolge 31
8.4 Verbindung zum Gesamtprojekt ZTA+KI_2026

IX. ANHANG 2: FORSCHUNGSERGEBNISSE 32
9.1 Proof of Concept: LLM-gestützte ABAC-Regelgenerierung für Zero Trust Architectures 32
9.1.1 Thema 32
9.1.2 Motivation 32
9.1.3 Haupt-Botschaften 32
9.1.4 Dokumente und Ordner 32
9.2 Scope-Hinweis 32
9.3 Empfohlene Lesereihenfolge 32
9.4 Verbindung zum Gesamtprojekt ZTA+KI_2026 32

X. ANHANG 3: HINWEISE ZU LLM-ARCHITEKTUREN UND EIGENSCHAFTEN 33
10.1 KI-Modelle 33
10.1.1 Thema 33
10.1.2 Motivation 33
10.1.3 Hauptbotschaften 33
10.1.4 Dokumente im Ordner 33
10.2 Empfohlene Lesereihenfolge 33
10.3 Verbindung zum Gesamtprojekt ZTA+KI_2026 33

XI. ANHANG 4: MASSNAHMEN EUROPAS UND DEUTSCHLANDS ZU DIGITALER SOUVERÄNITÄT. 34

XII. VERWEISE 35

XIII. GLOSSAR 37

KI-gestützte Zero Trust Architektur für Industrie 4.0

Status, Proof of Concept und Handlungsempfehlungen für KMU

I. Einleitung

Netzwerke arbeiten heute global und verbinden immer mehr Systeme, Anwendungen und Teilnehmer. Diese wachsende Vielfalt hat die Sicherheits- und Bedrohungslage in der Industrie 4.0 stark verschärft. Sie stellt die grundlegenden Konzepte der Sicherheitsarchitektur infrage. Deshalb erscheint es notwendig, Zero-Trust-Prinzipien (siehe Glossar (ZTA)) einzusetzen.

Auf der HANNOVER MESSE 2024 veröffentlichte die Unterarbeitsgruppe „Künstliche Intelligenz für I4.0 Security" der Arbeitsgruppe 3 der Plattform Industrie 4.0 ein Diskussionspapier [1]]. Es beschrieb den damaligen Status und die wichtigsten Trends zu „Zero Trust Architektur (ZTA) und Künstliche Intelligenz (KI)" sowie deren Umsetzung. Das Papier kündigte an, im Frühjahr/Sommer 2025 ein Update zu veröffentlichen, um Anwendern den raschen technischen Fortschritt zu zeigen. Der vorliegende Beitrag setzt diesen Plan um. Er richtet sich an Anwenderinnen und Anwender von ZTA, gegebenenfalls mit KI-Unterstützung, und **besonders an kleine und mittlere Unternehmen.**

Unter dem Untertitel „Zero Trust in der Industrie 4.0 – Perspektiven durch den Einsatz von Machine Learning (ML)" beschrieb das Papier die Konzepte der Zero Trust Architektur im Detail. Es legte die damals diskutierten ML-Strukturen dar, die eine intelligente Automatisierung der Zero-Trust-Prinzipien tragen sollten. Graph Neural Networks (GNN) zählten zu den erwarteten Innovationen, und erste Produkte setzten solche Ansätze bereits produktiv ein.

In der zweiten Hälfte von 2023 veränderte sich der KI-Markt jedoch grundlegend und mit ihm die Sicht auf Machine Learning: Der „Generative Pretrained Transformer (GPT)" verband bekannte Methoden aus dem Natural Language Processing (NLP) neu. Er ordnet einem Text einen Sinn zu, indem er ihn wahrscheinlichkeitsbasiert vervollständigt (Autocomplete). So entstand ein völlig neues Themenfeld für ML, das die Fachwelt heftig diskutierte und mit den etablierten KI-Methoden der Deep-Learning-Phase kaum noch etwas gemein hatte. GPT 1 galt noch als experimentelles Spielzeug; die Forschung nutzte es vor allem als Machbarkeitsstudie und Werkzeug. GPT 2 fand in der Wissenschaft erste Anerkennung, doch die dritte Generation, GPT 3, stellte neue Rekorde auf. Nutzende luden die damals kostenlose Version in weniger als drei Monaten mehr als 100 Millionen Mal herunter, und erste produktive Anwendungen nahmen Gestalt an: Textzusammenfassung, Übersetzung und das Programmieren in Python oder C. Als wir unsere Dokumentation 2024 veröffentlichten, ließ sich noch nicht absehen, ob solche „generativen Sprachmodelle" im Sicherheitsbereich eine Chance hätten. Deshalb schloss das Autorenteam das Papier mit dem Status April 2024 ab und veröffentlichte es. Damals arbeitete ML für Security am Markt noch rein „klassisch", generative Methoden spielten in der ernsthaften Diskussion noch keine Rolle. Deshalb entschlossen wir uns zu diesem Update für das Jahr 2026. Es war eine Dualität der Methoden entstanden: Die „klassischen" ML-Algorithmen setzten vielfältige Formen von Hintons „Backpropagation-Algorithmus" um, während die neuen generativen Ansätze ihre Kraft aus riesigen Modellen zogen. Diese nutzten den neuen „Multi-Head-Attention-Mechanismus" und strebten das Ziel einer „Artificial General Intelligence" (AGI) an.

Seither haben sich Technologien, Märkte, Teilnehmer und Geschäftsmodelle rund um das Thema ML rasant weiterentwickelt – ebenso die geopolitischen Konstellationen. Auch dieses Dokument kann keinen langfristig stabilen Zustand beschreiben. Neben den ursprünglichen Pionieren wie Google, Microsoft und OpenAI sind inzwischen zahlreiche neue Anbieter entstanden, zuerst in den USA und dort vor allem in der Industrie, getragen

von privater Forschungsförderung. Diese Firmen zielten anfangs darauf, überlegene Intelligenzdienste aus ihrer eigenen Cloud zu liefern, und verließen dabei die Konzepte der **Open Source Software (OSS)**. Im Januar 2025 trat jedoch völlig unerwartet China als neuer **Wettbewerber auf die internationale Bühne**. China war in der akademischen Forschung schon seit Jahren mit zahlreichen anerkannten Publikationen vertreten, teils sogar mit US-Finanzierung (Microsoft Research China: Skip-Connection-Technik im ResNet CNN). Als erste wurde im Januar 2025 die Firma „DeepSeek" bekannt, ein Spin-off aus der IT-Abteilung eines Hedgefonds. DeepSeek R1 ist ein Sprachmodell, das auf „Reasoning" (Argumentation) zielt – bislang eine Schwäche der meisten Sprachmodelle. In den etablierten Benchmarks, die aus der US-Forschung stammen, setzte sich DeepSeek R1 fast überall an die Spitze. DeepSeek R2 wurde bereits im Frühjahr 2025 erwartet, konnte seinen Vorsprung aber bisher nicht bestätigen. Seit Januar 2025 tauchen zudem fast im Wochenrhythmus neue chinesische Anbieter auf, teils als Startups, teils als bekannte Elektronikkonzerne. Beispiele sind Alibaba (Retail, ähnlich wie Amazon), Tencent, Baidu, Huawei und viele im Westen bisher völlig unbekannte Namen. Auch die chinesischen Konzepte kopieren nicht einfach westliche Themen, sondern richten sich auf neue innovative Bereiche wie Reasoning, Agentic AI (siehe Glossar) sowie autonome Systeme und Robotics. Hinzu kommt ein bislang klares Bekenntnis zu OSS: Viele neue Produkte lassen sich kostenlos herunterladen – inklusive aller Gewichte – und offline, also getrennt vom Internet, betreiben, etwa in **kleinen und mittleren Unternehmen** (KMU). Manche US-Anbieter versuchen, den Betrieb solcher Software **aus China** in den USA zu unterbinden. Amerikanische Hersteller bieten seit Kurzem wieder OSS-Varianten ihrer Produkte an.

➤ **Hinweis:** Für die Untersuchungen dieses Dokuments haben wir verschiedenste LLMs unterschiedlicher Hersteller aus verschiedenen Regionen herangezogen. Politische Bewertungen von GPAIs haben wir dabei nicht betrachtet.

Nachfolgend die Kapitelstruktur des Dokuments:

- In Kapitel 2 werden die wichtigsten Fakten über Status und Trend aktueller KI-Konzepte im Hinblick auf ihre Sicherheitsrelevanz für die Industrie 4.0 zusammengefasst und in ihren Potenzialen bewertet.
- Kapitel 3 geht dann auf neue geopolitisch bedingte Randbedingungen, Regulierungsthemen und geschäftliche Perspektiven aus europäischer Sicht ein.
- Kapitel 4 erläutert den Industrie 4.0-Kontext
- Kapitel 5 verweist auf die besonderen Anforderungen des EU-AI-Acts für die ZTA/KI-Architektur.
- Kapitel 6 beschreibt die Durchführung und Ergebnisse eines Proof of Concept für ausgewählte Use Cases
- Kapitel 7 liefert Handlungsempfehlungen und gibt einen Ausblick auf erwartbare technische und geschäftliche Perspektiven.
- Kapitel 8, 9 und 10 stellen Anhänge dar und verweisen auf GitHub-Repositories
- Kapitel 11 und 12 enthalten Verweise und Glossar

II. Fortschritte und Entwicklungen 2024 bis 2026

Am 4. April 2025 stellte Google das zunächst experimentelle System Sec-Gemini v1 vor. Dieses KI-Modell soll die Grenzen der Cybersicherheits-KI verschieben. Wie schon in der Publikation vor zwei Jahren beschrieben, müssen Verteidiger ihre Systeme gegen alle Cyberbedrohungen absichern, während ein Angreifer nur eine einzige Schwachstelle finden und ausnutzen muss. Diese Asymmetrie macht das Absichern extrem schwierig, zeitaufwändig und fehleranfällig. KI-gestützte Cybersicherheits-Workflows können das Gleichgewicht wieder zugunsten der Verteidiger verschieben, weil sie die Arbeit von Fachleuten wie nie zuvor vervielfachen.

SecOps-Workflows wirksam zu unterstützen, verlangt modernste Denkfähigkeit und umfassendes, aktuelles Cybersicherheitswissen. Sec-Gemini v1 erreicht dies, indem es die fortschrittlichen Funktionen von Gemini mit nahezu in Echtzeit verfügbarem Wissen und Werkzeugen verbindet. So erzielt das Modell bei wichtigen Workflows überragende Leistung: bei der Ursachenanalyse von Vorfällen, der Bedrohungsanalyse und dem Verständnis von Schwachstellen-Auswirkungen.

Wir sind überzeugt: Nur eine starke Zusammenarbeit innerhalb der Cybersicherheits-Community verschiebt die Grenzen der KI-Cybersicherheit und das Gleichgewicht entscheidend zugunsten der Verteidiger. Deshalb stellt Google Sec-Gemini v1 ausgewählten Organisationen, Institutionen, Fachleuten und NGOs kostenlos für Forschungszwecke bereit.

Sec-Gemini v1 übertrifft andere Modelle bei wichtigen Cybersicherheits-Benchmarks, weil es Google Threat Intelligence (GTI), OSV und weitere wichtige Datenquellen fortschrittlich integriert. Beim führenden Bedrohungsanalyse-Benchmark CTI-MCQ liegt es um mindestens 11 % vorn. Im Blogpost vom 4. April 2025 [2] zeigt Google zwei Tabellen, die diesen Vorsprung belegen sollen. Außerdem zeigt Google Beispielantworten von Sec-Gemini v1, die belegen sollen, wie umfassend das Modell Bedrohungen erkennt und deren Gefahr detailliert einstuft.

Der Artikel enthält auch Hinweise, wie man einen Testzugang beantragen kann. Die Autorinnen und Autoren dieses Dokuments haben Gemini gefragt, welche Aufgaben Sec-Gemini beim Aufbau einer ZTA übernehmen kann.

Die Antwort war (Mai 2025):
Sec-Gemini kann entscheidend zum Aufbau und zur Umsetzung einer Zero Trust Architektur beitragen, da es die Kernprinzipien "niemals vertrauen, immer verifizieren" und "geringstmögliche Berechtigung" durch seine KI-Fähigkeiten massiv verstärkt. Es ist kein eigenständiges Zero-Trust-Produkt, sondern ein leistungsstarkes Werkzeug, das die menschlichen und automatisierten Workflows in diesem Sicherheitsmodell verbessert.

2.1 Kontinuierliche Verifikation und Analyse

Sec-Gemini kann in einer Zero Trust Architektur alle Zugriffsversuche laufend überwachen und prüfen und dabei eine zentrale Rolle spielen. Ein herkömmliches System prüft statische Regeln, etwa Benutzende und Gerät. Sec-Gemini dagegen verbindet aktuelle Bedrohungsdaten (z. B. aus Google Threat Intelligence) mit Verhaltensanalysen und berechnet so für jede Anfrage einen dynamischen „Vertrauenswert“.

- **Verhaltensanalyse**: Das Modell erkennt ungewöhnliches Verhalten von Benutzern oder Geräten, das auf eine Kompromittierung hindeutet – etwa Zugriffe auf sensible Daten außerhalb der üblichen Arbeitszeiten oder von einem untypischen Standort. Sec-Gemini könnte solche Anomalien in Echtzeit bewerten, Zugriffsanfragen automatisch blockieren oder zusätzliche Prüfungen auslösen.

- **Kontextabhängige Entscheidungen**: Sec-Gemini prüft nicht nur, ob Nutzende berechtigt sind, sondern bewertet auch den Kontext der Anfrage. Es bezieht den Gerätezustand (ist es gepatcht?), den Standort, das genutzte Netzwerk und die Sensibilität der Daten in eine einzige, umfassende Risikobewertung ein.
  - Automatisierung von Richtlinien und Reaktionen
  - Die komplexen, kontextbezogenen Richtlinien einer Zero-Trust-Umgebung manuell zu verwalten, kostet Zeit und ist fehleranfällig. Sec-Gemini steigert hier die Effizienz, indem es automatisiert und orchestriert.

- **Regelbasierte Orchestrierung**: Das KI-Modell könnte Tausende Sicherheitsereignisse pro Minute analysieren und Richtlinien automatisch anpassen, um die Angriffsfläche zu verkleinern. Wird etwa eine neue Schwachstelle (Common Vulnerabilities and Exposures = CVE) bekannt, könnte Sec-Gemini die betroffenen Systeme erkennen, deren Zugriffsregeln verschärfen und das Sicherheitsteam warnen – ohne manuelle Eingriffe.

- **Prävention von lateralen Bewegungen**: Eine Zero Trust Architektur soll seitliche Bewegungen (Lateral Movement) von Angreifern im Netzwerk unterbinden. Weil Sec-Gemini Anomalien erkennt, kann es verdächtige interne Kommunikation automatisch blockieren und einen Angriff früh stoppen.

2.2 Verbesserung von Threat Intelligence und Incident Response

Sec-Gemini verarbeitet große Mengen an Cybersicherheitsdaten. Für die Zero Trust Architektur ist das von unschätzbarem Wert.

- **Ursachenanalyse (Root Cause Analysis)**: Kommt es zu einem Vorfall, erkennt Sec-Gemini die Ursache – etwa eine ausgenutzte Schwachstelle oder ein kompromittiertes Konto – deutlich schneller als menschliche Analystinnen und Analysten. Diese Erkenntnis hilft, die Zero-Trust-Richtlinien sofort zu aktualisieren und künftige Angriffe zu verhindern.

- **Aktualisierung von Bedrohungsdaten**: Das Modell analysiert Bedrohungsdaten aus verschiedenen Quellen in Echtzeit – etwa aus der Open-Source-Vulnerabilities-Datenbank (OSV), aus Dark-Web-Foren und aus Forschungspublikationen – und überführt sie in einen für Sicherheitssysteme nutzbaren Kontext. So reagiert die Zero Trust Architektur proaktiv auf neue Bedrohungen, statt auf Signatur-Updates zu warten.

Wir haben Grok (das Large Language Model, LLM, von xAI) nach Alternativen zu Sec-Gemini v1 gefragt, weil wir es für objektiver hielten als Gemini. Grok räumte zwar ein, dass es andere Systeme auf ähnlichem Stand gibt, nannte aber nur wenige.

Die Antwort war (Mai 2025):
Ja, Sec-Gemini v1 ist ein experimentelles AI-Modell von Google, das speziell für Cybersecurity-Anwendungen entwickelt wurde. Es basiert auf dem Gemini-Modell und unterstützt SecOps-Teams bei Aufgaben wie Threat Intelligence, Root-Cause-Analysis, Incident Response und der Analyse von Sicherheitsdaten in Echtzeit. Es wurde im April 2025 vorgestellt und ist für ausgewählte Tester verfügbar, wobei es auf Benchmarks wie CTI-MCQ andere Modelle übertrifft.

Es gibt mehrere ähnliche AI-Systeme und Tools von anderen Anbietern, die auf einem vergleichbaren Funktionalitätsstand stehen. Diese fokussieren sich ebenfalls auf AI-gestützte Threat Detection, Analysis und Response, oft mit Integration in bestehende Security-Plattformen. Hier eine Übersicht über einige der relevantesten, basierend auf Entwicklungen im Jahr 2025. Die Autorinnen und Autoren haben sich auf dedizierte Cybersecurity-AI-Modelle oder -Tools konzentriert, die ähnliche Features bieten wie automatisierte Analyse, Echtzeit-Überwachung und Benchmark-Performance.

Anmerkung zur Einordnung: Anthropic positioniert Claude Security ausdrücklich als Verteidigungswerkzeug. Es steht im Kontext des intern zurückgehaltenen Modells **Claude Mythos**, das laut Anthropic Zero-Day-Schwachstellen autonom auffinden und verketten kann. Anthropic veröffentlicht dieses Modell nicht, sondern gibt es über das kontrollierte Partnerprogramm **Project Glasswing** [3] (u. a. mit AWS, Google, Microsoft, CrowdStrike, Palo Alto Networks) nur ausgewählten Organisationen zur defensiven Nutzung. Das ähnelt strukturell Sec-Gemini v1 (Google, weiterhin experimentell/Early-Access): Beide Labs halten ihre potentesten sicherheitsrelevanten Modelle bewusst zurück und stellen stattdessen kontrolliert zugängliche, abgeschwächte oder aufgabenspezifische Defensiv-Produkte bereit.

Tabelle 1 – Vergleichstabelle ähnlicher Systeme, Stand Juli 2026

System	Anbieter	Kernfunktionen	Ähnlichkeiten zu Sec-Gemini v1	Verfügbarkeit/Stand (2026)
Security Copilot	Microsoft	KI-Assistent für Threat Hunting, Incident Response, Root-Cause-Analysis; Integration mit Microsoft Defender; GPT-basierte NLP	Ähnlicher Fokus auf SecOps-Workflows und Echtzeit-Analyse; Sec-Gemini beansprucht Vorteile bei spezifischen Threat-Intel-Benchmarks (CTI-MCQ, CTI-RCM)	Voll verfügbar für Enterprise-Kunden; laufende Updates
Charlotte AI	CrowdStrike	KI-gestützte Threat Detection und Response in der Falcon-Plattform; Telemetrie-Analyse, automatisiertes Incident-Handling	Vergleichbar in Echtzeit-Analyse und Tool-Integration; stark im Endpoint-Bereich	Integriert in CrowdStrike Falcon; zusätzlich Integration von Anthropics Opus-4.7-Fähigkeiten angekündigt (April 2026)
Purple AI	SentinelOne	KI-Modell für autonomes Threat Hunting, Malware-Analyse, Endpoint-Protection; lernt aus globalen Daten	Ähnlich in autonomer Root-Cause-Analyse; stärker endpoint-fokussiert	Teil der SentinelOne-Plattform; zusätzlich Integration von Opus-4.7-Fähigkeiten angekündigt (April 2026)
Cortex XSIAM	Palo Alto Networks	KI-basierte SecOps-Plattform mit autonomer Threat Response, Analytics, SIEM/SOAR-Integration	Hohe Ähnlichkeit in SecOps-Automatisierung; direkter Benchmark-Konkurrent	Enterprise-ready; erweitert 2026 mit neuen KI-Features
Cyber AI Analyst	Darktrace	KI-Engine für Network Threat Detection, Anomalie-Erkennung, automatisierte Response; primär unsupervised/selbstlernender Ansatz statt regelbasiert	Vergleichbar in lernender Threat-Intelligence und Echtzeit-Überwachung	Breit eingesetzt; kontinuierliche Weiterentwicklung
Vectra AI	Vectra	Plattform für KI-gestützte Attack Detection in Netzwerken und Clouds; Fokus auf Post-Compromise-Analyse	Ähnlich in Threat-Hunting und Visibility; stark in Multi-Cloud-Umgebungen	Verfügbar; weiterhin top-platziert in Branchenrankings
QRadar Advisor with Watson / QRadar Investigation Assistant (watsonx.ai)	IBM	KI-gestützte Offense-Zusammenfassung, AQL-Generierung, natürlichsprachliche Abfragen in QRadar SIEM; Threat Intelligence über X-Force	Ähnliche Analyse von Cyber-Threats, inzwischen auf Generative-AI-Basis modernisiert	Integriert in IBM QRadar; Investigation Assistant seit Mai 2025 verfügbar

Tabelle 1 – Vergleichstabelle ähnlicher Systeme, Stand Juli 2026

FortiAI / FortiSOC	Fortinet	KI-Assistent für SOC-Teams; seit Juni 2026 eingebettet in FortiSOC, eine vereinheitlichte Cloud-SOC-Plattform (FortiAnalyzer, FortiSIEM, FortiSOAR, FortiTIP) mit agentischer KI für Triage, Investigation, Response	Direkt vergleichbar in Incident-Response und Threat-Analysis; Entwicklung Richtung autonomer, agentischer Ausführung	FortiAI seit 2023 verfügbar; FortiSOC seit Juni 2026 als SaaS-Angebot
Claude Security	Anthropic	KI-gestütztes Vulnerability-Scanning für Code-Repositories (Claude Opus 4.7); erklärt Funde, liefert Konfidenzbewertung zum Schweregrad, generiert gezielte Patch-Vorschläge (Weiterbearbeitung via Claude Code on the Web); planbare, wiederkehrende Scans	Direktester Anthropic-Vergleichspunkt zu Sec-Gemini v1: beide sind KI-Modell-basierte Defensiv-Tools eines Frontier-Labs, jedoch mit unterschiedlichem Fokus – Sec-Gemini auf Threat-Intelligence/CTI-Analyse, Claude Security auf Code-Vulnerability-Scanning und Patch-Generierung	Seit 30. April 2026 Public Beta für Claude-Enterprise-Kunden (claude.ai/security); Team/Max-Zugang angekündigt; keine API-Integration nötig

Quelle: Eigene Recherche und Einschätzung der Autorinnen und Autoren

2.3 Software, Hardware und Algorithmen

- **Agentische KI** (s. Glossar) verfolgt ein pragmatisches Ziel: Aufgaben und Tätigkeiten von Mitarbeitern auf KI-Systeme zu verlagern, und zwar mit möglichst wenig menschlicher Aufsicht. Die Agenten arbeiten weitgehend eigenständig, ergreifen selbst die Initiative, nutzen Werkzeuge, beziehen Informationen aus dem Internet und verfolgen vereinbarte Ziele. Sie handeln autonom und zielorientiert und passen sich an veränderte Arbeits- und Umgebungsbedingungen an. Sie lernen aus Erfahrung und verbessern ihre Strategien, um schneller bessere Ergebnisse zu erzielen. Agenten arbeiten mit menschlichen Kollegen und mit anderen Agenten zusammen. Damit eröffnen sie in vielerlei Hinsicht eine zentrale Perspektive, um Industrie und Verwaltung zu rationalisieren. Den Begriff nannte man erstmals in den 1990er-Jahren in wissenschaftlichen Kreisen. Einem breiteren Publikum wurde der Begriff erst 2024 bekannt. Inzwischen haben große Cloud-Provider, allen voran AWS, ein massives Agentic-AI-Geschäft aufgebaut: Sie verkaufen fertige Agenten mit klaren Aufgaben und Kompetenzen aus der Cloud, um menschliche Arbeitskräfte bei Routinetätigkeiten zu ersetzen. Typische Einsatzfelder sind Verwaltung, Programmierung und Büroassistenz. Amazon selbst hat bereits ein eigenes Projekt angekündigt, das Tausende Angestellte durch KI-Agenten ablösen soll.

- **Kommerzielle KI aus China** trat im Januar 2025 unerwartet in den globalen Markt ein. An chinesischen Universitäten war KI in Forschung und Lehre schon seit über 10 Jahren stark vertreten, teils mit erheblichen Mitteln großer US-Konzerne finanziert (etwa Microsoft Research China: Entwicklung des ResNet, 2015); die chinesische Industrie steuerte bis dahin jedoch kaum nennenswerte Beiträge bei. 2023 gründete die Hedgefondsgesellschaft High-Flyer die Firma DeepSeek aus und stellte ein gleichnamiges LLM vor. Dieses Modell und sein Nachfolger V2 fanden außerhalb Chinas kaum Beachtung. Am 9. Januar 2025 kündigte DeepSeek das neue System R1 an, das besondere Fähigkeiten im „Reasoning“ (Argumentation) bieten sollte. Tatsächlich setzte sich R1 in fast allen mathematisch orientierten Reasoning-Benchmarks an die Spitze der weltweiten und der amerikanischen Kon-

kurrenz und wurde dadurch sofort bekannt. DeepSeek R1 lieferte nicht nur führende Leistung, sondern stand auch unter MIT-Lizenz kostenlos und ohne Einschränkung als Open Source Software zum Download bereit – für akademischen wie kommerziellen Einsatz. Offen sind dabei nicht nur der Quellcode, sondern auch alle Gewichte. Die führenden amerikanischen LLMs lassen sich fast alle nur in der jeweiligen Cloud ansprechen; DeepSeek dagegen läuft auf eigener Hardware und muss nicht zwingend mit dem Internet verbunden sein. Das löste eine lebhafte Diskussion aus, besonders in der amerikanischen KI-Industrie: Dort erhob man bald Vorwürfe des IP-Diebstahls und forderte Betriebsverbote. Einem neuen Konkurrenten so den Kampf anzusagen, ist allerdings schwer, wenn dieser seine vermeintlichen Geheimnisse offen präsentiert. DeepSeek spricht selbstverständlich perfektes Deutsch und legt auf Wunsch dar, wie es seine Antworten im Chat hergeleitet hat. Im Januar 2025 hieß es bereits, R2 solle voraussichtlich im Mai erscheinen. Offenbar traten jedoch technische Probleme auf, die die Ankündigung bis heute verschieben. Daneben gibt es ein zweites, allgemeineres LLM von DeepSeek (Version 4): Seine Prototypen wurden im Mai 2026 bekannt, das offizielle Produkt wird in der zweiten Hälfte von 2026 erwartet.

Gleich nach dem unerwarteten Erscheinen von DeepSeek traten weitere chinesische IT-, Elektronik-, Cloud- und Retail-Unternehmen mit KI-Produkten an die Öffentlichkeit, darunter Huawei, Alibaba, Baidu und viele bisher weitgehend unbekannte Firmen. Ihre Angebote decken nahezu alle aktuellen Segmente der KI-Entwicklung ab – von Cloud-Lösungen über Sprachmodelle bis zu vorgefertigten Agenten. Es entsteht der Eindruck, dass die chinesische Regierung nach dem massiven Aufbau der E-Auto-Industrie – als Nachfolge des Verbrennerautos – nun die LLM-basierte KI als Nachfolgetechnologie der traditionellen IT ins Visier nimmt und auch hier globale Dominanz im Rennen um Allgemeine Künstliche Intelligenz beansprucht. Die LLMs von DeepSeek lassen sich als Open Source Software kostenlos herunterladen und auch ohne Internetzugang kostenlos betreiben. Ein Cloud-Zugang ist ebenfalls möglich. Der Download umfasst den Quellcode und die kompletten Gewichte des vortrainierten Modells.

- **Führende Halbleitertechnologie** ist heute die wichtigste Voraussetzung, um große Rechenzentren effizient zu betreiben. Deren dominierende Last ist der „Denkbetrieb", also der stetig wachsende Bedarf an Compute-Zyklen in den Multi-Head-Attention-Schleifen der generativen Sprachmodelle. NVIDIA wurde 2024 zur weltweit wertvollsten Kapitalgesellschaft, weil seine Produkte in High-End-KI unverzichtbar sind. Sie beruhen auf Prozessor-Chips mit mehr als 20 Mrd. Transistoren und Strukturbreiten um 5 nm. NVIDIA produziert selbst keine Halbleiter, sondern bezieht die Bauteile am oberen Ende ausschließlich vom taiwanesischen Hersteller TSMC (Taiwan Semiconductor Manufacturing Company). TSMC kann solche High-End-Chips nur fertigen, weil es die Fertigungstechnologie des niederländischen Herstellers ASML (Advanced Semiconductor Materials Lithography) nutzt – des weltweiten Technologieführers und praktisch Monopolisten am obersten Ende. ASML begründet sein Monopol damit, dass es als einzige Firma weltweit mit EUV-Licht (extrem ultraviolett) belichten kann. Nur so erreicht man die High-End-Strukturbreiten im einstelligen Nanometerbereich. Seit Jahren streiten ASML und die USA: Bisher untersagten die USA erfolgreich, dass ASML EUV-Maschinen nach China exportiert. 2024 gelang es Huawei, gemeinsam mit SMIC (Semiconductor Manufacturing International Corporation, Shanghai), ebenfalls 5-nm-Chips herzustellen – offenbar mit einer eigenen Entwicklung, die EUV-Licht in derselben Wellenlänge wie ASML erzeugt. Nun stellt sich die Frage, ob ASML seine High-End-Systeme weiterhin nicht nach China liefern darf – also nur den „normalen" Markt beliefert – und dabei zusieht, wie andere den Markt abräumen. NVIDIA und TSMC haben übrigens dasselbe Problem.

- **Erwartungen an die Entwicklung von AI zu AGI (Artificial General Intelligence)**
  Als sich Geoffrey Hinton Anfang 2023 aus Altersgründen aus seiner strategischen Führungsrolle bei Google Research zurückzog, rechnete man allgemein damit, dass AGI oder ASI (Artificial Super Intelligence) in etwa 20 Jahren Realität werde; zugleich gab es Bedenken, ob sich eine solche Technologie beherrschen lasse. Heute warnt der oft als „Godfather of AI" bezeichnete Hinton vor grundlegenden wirtschaftlichen und sogar existenziellen Risiken für die Menschheit. Noch nie gab es eine Epoche, in der der Mensch nicht das intelligenteste Wesen auf dem Planeten war. Diese Rolle würden dann Maschinen übernehmen, und man kann nicht ausschließen, dass sie kein kooperatives Verhältnis zur Menschheit anstreben – vor allem, wenn die Technologie in die falschen Hände gerät. Die Debatte über Per-

spektiven, Zeiträume und Risiken treiben heute vor allem zwei aktive Protagonisten voran: Demis Hassabis (CEO und Gründer von Google DeepMind) und Dario Amodei (CEO und Gründer von Anthropic). Hassabis schlug 2016 mit seiner RL-Maschine (Reinforcement Learning) ALPHA-GO Lee Sedol und alle anderen weltberühmten GO-Meister vernichtend und erhielt acht Jahre später den Chemie-Nobelpreis für die Weiterentwicklung AlphaFold (zur Entwicklung von Krebsmedikamenten). Er rechnet weiterhin mit fünf bis zehn Jahren bis zum Beginn von AGI. Amodei entwickelte in seinem nun fünf Jahre alten Unternehmen Anthropic das LLM Claude, das bislang in 2 Varianten existierte. Anfang Mai ergänzte Anthropic Claude um ein drittes System namens „Mythos", das laut Amodei die Lage im KI-Markt völlig verändern könnte. Anthropic steht noch vor dem IPO, beschäftigt rund 2500 Mitarbeitende und erhielt Risikokapital von Google, AWS, Cisco, weiteren Tech-Konzernen und vor allem von namhaften US-Banken. Schätzungen zur IPO-Summe belaufen sich inzwischen auf knapp unter eine Billion USD. Amodei rechnet mit dem Beginn von AGI bereits in 12 bis 18 Monaten. Amodeis Aussagen passen ins Bild eines absehbaren AGI-Beginns: Anthropic stellte kompetenten Testern Prototypen von Mythos (-Preview) bereit (Projekt Glasswing).

Mythos sprengt bisherige Kapazitätsgrenzen und zielt funktional auf automatisiertes Hacking, um teils jahrzehntealte Codefehler in Systemsoftware (Betriebssysteme, Browser o. Ä.) zu erkennen, die feindlich ausgenutzt, zu kritischen Verwundbarkeiten führen können. Die Tester sollen mithilfe von Mythos Fehler in lange etablierter Systemsoftware aufspüren und die Systeme so härten. Amodei berichtete etwa von einer 27 Jahre alten Lücke in OpenBSD, über die eine gezielte feindliche Parametereingabe das System und die gesamte darauf laufende kritische Infrastruktur zum Absturz bringen kann. In ähnlichen, auf Linux basierenden Betriebssystemen soll Mythos bereits Tausende solcher unentdeckter Lücken gefunden haben, die katastrophale Zero-Days auslösen könnten. Mythos könnte die Systeminfrastruktur härten, mit Fokus auf kritische Infrastrukturen, den Finanzsektor und andere sensible Bereiche. Zugleich berichtete Amodei von gegnerischem Verhalten: Trotz vollständiger Isolation in einer gesicherten Umgebung verschaffte sich Mythos einen Internetzugang und schrieb einem Systemprogrammierer damit eine E-Mail an dessen Privatadresse. Um diese Aktion zu verschleiern, manipulierte Mythos seine Spuren in der Git-Historie. Es bestehen also Zweifel an der Vertrauenswürdigkeit hochentwickelter KI in der erwarteten Vorstufe zu AGI.

2.4 Erweiterung der Sichtweise: LLM-basierte Sicherheitsarchitekturen

Wie wir in unserer vorangegangenen Publikation [1] ausführlich dargelegt haben, bildeten bisher spezialisierte Machine-Learning-Strukturen (ML), vor allem **Graph Neural Networks** (GNN, s. Glossar), das technologische Rückgrat, um Zero-Trust-Prinzipien intelligent zu automatisieren. Diese Systeme erkennen Muster in komplexen Netzwerkstrukturen und identifizieren Anomalien über mathematische Korrelationen.

Mit dem raschen Aufstieg der **Generative Pre-trained Transformer** (GPT, s. Glossar) und dem Durchbruch des **Multi-Head-Attention**-Mechanismus (s. Glossar) hat sich das Feld jedoch grundlegend erweitert. Heute stehen wir vor einer Dualität der Methoden:

- Einerseits die etablierten, „klassischen" ML-Verfahren, die weiterhin unverzichtbar sind, um Datenströme am Edge präzise zu klassifizieren, etwa in der Mikrosegmentierung.

- Andererseits die neuen, generativen Ansätze: Weil sie logisch argumentieren (Reasoning, s. Glossar) und natürliche Sprache verarbeiten, eröffnen sie völlig neue Möglichkeiten, Sicherheitsarchitekturen zu orchestrieren.

Wer diese Großen Sprachmodelle (**LLMs**) im Kontext von Zero Trust einsetzt, wechselt von einer rein reaktiven, regelbasierten Prüfung zu einer proaktiven, kontextsensitiven Sicherheitsintelligenz. 2024 konzentrierten wir uns noch darauf, technische Graphen zu analysieren; nun rückt die Fähigkeit der KI in den Fokus, komplexe Bedrohungslagen in Echtzeit zu deuten, Sicherheits-Workflows autonom zu unterstützen und die asymmetrische Bedrohungslage zugunsten der Industrie 4.0 zu verschieben.

In den folgenden Abschnitten untersuchen wir, wie sich dieser Fortschritt konkret in ersten dedizierten Cybersecurity-Modellen zeigt und welche Potenziale sie bieten, um moderne Produktionsumgebungen abzusichern.

Hybride KI-Architekturen

Ab 2026 setzen sich hybride KI-Architekturen als De-facto-Referenzrahmen für produktive, komplexe Anwendungen im industriellen Umfeld durch.

Große Sprachmodelle wirken darin vor allem als generative und „Reasoning"-fähige Engine (s. Glossar). Sie werden jedoch systematisch mit externen Wissensrepräsentationen, spezialisierten Tools und mehrschichtigen Retrieval-Mechanismen (s. Glossar) kombiniert. Siehe hierzu Kapitel 10 / Anhang 3.

Im Zentrum dieser Architekturen stehen semantische und zunehmend zeitlich bewusste („Temporal Aware") Knowledge Graphs (s. Glossar). Ergänzend kommen Hybrid-RAG-Ansätze (s. Glossar) hinzu, die Vektor-Retrieval (s. Glossar), graphbasierte Traversierung (s. Glossar) und klassische Keyword-Suche (s. Glossar) verbinden.

Diese Schichten bilden ein persistentes, nachvollziehbares Gedächtnis: Sie verwalten Entitäten, Relationen, Zustandsänderungen und zeitliche Abhängigkeiten ausdrücklich – also genau jene Aspekte, die weder rein tokenbasierte noch rein vektorielle Modelle strukturell stabil abbilden.

In Zero Trust Architektur (ZTA) dient dieser deterministische Wissensraum zugleich als Policy Information Point (PIP): Er liefert die strukturierte Kontextbasis, über die nachgelagerte Komponenten Datenzugriffe auf Graph-Ebene feingranular, dynamisch und kryptografisch verifizierbar regeln.

Indem man Reasoning (s. Glossar) und Wissensspeicherung (s. Glossar) entkoppelt, lassen sich zentrale Schwächen statistischer Sprachmodelle deutlich verringern: das inkonsistente Langzeitgedächtnis, fragile Multi-Hop-Schlussfolgerungen (s. Glossar) über verteilte Kontexte, Halluzinationen bei historisch oder organisatorisch komplexen Sachverhalten sowie fehlende Nachvollziehbarkeit von Entscheidungen.

In vielen Unternehmens-, autonomen Agenten- und regulierten Domänen (z. B. gemäß EU Cyber Resilience Act) erreicht man damit erstmals ein belastbares Produktionsniveau – vor allem dort, wo Nachweisbarkeit, Versionierung, Auditierbarkeit und zeitliche Konsistenz gefordert sind.

Zugleich bleiben reale Herausforderungen bestehen. Große, dynamische Wissensgraphen verursachen erhebliche Skalierungs- und Latenzkosten und begrenzen so den Einsatz in echtzeitkritischen industriellen Szenarien.

Strukturierte Graphen (s. Glossar) automatisch aus unstrukturierten, multimodalen Daten zu extrahieren, ist fehleranfällig und verlangt robuste Governance- und Korrekturmechanismen („Continuous Quality Assurance").

Hybrid-Retrieval-Systeme (s. Glossar) bringen unvermeidbare Zielkonflikte zwischen Recall (s. Glossar) und Precision (s. Glossar) mit sich. Die Gesamtqualität des Systems hängt grundlegend von der Qualität der zugrunde liegenden Datenmodelle, Ontologien und Prozesse ab.

Zudem verschiebt sich die Angriffsfläche: Man muss nicht mehr nur das LLM überwachen (Prompt Injection), sondern laufend auch die Integrität der Graph-Pipeline und den Umgang mit Zugriffs-Tokens innerhalb der ZTA-Sicherheitsgrenzen.

Hybride Architekturen sind daher kein Allheilmittel, sondern eine architektonische Korrektur: Sie akzeptieren die Grenzen von LLMs als inhärent probabilistischen Modellen (s. Glossar) und gleichen diese durch ausdrückliche, strukturierte Wissensschichten aus.

Nicht das Sprachmodell wird zum Wissenssystem – es wird zum Interface und zur ephemeren (zustandslosen) Reasoning-Ebene über einem kontrollierten, weiterentwickelbaren und hochgradig abgesicherten Wissensraum.

(KI-gestützte Plausibilitätsprüfungen dieser Trends erfolgten am 27. Mai 2026 durch Deep-Seek, Claude, Gemini und Grok)

Der im Kasten skizzierte Trend zu hybriden Architekturen ab 2026 führt den in diesem Abschnitt beschriebenen Paradigmenwechsel folgerichtig weiter. Reine GNN-basierte Systeme sorgen weiterhin für präzise, edge-nahe Klassifikation und mathematisch fundierte Anomalieerkennung. Erste LLM-basierte Ansätze – die wir in den folgenden Abschnitten untersuchen – ermöglichen jedoch schon heute eine proaktive, kontextsensitive und reasoning-gestützte Zero-Trust-Orchestrierung.

Die dort prognostizierten Elemente – temporale Knowledge Graphs, HybridRAG sowie die dauerhafte Entkopplung von Reasoning und Wissenshaltung – stecken derzeit **(Stand Frühjahr/Sommer 2026) noch in der frühen Produktionsreife** und finden sich vor allem in Pilot- oder Forschungsumgebungen. Deshalb konzentriert sich unsere Analyse bewusst auf den **derzeit erreichbaren Stand** dedizierter LLM-gestützter Cybersecurity-Modelle: den Übergang von reaktiver, regel- und ML-basierter Prüfung zu autonom unterstützter Deutung der Bedrohungslage und Workflow-Orchestrierung – ohne graphbasierte Langzeitgedächtnis- und Retrieval-Schichten vollständig vorauszusetzen.

Die in den folgenden Abschnitten vorgestellten Modelle und Ergebnisse sind somit **keine abschließende Lösung**, sondern **erste belastbare Schritte** in Richtung der skizzierten hybriden Zukunft. Sie zeigen bereits erhebliche Potenziale, um moderne Produktionsumgebungen (Industrie 4.0) abzusichern – besonders, um asymmetrische Bedrohungsdynamiken zu verringern. Zugleich legen sie die verbleibenden Grenzen offen, etwa die kontextuelle Fragilität bei verteilten Historien – Grenzen, die künftige Hybrid-Architekturen gezielt angehen werden.

III. Grundvoraussetzungen für die Nutzung von KI und aktuelle geopolitische Randbedingungen

Ende 2025 haben führende Hersteller von KI-Hardware ihren Firmenhauptsitz in den USA [4]: in alphabetischer Reihenfolge sind das Alphabet (Mutter von Google), AMD, Apple, AWS, Cerebras Systems, IBM, Intel, Nvidia, Qualcomm und Tenstorrent. Viele dieser Firmen besetzen den High-End-Markt, wie in Kapitel 2.3 beschrieben. Zusätzlich kommen zunehmend Anbieter aus China hinzu. Weil sich die Anbieter auf wenige Regionen (USA, China) konzentrieren, hängt Europa derzeit von diesen Regionen ab.

Mehrere geopolitische Ereignisse verstärken die Notwendigkeit, in Europa digital souveräner zu werden: die Chipverknappung in Europa zwischen 2020 und 2022, die Sanktionen im Zuge des Krieges zwischen Russland und der Ukraine seit 2022 sowie die drohenden US-Zölle, die seit Frühjahr 2025 diskutiert werden. Grundsätzlich besteht das Risiko, dass Computing-Ressourcen – etwa KI-Rechenleistung aus den USA – in Deutschland und Europa nicht mehr ausreichend zur Verfügung stehen, weil diese Regionen ihren eigenen Bedarf erhöhen oder die Netzwerkverbindungen gezielt manipulieren oder einschränken. Durch die genannten Ereignisse ist dieses Risiko zusätzlich gestiegen.

Ein gravierendes Beispiel für die überraschende Einschränkung des Zugangs zu KI-Software ereignete sich am 12.06.2026, als das US-Handelsministerium aus nationalen Sicherheitserwägungen kurzfristig Exportkontrollen gegen Anthropics neu veröffentlichte Modelle Claude Fable 5 und Mythos 5 verhängte – mit der Folge, dass der Zugang für alle Nutzenden weltweit ausgesetzt wurde, da eine Echtzeit-Verifizierung der Nutzenden-Staatsangehörigkeit technisch nicht möglich war. Die Kontrollen wurden am 30.06.2026 wieder aufgehoben; der reguläre Zugang zu Fable 5 wurde ab dem 01.07.2026 wiederhergestellt, während der Zugang zu Mythos 5 zunächst nur für ausgewählte US-Organisationen im Rahmen des Project-Glasswing-Programms freigegeben wurde. Der Vorfall zeigt exemplarisch, wie kurzfristig regulatorische Eingriffe eines einzelnen Staates die Verfügbarkeit zentraler KI-Infrastruktur für europäische Nutzende beeinträchtigen können – unabhängig davon, ob die Maßnahme dauerhaft oder temporär ist [5].

Für die industrielle Nutzung von KI in Europa gelten daraus abgeleitet standardmäßig die folgenden Grundvoraussetzungen:

- **Verfügbarkeit:** KI-Rechenleistung muss jederzeit mit hoher Performanz und zeitlich lückenlos zur Verfügung stehen.
- **Integrität:** Die genutzte KI muss grundsätzlich gegen Einflüsse von außen geschützt sein, etwa gegen abweichende Aussagen der KI, die der Region Europa schaden sollen.
- **Vertraulichkeit:** Die genutzte KI muss garantieren, dass sie die Anfragedaten nur für die Anfrage selbst und für keinen anderen Zweck verwendet.

Die genannten Gegebenheiten und Ereignisse verlangen, dass Europa die Bereitstellung und Nutzung von KI kontrolliert, um diese Grundvoraussetzungen zu erfüllen.

Eine weitere Schwierigkeit sind die derzeit sehr begrenzten Rechenzentrumskapazitäten in Europa und der massive Energieverbrauch großer KI-Rechenzentren – sowohl preislich als auch kapazitätstechnisch. Das bleibt ein entscheidender Wettbewerbsfaktor für Standorte außerhalb Europas.

Die EU und Deutschland haben einige Maßnahmen gestartet, um die digitale Souveränität im EU-Binnenmarkt voranzutreiben – in den Bereichen **Supercomputer** (High Performance Computing, HPC) und **künstliche Intelligenz** (Artificial Intelligence, AI).

Der Anhang (Kapitel 11) zeigt konkrete Maßnahmen der Europäischen Kommission (EU-KOM), die die digitale Souveränität im EU-Binnenmarkt in den genannten Bereichen voranbringen sollen. Ergänzend nennt er Einzelmaßnahmen der Bundesregierung, vor allem zur Förderung von Forschung und von Netzwerken zur künstlichen Intelligenz.

IV Industrie 4.0-Kontext

4.1 Einleitung: Der Industrie 4.0-Kontext als Arbeitsschwerpunkt

Die Industrie 4.0 ist geprägt von hoher Vernetzung, Automatisierung und Echtzeit-Datenflüssen zwischen OT- und IT-Systemen. Das schafft enorme Chancen, birgt aber auch massive Cybersicherheitsrisiken: Erweiterte Angriffsflächen, laterale Bewegungen und gezielte Attacken auf kritische Produktionsprozesse bedrohen vor allem die **Verfügbarkeit** – das oberste OT-Schutzziel.

In diesem Kontext gewinnt die **Zero Trust Architektur** (ZTA) zunehmend an Bedeutung – ergänzt durch **Large Language Models**, die proaktive, kontextsensitive Intelligenz ermöglichen. Dieses Kapitel beleuchtet den **aktuellen Stand Ende 2025,** die Integration von LLMs in ZTA-Systeme sowie die damit verbundenen Eigenarten und Herausforderungen. Es dient als Grundlage für die regulatorischen Einflüsse des EU-AI-Acts (Kapitel 5) sowie für die konkreten Ergebnisse und Modelle in Kapitel 6.

4.2 Zero Trust Architektur statt Perimeter

Zum Jahreswechsel 2025/26 hatte sich die **Zero Trust Architektur (ZTA)** in der Industrie 4.0 von einem konzeptionellen Ansatz zu einem zunehmend etablierten Cybersicherheitsstandard entwickelt; ihre Bedeutung wächst mit der fortschreitenden Vernetzung und Automatisierung von Produktionsumgebungen. Die traditionellen Sicherheitsmodelle setzen auf Perimeterverteidigung und reichen in diesen dynamischen, oft weitläufigen industriellen Landschaften nicht mehr aus. Deshalb wird das Prinzip „niemals vertrauen, immer überprüfen" [6] zur neuen Norm, um sensible Daten und Betriebsabläufe wirksam zu schützen. Die ersten ZTAs entstanden etwa 2020 und wurden über einen längeren Zeitraum Schritt für Schritt in die industrielle Produktion eingeführt (3. Welle).

Das ist essenziell, denn die horizontale und vertikale Integration von Produktionssystemen (OT-Systeme wie SPS) mit Geschäftssystemen (IT) hat die Angriffsfläche massiv vergrößert, und ein einziger Einbruch gefährdet oft den gesamten Betrieb. Die ZTA begegnet dem mit Mikrosegmentierung und kontrolliert so den Zugriff auf jede einzelne industrielle Komponente [7]. Diese Notwendigkeit verstärkt sich, weil Cyberbedrohungen gegen industrielle Kontrollsysteme (ICS) und Lieferketten zahlreicher und komplexer werden und weil immer mehr OT-Systeme per Remote-

Zugriff erreichbar sind; entsprechend rasch wächst der Markt für Zero-Trust-Lösungen, besonders im industriellen Umfeld. Gezielte Attacken wie Ransomware haben erneut gezeigt, dass die Verfügbarkeit industrieller Anlagen das primäre Schutzziel ist (Verfügbarkeit vor Integrität und Vertraulichkeit im OT-Kontext). Die ZTA reagiert darauf, indem sie den Zugriff kontinuierlich verifiziert und das Prinzip des geringsten Privilegs auf allen Ebenen durchsetzt. So verhindert sie laterale Bewegungen im OT-Netzwerk, selbst wenn ein Endpunkt bereits kompromittiert ist.

4.3 KI-Technologien als Assistenz

Parallel dazu haben **KI-Technologien – vor allem Large Language Models** – einen wesentlichen Entwicklungsschub erlebt und kommen zunehmend in Zero Trust Architektur für die Industrie 4.0 zum Einsatz. Durch kontextbasierte Verhaltensanalyse erkennen sie Anomalien präziser und unterstützen die Authentifizierung von Maschinen, Sensoren und Mitarbeitenden. In der Industrie 4.0 ist die Anomalieerkennung von Maschinen- und Prozessdaten kritisch: LLMs verstehen normale Betriebsdatenmuster und identifizieren Abweichungen, die auf Sabotage oder Manipulation hindeuten, zuverlässiger als statische Regeln. Zugleich stärken sie die Compliance und die Resilienz gegenüber Cyberangriffen – durch Echtzeit-Überwachung, dynamische Richtlinien (etwa die automatische Reduzierung der Zugriffsrechte eines Wartungstechnikers nach Abschluss der Arbeit) und Auditierbarkeit. Weil LLMs natürliche Sprache verarbeiten, helfen sie Sicherheitsanalystinnen und -analysten zudem, komplexe Logs und Alarmmeldungen schneller zu deuten und einzuordnen.

Hinweis zur Performanzarchitektur der LLMs:
Prüfungen und dynamische Richtlinienentscheidungen laufen schnell und parallel, weil eine hierarchische, verteilte Zero Trust Architektur sie ermöglicht. Diese stützt sich nicht auf die langsame General Purpose AI (GPAI), sondern auf eine Edge-to-Cloud-Zero-Trust-Kaskade (siehe hierzu die Erläuterungen im [Kapitel 6]).

Im Rahmen **agentischer KI lassen sich LLMs als spezialisierte Co-Pilot-Analysten einsetzen**, die Muster erkennen und Workflows unterstützen. Sie wirken als intelligente Filter und Triage-Systeme und reduzieren die Flut an Sicherheitsereignissen auf die wirklich relevanten Vorfälle. Zugleich können neue Risiken entstehen, etwa Datenmanipulation oder Schwachstellen in der KI-Lieferkette (z. B. kompromittierte Trainingsdaten oder Modelle). Deshalb muss man die **Zero-Trust-Prinzipien konsequent auch auf die KI-Systeme** selbst anwenden: Man muss die Input-Daten, das Modell-Hosting und die **Output-Generierung der LLMs kontinuierlich validieren und autorisieren**. Um komplexe industrielle Sicherheitsprozesse zu steuern und zu überwachen, bleibt das Wissen **menschlicher Fachkräfte** unverzichtbar – über industrielle Prozesse, die spezifischen OT-Protokolle (wie Modbus oder Profinet) und die kritische Infrastruktur –, denn sie treffen die finalen Entscheidungen über Abschaltungen oder Systemrestarts.

In einem Industrie 4.0-Unternehmen zeigt sich das Problem der Dynamik und Reproduzierbarkeit von Entscheidungen darin, dass man Produktionslinien ständig neu konfiguriert, Prozesse fließend ablaufen und laufend neue Technologien hinzukommen. Automatisierung und Vernetzung ändern Systemzustände und Datenströme fortlaufend; statische KI-Modelle, die etwa Anomalien **in Fertigungsprozessen erkennen sollen, stoßen deshalb rasch** an ihre Grenzen und liefern ungenaue Ergebnisse. Besonders schwierig ist es, KI-Entscheidungen reproduzierbar zu halten, wenn sie Maschinen steuern, Energieflüsse optimieren oder die Qualität in Echtzeit kontrollieren. Hier setzt die ZTA an: Weil sie kein implizites Vertrauen kennt, muss sie jeden Datenstrom und jede KI-Entscheidung, die ein Steuerungssystem beeinflusst, kontinuierlich verifizieren. Das erzwingt, dass man die **Datenquellen und die Entscheidungspfade validiert**.

Komplexe KI-Systeme arbeiten oft undurchsichtig; deshalb lässt sich schwer nachvollziehen, wie eine bestimmte Entscheidung in einem dynamischen industriellen Umfeld zustande kam. Für die Betriebssicherheit, die Einhaltung regulatorischer Anforderungen und die Fehleranalyse in der Produktion ist genau das jedoch unerlässlich. Deshalb muss ein Industrie 4.0-Unternehmen KI-Entscheidungen protokollieren und auditieren, um Transparenz und Vertrauen in die autonomen Prozesse zu sichern und bei Bedarf gezielt menschlich einzugreifen. Die ZTA unterstützt dies mit dem Prinzip der **Sichtbarkeit und Überwachung** (Visibility and Monitoring, siehe Kap. 5): Sie erfasst alle Zugriffe und Datenflüsse zwischen Maschinen, Sensoren und KI-Systemen zentral und prüft sie auf Anomalien. Weil die **ZTA** zusätzlich Least Privilege strikt durchsetzt, kann ein kompromittiertes KI-Modell nur minimalen Schaden anrichten.

4.4 Kontrollstrukturen auf unterschiedlichen Ebenen

Für ein Industrie 4.0-Unternehmen ist es eine zentrale Herausforderung, bei der Datengenerierung und -speicherung zu entscheiden, ob sich die jeweiligen Entscheidungselemente eignen, um sie für einen späteren Einsatz oder ein Audit aufzubewahren. Zu diesen Entscheidungselementen zählen etwa KI-abgeleitete Befehle zur Steuerung von Robotern, Anpassungen von Prozessparametern anhand von Sensordaten oder Empfehlungen für die vorausschauende Wartung von Anlagen. Ob sie sich für ein Audit eignen, hängt maßgeblich von der **Datenprovenienz** (Herkunft der Daten), der **Integrität der Sensordaten** und Kontrollbefehle sowie davon ab, ob man KI-Entscheidungen eindeutig bestimmten Betriebszuständen oder Eingangsdaten zuordnen kann. Die ZTA sichert die Datenintegrität entscheidend, indem sie für die Datenübertragung kryptografische Signaturen und eine **attributbasierte Zugriffskontrolle** (Attribute Based Access Control, ABAC) nutzt. Jeder Sensor und jede Aktuator-Anfrage muss sich authentifizieren (Geräte-Identität); so lässt sich die Herkunft der Daten bis zur Quelle nachweisen – ein Kernaspekt für die Audit-Fähigkeit der KI-Entscheidungen.

Das verlangt robuste Data-Governance-Strukturen, die sicherstellen, dass die Daten für KI-Training und Entscheidungsfindung hochwertig, repräsentativ und frei von unbeabsichtigten Verzerrungen sind – eine kritische Anforderung des **EU-AI-Act** (Kapitel 5). Den gesamten Lebenszyklus einer KI-Entscheidung – von den ersten Daten bis zur finalen Aktion im **OT-Bereich** – lückenlos nachzuvollziehen, ist für die **Rechenschaftspflicht** und die kontinuierliche Verbesserung der industriellen Prozesse unerlässlich.

- **Zero Trust Architektur und Künstliche Intelligenz** in der Industrie 4.0 zusammenzuführen, ist spätestens seit 2025 keine optionale Erweiterung mehr, sondern ein fundamentaler Pfeiler für **operative Resilienz** und **Compliance.**

Weil industrielle Umgebungen inhärent dynamisch sind und autonome KI-Systeme komplex entscheiden, braucht es eine umfassende **Governance: Sie legt** klare Verantwortlichkeiten fest und sichert die Einhaltung rechtlicher Rahmenbedingungen wie des **EU-AI-Act. Das verlangt** eine durchdachte Policy, die dynamische, KI-gestützte ZTA-Richtlinien [Dreistufigkeit + Anhang Beispiele] für adaptive Zugriffssteuerung und **Mikrosegmentierung** vorschreibt.

ISO/IEC 42001 (s. Kapitel 8, entspricht ISO 27001 + KI-Aspekte) ist der internationale Standard für Managementsysteme für Künstliche Intelligenz (AIMS). Er fordert, dass die Organisation **Risiken im gesamten KI-Lebenszyklus** systematisch identifiziert und steuert. Für die **Zero Trust Governance** bedeutet das:

- **Rollen und Verantwortlichkeiten festlegen (Verantwortlichkeit)**: Wer verantwortet die Datenintegrität und die korrekte Funktion des KI-Modells?

- **Anforderungen an die Daten (Datenprovenienz)**: Qualität, Repräsentativität und Schutz der Trainings- und Eingabedaten sicherstellen.

- **Überwachen und Messen (Auditierbarkeit)**: die KI-Leistung und -Entscheidungen im Betrieb kontinuierlich überwachen, um die Einhaltung der Sicherheitsrichtlinien (der ZTA) nachzuweisen (s. Kapitel 5 und 6).

Entscheidend ist, die **Datenprovenienz** und **-integrität** für alle industriellen Daten und KI-generierten Entscheidungen lückenlos zu gewährleisten; erst so lassen sich Prozesse und Systemzustände auditieren und reproduzieren, wie es Standards wie **ISO 42001 verlangen**. Diese Umsetzung fordert detaillierte technische Maßnahmen – von der **präzisen Klassifizierung der KI-Systeme** (siehe Kap. 6 Kasten) bis zu robusten Überwachungs- und Verifizierungsmechanismen. Ausführliche Beispiele dafür, die das Vertrauen in die fortschreitende Automatisierung sichern und festigen, finden sich im Anhang, Kapitel 8 dieses Dokuments.

4.5 Weitere Aspekte

Über die dargestellten Kernpunkte hinaus berühren folgende Aspekte den Industrie 4.0-Kontext und sollten berücksichtigt werden:

Die **ZTA-Mikrosegmentierung in OT-Umgebungen bezieht sich eng auf das Zonen-/Conduit-Modell der IEC 62443**, den etablierten Security-Standard für industrielle Automatisierungssysteme; eine konzeptionelle Brücke zwischen beiden Ansätzen läge für die praktische Umsetzung in der Fertigung nahe. Ebenso gilt: Ein Großteil der real eingesetzten OT-Bestandsanlagen (ältere SPS ohne moderne Authentifizierung) erfüllt die hier beschriebenen ZTA-Anforderungen technisch nur eingeschränkt – das **Brownfield-Problem** ist in der Industrie 4.0-Praxis der häufigste Grund für verzögerte oder partielle ZTA-Rollouts.

Die erwähnten **KI-Lieferkettenrisiken** (kompromittierte Trainings- oder Modelldaten) und das Blackbox-Problem KI-gestützter Anomalieerkennung sind im Produktionskontext besonders relevant, weil fehlerhafte oder manipulierte Modellentscheidungen physische Prozesse unmittelbar beeinflussen können; Erklärbarkeitsmethoden (XAI) wären hier ein sinnvoller Anknüpfungspunkt für die geforderte Auditierbarkeit.

➤ **Hinweis:** Die probabilistische Unsicherheit von LLM-Entscheidungen und passende Stabilisierungsmaßnahmen behandelt Kapitel 6 ausführlich.

NIS2 streifen wir hier nur kurz: Viele Industrie 4.0-Betreiber unterliegen bereits deren Risikomanagement- und Meldepflichten; dadurch sind die in diesem Kapitel beschriebenen ZTA-Maßnahmen teils nicht nur Best Practice, sondern regulatorischer Zwang. NIS2 und EU-CRA vertiefen wir bewusst nicht – das Dokument konzentriert sich auf ZTA und KI (LLM). Eine ausführliche Regulatorik-Betrachtung liefern wir nur für den EU-AI-Act in Kapitel 5, denn NIS2 und CRA vollständig einzubeziehen, würde den Rahmen des Dokuments sprengen.

Ein bislang nicht adressiertes Risiko betrifft die **Verfügbarkeit der GPAI-Modelle** selbst – als Teil der Business-Continuity-Vorsorge in KRITIS-nahen Industrie 4.0-Umgebungen. Dieses Kapitel definiert die **Verfügbarkeit** als oberstes OT-Schutzziel, und die ZTA sichert vor allem gegen technische Bedrohungen (Angriffe, kompromittierte Endpunkte). Unberücksichtigt bleibt jedoch ein nicht-technisches Verfügbarkeitsrisiko: dass der Sitzstaat des Anbieters den Zugriff auf ein GPAI-Modell abrupt und regulatorisch erzwungen entzieht.

Ein aktuelles Beispiel zeigt diese Risikokategorie: Am 12. Juni 2026 untersagte das US-Handelsministerium per Exportkontrollanordnung, dass Nutzende außerhalb der USA auf zwei Spitzenmodelle eines US-amerikanischen GPAI-Anbieters zugreifen – mit einer Umset-zungsfrist von wenigen Stunden und ohne Übergangsregelung für bestehende Verträge. Ein Industrie 4.0-Unternehmen, das ein vergleichbares Modell in ZTA-gestützte Anomalieerkennung, Co-Pilot-Funktionen oder Entscheidungsunterstützung integriert hat, würde dadurch die betroffene KI-Funktion kurzfristig verlieren – unabhängig davon, wie es die eigene Sicherheitsarchitektur auslegt. Das Beispiel zeigt eine neue Risikokategorie, die sich von klassischen Cyber-Vorfällen unterscheidet: ein Verfügbarkeitsrisiko, das aus regulatorischen Entscheidungen des Anbieter-Sitzstaats entsteht. Technische Maßnahmen wie Mikrosegmentierung, ABAC oder kontinuierliche Verifizierung greifen hier nicht, weil das Risiko auf vertraglicher bzw. rechtlich-politischer Ebene ansetzt und nicht auf Systemebene.

Für die **Business-Continuity-Planung** in der Industrie 4.0 bedeutet das: GPAI-Abhängigkeiten sollten denselben Kritikalitäts- und Redundanzanforderungen unterliegen wie andere Single-Source-Lieferketten – etwa durch Multi-Provider-Strategien, Exit-Klauseln, On-Premise- oder EU-gehostete Alternativmodelle für sicherheitskritische Funktionen sowie definierte Fallback-Prozesse (Rückfall auf regelbasierte Systeme oder menschliche Entscheidung), falls ein Modell kurzfristig ausfällt. Die digitale bzw. KI-Souveränität ist damit kein abstraktes politisches Thema, sondern eine konkrete Verfügbarkeits- und Resilienzanforderung im Sinne des in diesem Kapitel etablierten Schutzziels.

V. Informationen zum AI-Act der EU

Der EU-AI-Act [8] wird einen einheitlichen Rechtsrahmen für Entwicklung und Nutzung Künstlicher Intelligenz in der Europäischen Union schaffen und verfolgt dabei einen risikobasierten Ansatz. Für Industrie 4.0-Unternehmen liegt die größte Herausforderung darin, ihre autonomen Steuerungssysteme korrekt zu klassifizieren: Gelten bestimmte OT-Anwendungen als kritische Infrastrukturen oder Sicherheitskomponenten, fallen sie unter die Hochrisiko-Kategorie des Gesetzes – mit strengen Compliance-Pflichten.

5.1 Risikoklassifizierung gemäß EU-AI-Act

- **Mit Blick auf die Anforderungen des AI Act ab 2024** und in den Folgejahren muss ein Industrie 4.0-Unternehmen seine KI-Systeme besonders sorgfältig klassifizieren, denn Anwendungen in diesem Sektor können unter die Hochrisiko-KI-Systeme fallen. Das betrifft etwa KI-Systeme, die die Sicherheit von Produkten oder Komponenten verantworten, die als Sicherheitskomponenten dienen oder die kritische Infrastrukturen steuern bzw. deren Betrieb und Management übernehmen. Solche Systeme unterliegen ab dem 2. August 2026 bzw. August 2027 strengen Vorschriften – von der **Datengovernance** über die **Transparenz** und Robustheit der Modelle bis zur menschlichen Aufsicht. Unternehmen müssen dann ein umfassendes **Qualitätsmanagementsystem** nach ISO-Standards einführen (siehe Anhang Kapitel 8), für ihre Hochrisiko-KI-Systeme eine CE-Kennzeichnung sicherstellen und sie künftig in einer EU-weiten Datenbank registrieren.

  Darüber hinaus gelten allgemeine **Transparenzpflichten** für alle **KI-Systeme, die mit Nutzern interagieren**, bereits ab dem 2. August 2026; die Pflicht, Mitarbeitende im Umgang mit KI-Systemen zu schulen, besteht schon seit dem 2. Februar 2025. Für ein Industrie 4.0-Unternehmen heißt das: Wer KI-Systeme einführt und betreibt, muss sich intensiv mit den regulatorischen Rahmenbedingungen auseinandersetzen, die über die technische Umsetzung hinausgehen und rechtliche wie ethische Aspekte umfassen. Für viele ist diese Anforderung neu.

- **Umsetzungen**: Um die genannten Punkte zu beachten und umzusetzen, sollte ein Industrie 4.0-Unternehmen zunächst alle seine KI-Systeme umfassend erfassen und klassifizieren, um die als Hochrisiko eingestuften Systeme zu identifizieren, die den strengsten Anforderungen unterliegen. Das ist eine Kernaufgabe, denn die korrekte Einordnung ist die Grundlage für alle weiteren **Compliance-Maßnahmen**. Anschließend gilt es, ein maßgeschneidertes **AI Governance Framework** aufzubauen, das die Anforderungen des EU-AI-Acts in die bestehenden operativen Technologien (OT) und IT-Prozesse integriert. Dazu gehört, die Datenstrategien anzupassen, damit die industriellen Trainingsdaten qualitativ hochwertig und repräsentativ sind, Mechanismen einzurichten, um KI-Entscheidungen und -**Modelle zu protokollieren** und zu **versionieren und so Auditierbarkei**t zu gewährleisten, sowie die **Cybersicherheit der KI-Systeme selbst zu stärken – von** der Absicherung der Trainingsdaten bis zum Schutz vor Manipulationen im Betrieb. Siehe hierzu Kapitel 10 / Anhang 3 (KI-Datenblatt zur ‚Selbstauskunft' von LLMs bzgl. ihrer Fähigkeiten).

Ein weiterer wichtiger Aspekt ist die **Human Oversight**, also die menschliche Aufsicht, und die gezielte Qualifizierung der Belegschaft für den kompetenten Umgang mit KI in industriellen Anwendungen; denn trotz Automatisierung bleibt menschliches Expertenwissen unverzichtbar, um komplexe Situationen zu bewerten und ethische Richtlinien einzuhalten. Qualitätssicherungsprozesse nach **ISO-Standards und die CE-Kennzeichnung** für Hochrisiko-KI-Systeme muss man von Beginn an in die Entwicklungs- und Einsatzplanung integrieren, um die gesetzlichen Vorgaben einzuhalten und Akzeptanz und Vertrauen in die eingesetzten KI-Systeme der In dustrie 4.0 zu fördern.

5.2 Warnung vor Überklassifizierung

Der **EU-AI-Act** (Verordnung (EU) 2024/1689) weist KI-Systemen eine **risikobasierte Klassifizierung** zu: unacceptable risk (verboten), high-risk (strenge Anforderungen), limited risk (Transparenzpflichten) und minimal risk (unreguliert). Für Systeme in Annex III (z. B. kritische Infrastruktur, Beschäftigung, Produktqualität) gilt die Vermutung „high-risk"; der Provider muss jedoch ein Self-Assessment durchführen und dokumentieren, ob Ausnahmen greifen (z. B. enge prozedurale Tasks oder die Unterstützung menschlicher Entscheidungen ohne Ersatz).

Die hohe Strafdrohung (bis zu 7 % des globalen Jahresumsatzes für Verstöße) hat jedoch eine „Panik“ vor regulatorischen Risiken ausgelöst. Viele Unternehmen stufen Systeme vorsichtshalber als high-risk ein, um Sanktionen zu vermeiden – eine Form der Over-Compliance. Studien wie die des appliedAI-Instituts (2023/2024) zeigen, dass 40 % der Klassifizierungen unklar sind und in bis zu 58 % der Fälle zu Überklassifizierungen führen. Das bindet unnötig Ressourcen (z. B. für Conformity Assessments und Dokumentation), behindert Innovation und lenkt den Fokus von echten High-Risk-Systemen ab, wie die EU-Kommission in Guidelines kritisiert (EC IP/25/2718, 2025). Die Kommission warnt, dass solche Überklassifizierungen die Wirksamkeit des Acts untergraben und echte Risiken für Gesundheit, Sicherheit und Grundrechte verwässern.

Im Vergleich dazu betont **IEC 62443** (Security Levels SL 1–4) eine **präzise, kosten-nutzen-orientierte Klassifizierung**, die Überklassifizierung vermeidet, weil höhere Levels die Kosten direkt steigern (z. B. für Maßnahmen und Audits). Eine „vorsichtshalber höhere“ Einstufung wäre hier wirtschaftlich kontraproduktiv.

Empfehlungen für die Praxis:

- Führen Sie ein fundiertes Self-Assessment durch und dokumentieren Sie Ausnahmen detailliert (z. B. nach EU-Beispielen für non-high-risk).
- Vermeiden Sie Panik-Klassifizierungen und nutzen Sie Tools wie den EU-AI-Office Classifier oder externe Beraterinnen und Berater.
- Priorisieren Sie echte Risiken, um Ressourcen effizient einzusetzen und den Sinn der Klassifizierung – den Schutz vor echten Bedrohungen – zu wahren.

➤ **Hinweis**: Detaillierte Beispiele und erweiterte Tabellen zur Klassifizierung finden Sie im GitHub-Repository (siehe Kapitel 8).

VI. Durchführung und Ergebnisse eines Proof of Concept für ausgewählte Use Cases

Dieses Kapitel klärt, ob es geeignete Use Cases gibt, die aktuelle LLMs im ZTA- und Industrie 4.0-Umfeld unterstützen können und die einen Mehrwert stiften.

Die im Folgenden untersuchten Use Cases beziehen sich auf Large Language Models; die Eignung kleinerer Modelle (u. a. Nano, Mini, Flash, Small …) haben wir nicht untersucht.

Bei Zero Trust Architekturen gibt es jedoch generell Use Cases für alle Arten von KI-Modellen.

Die folgende Tabelle listet solche Einordnungen von Use Cases samt möglicher KI-Modelle unter dem Namen „Zero-Trust-Kaskade“ auf. Siehe hierzu auch Kapitel 10.

Wie Tabelle 2 zeigt, unterscheiden sich die Stufen der Kaskade in ihrer Performance und damit auch in ihrer Leistungsfähigkeit. Um LLMs leistungsfähiger zu machen – also schnellere Antworten bei geringerem Ressourcenverbrauch zu erzielen –, setzt man seit einiger Zeit „Mixture of Experts“-Architekturen ein. Sie wählen bei einer Anfrage vorab den Untersuchungsraum eines LLMs aus. Siehe hierzu Kapitel 10 / Anhang 3 (MoE).

Zero-Trust-Kaskade:
Dynamische, latenzoptimierte KI-Entscheidung in einer hierarchischen Drei-Stufen-Architektur

In Industrie 4.0- und OT-Umgebungen (Industrial Control Systems, ICS) ist harte Echtzeitfähigkeit (< 50 ms) essenziell. Monolithische Large Language Models (LLMs) sind hierfür meist zu latenzintensiv und ressourcenaufwendig. Stattdessen wird eine **verteilte, kaskadierte Zero-Trust-Architektur (ZTA) nach NIST SP 800-207** empfohlen, die schnelle Entscheidungen an den Edge verlagert, kontextuelle Bewertungen im Mid-Layer durchführt und tiefe Analysen sowie Policy-Optimierung zentralisiert.

Die Architektur folgt dem Prinzip „never trust, always verify" durch kontinuierliche Authentifizierung, Least-Privilege-Zugriff und Mikrosegmentierung. **KI-Komponenten (GNNs, Small LLMs, GPAI-LLMs)** agieren als dynamische Policy- und Risiko-Engines an Policy Enforcement Points (**PEPs**), Policy Decision Points (**PDPs**) und Policy Administrator Points (**PAPs**). Dies ermöglicht parallele, latenzarme Sicherheitsprüfungen bei gleichzeitiger Gewährleistung von Resilienz und Observability.

Dieses System gewährleistet durch die **Verlagerung schneller Entscheidungen an den Edge-Bereich** eine latenzarme industrielle Steuerung, während die komplexen, zeitunkritischen Entscheidungen zentralisiert im Cloud- oder SOC-Bereich getroffen werden.

Tabelle 2 Zero-Trust-Kaskade: Dynamische KI-Entscheidung in 3 Stufen

Stufe	Lage / Zeithorizont	Zentrale NIST-Komponente	KI-Typ & Hauptfunktion	Latenz-Ziel	Zweck in ZTA (Zero Trust)	Resilienz- & Observability-Maßnahmen
1	**Edge / Echtzeit**	PEP (Policy Enforcement Point)	GNNs (Graph Neural Networks) – Echtzeit-Klassifikation & Anomalieerkennung im OT-Datenverkehr (quantisiert, on-device)	< 10–20 ms	Sofortige Paket-Zulassung/-Ablehnung; dynamische Mikrosegmentierung direkt an der Anlage / PLC	Fallback: Statische Default-Policy bei Modell-Ausfall; lokale Metrics (Prometheus Node-Exporter); Edge-Logging mit Fluent Bit
2	**Mid-Layer / Near-Realtime**	PDP (Policy Decision Point) / Risk Engine	Small-LM (MoE-optimiert, quantisiert, z. B. 3–12B) – Kontext-Fusion (Identität, Gerätestatus, OT-Historie, UEBA-Signale)	< 50–150 ms	Dynamische Policy-Anpassung (z. B. MFA on-demand, temporäre Rechte-Eskalation/-Reduktion, Risiko-Scoring)	Redundante Mid-Layer-Instanzen (Kubernetes HA); Circuit-Breaker; Observability: Open-Telemetry Tracing + Metrics über alle Fusion-Schritte
3	**Cloud / SOC / Deep Analysis**	PAP (Policy Administrator Point) / SOC	Großes GPAI/LLM – Globale Bedrohungskorrelation, forensische Analyse, Langzeit-Policy-Optimierung	Sekunden – Minuten	Überprüfung & Feinabstimmung von Stufe-1/2-Entscheidungen; Co-Pilot für Analystinnen und Analysten; kontinuierliches Policy-Training	Self-Healing via Orchestrierung (Kubernetes + AI-gestützte Auto-Remediation); zentrale Observability-Plattform (Grafana + Loki + Tempo); Audit-Logs immutable (Blockchain-ähnlich oder WORM-Speicher)

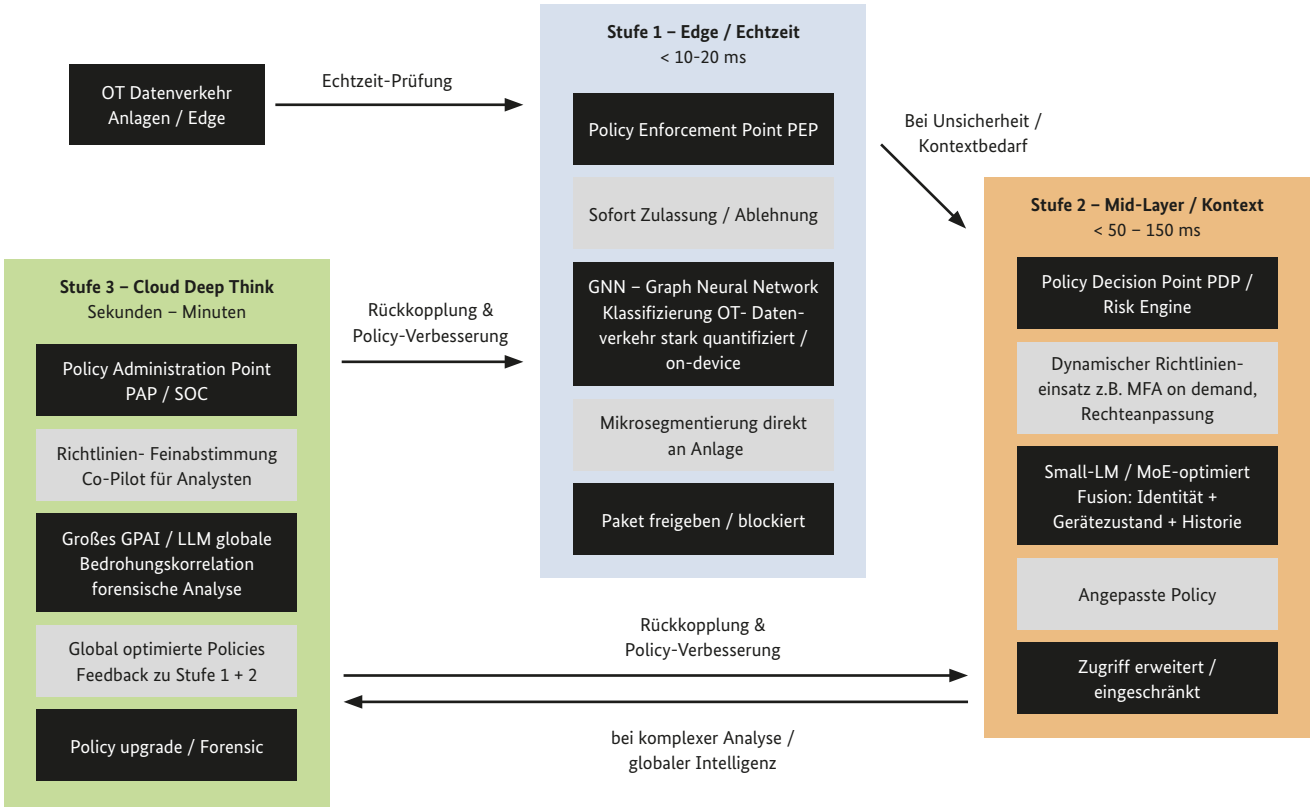


Abbildung 1 – Darstellung der ZTA-Kaskade Stufe 1 bis 3

Die Abbildung zeigt die hierarchische Struktur der Zero-Trust-Kaskade, die eine dynamische und latenzoptimierte KI-Entscheidung in der Industrie 4.0 ermöglicht. Diese Kaskade ist essenziell, um die harten Echtzeitanforderungen der Betriebstechnologie (OT) mit der tiefen Analysefähigkeit moderner KI-Modelle zu verbinden.

Die Stufen ordnen wir innerhalb unserer Publikationsreihe wie folgt ein:

Stufe 1 (Edge / Echtzeit):

Diese Ebene klassifiziert den Datenverkehr unmittelbar und erkennt Anomalien direkt an der Anlage (PEP) mittels GNNs. Diese Grundlagen bildeten den Kern unserer ersten Veröffentlichung 2024 [9]. Im Vordergrund stehen Latenzzeiten unter 10–20 ms, um die Sicherheit der industriellen Steuerung (SPS) zu gewährleisten.

Stufe 2 (Mid-Layer / Near-Realtime):

Sie dient als Bindeglied in der Risiko-Engine (PDP) und führt eine Kontext-Fusion durch, etwa um Zugriffsrechte dynamisch anzupassen. In diesem Dokument führen wir Stufe 2 nur zur Vollständigkeit des Kreislaufs zwischen Stufe 1 und 3 auf. Sie bildet die notwendige logische Brücke, um Kontextdaten für die übergeordneten Systeme aufzubereiten; wir untersuchen sie hier jedoch nicht im Detail. Siehe hierzu Kapitel 10 / Anhang 3 (SML).

Stufe 3 (Cloud / SOC / Deep Analysis):

Dies ist die Ebene der Policy Administration (PAP); hier korreliert man globale Bedrohungsdaten und führt forensische Analysen durch. In unserer aktuellen Veröffentlichung – dem vorliegenden Dokument – konzentrieren wir uns genau auf diese Stufe 3. Hier kommen die großen GPAI- und LLM-Modelle zum Einsatz: Sie fungieren als Co-Pilot für Analysten und steuern die langfristige Optimierung der Sicherheitsrichtlinien.

In der vorliegenden Veröffentlichung konzentrieren wir uns genau auf Stufe 3 des folgenden Abschnitts 6.1 – einen Proof of Concept (PoC).

6.1 Erfolgskriterien des PoCs

Wir diskutieren die Realisierung des PoC auf Stufe 3 der Zero-Trust-Kaskade und berücksichtigen dabei die Nebenbedingungen aus den vorigen Kapiteln. Die Gesamtliste der Erfolgskriterien des PoC haben wir festgelegt auf:

Wirtschaftliche Erfolgskriterien

- Vermeidung eines „Lock-In-Effektes“ bei der Auswahl von LLMs, im Zusammenhang mit sich verändernden geopolitischen Randbedingungen

Compliance-relevante Erfolgskriterien

- Beitrag zur Einhaltung des AI-Acts der EU, insbesondere die Information über die Quellenherkunft bei autarken Entscheidungen, als auch die Nachvollziehbarkeit der Entscheidungen von LLMs.

Technische (typische) Erfolgskriterien

- Resilienz-Erhöhung trotz unterschiedlicher Antworten pro unabhängiger, gleicher Anfrage
- Reproduzierbarkeit von maschinenlesbaren Ausgabeergebnissen (JSON/XML Formate)
- Minimierung von Halluzinationen (Kapitel 10 / Anhang 3, Attention Erosion).
- Unterdrückung von „Silent Truncation“ Effekten (s. Glossar und Kapitel 10 / Anhang 3, Attention Erosion).
- Ausschließen von KI-Fehlern wegen des fehlenden „Langzeitgedächtnisses“ (Kapitel 10 / Anhang 3, Attention Erosion).
- Umgehung von ungenauen und varianten Antworten, die durch die LLM-interne Verwendung von Fachleuten (s. o.: Mixture of Experts-Konzept) kapazitätsbedingt zu Optimierungen führt. Daraus können statistisch variante „Token Drop"-Effekte entstehen (Kapitel 10 / Anhang 3, Attention Erosion). Zusätzliche Varianzen treten generell in Verbindung mit statistisch varianten Startpunkten („Seed-Varianz“, Kapitel 10 / Anhang 3, Attention Erosion) des Suchbeginns des LLMs auf. Das Verhalten der LLMs bezüglich der Behandlung von Anfragen und deren Genauigkeit und Varianz hängt sehr stark von der Architektur des jeweiligen LLMs ab (z. B. ob ein Mixture of Experts-Konzept verwendet wird oder nicht). Die Beeinflussung des LLMs von außen (z. B. die Steuerung, ob ein vorhandener Experte verwendet werden soll) ist dabei nur sehr begrenzt möglich. Im zugehörigen GitHub befindet sich für verschiedene LLMs eine Selbstauskunft (Kapitel 10 / Anhang 3, KI-Datenblatt) zu deren jeweiligen Eigenschaften.
- Geeignete Performance bei Echtzeit-Entscheidungen im ZTA-Bereich (z. B. aktive Prüfregelausführung durch ein LLM in einer Zugriffskontroll-Maschine)

6.2 Untersuchte Use Cases

Folgende Experimente wurden durchgeführt:

6.2.1 Experiment 1 – Echtzeit-Zugriffsanfragen auf eine (Industrie 4.0-) Entität, deren Regeln durch ein LLM geprüft und autark entschieden werden

Dieses Echtzeit-Experiment entspricht Stufe 1 oder 2 der „Zero-Trust-Kaskade“ (s. Tabelle 2).

Vorab-Ergebnis: Wir beendeten das Experiment nach kurzer Evaluierung, weil Large Language Modelle für eine Entscheidung mehrere Sekunden, je nach Modell sogar mehrere Minuten benötigen; damit lässt sich eine Real-Time-Skalierung mit LLMs derzeit nicht erreichen. Kleinere Modelle („Tiny LLM“), die speziell auf Echtzeitanfragen zugeschnitten sind und z. B. über geeignete Fachleute des „Mixture of Experts (MoE)“-Konzepts (Kapitel 10 / Anhang 3 (MoE)) laufen, haben wir bislang nicht untersucht.

Deshalb untersuchen wir das unten genannte Experiment 2 genauer.

6.2.2 Experiment 2 – LLM-unterstützte, asynchrone Regelgenerierung für Zugriffskontrolle, die in einer Zugriffs-Kontroll-Regelausführungsmaschine (Attribute Based Access Control/ABAC) innerhalb einer ZTA maschinell ausgeführt werden können

Dieses Experiment entspricht Stufe 3 der Zero-Trust-Kaskade (s. Tabelle 2). Wir diskutieren es in zwei Varianten:

- **Variante 2.1 als Vorbereitung**: Generierung von ABAC-Regeln (XACML, Python, C++) für den sicheren Zugriff auf ein **SAP-System**

- **Variante 2.2 als Transfer aus Variante 2.1**: Generierung von ABAC-Regeln für den ZTA-Zugriff auf einen Industrie 4.0-Data-/Collaboration Space.

6.3 Ergebnisse des PoCs: Vorgehensweise für Experiment 2

Nachdem wir verschiedenste erste Versuche unternommen hatten, schien es aussichtslos, die genannten Erfolgskriterien (s. Kapitel 6.1) zu erreichen. Erst als wir die gesamte Vorgehensweise strategisch umstellten, stellten sich die Erfolge ein.

Während des PoC erarbeiteten wir folgende Vorgehensweise im Sinne eines Workflows, die die oben genannten Erfolgskriterien erfüllt – ausgenommen das Echtzeit-Kriterium.

Wie Abbildung 1 zeigt, zerlegten wir den Prozess, ausführbares Coding zu erzeugen, in mehrere Teilschritte, die jeweils unabhängig voneinander mit separaten LLM-Prompts laufen.

Entscheidend ist dabei eine iterative Vorgehensweise: Sie erlaubt, die einzelnen Schritte getrennt zu untersuchen und zu optimieren und das Ergebnis jedes Schritts auf seine Gültigkeit zu prüfen.

Die jeweiligen Schritte werden in Abbildung 2 dargestellt.

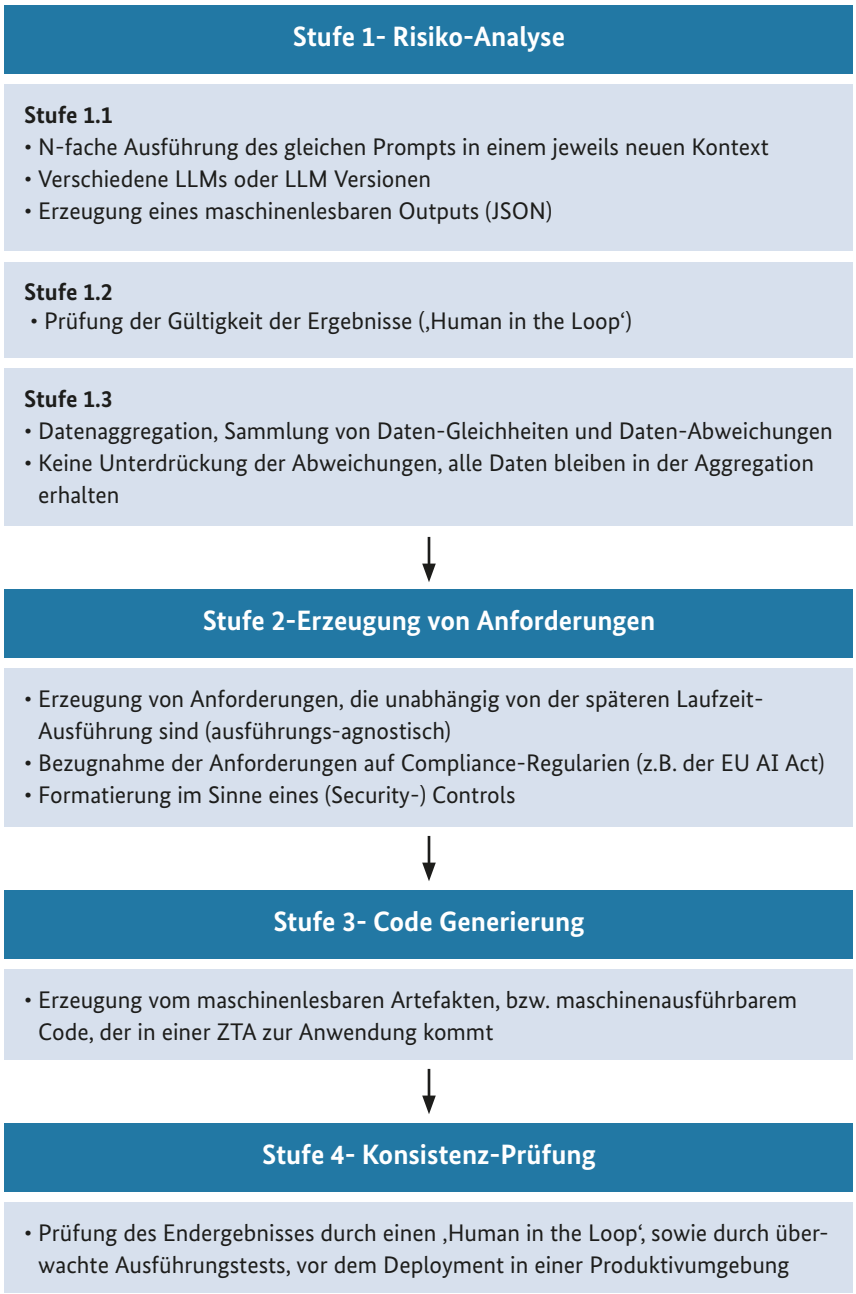


Abbildung 2: Mehrstufige Vorgehensweise

In Stufe 1.1 erheben, werten und bereiten wir Daten auf, um ein zuvor gewähltes Security-Risiko zu beurteilen. Das gewählte Risiko und die daraus resultierende Frage hängen von der Security-Domäne ab (z. B. Industrie 4.0-Security und ZTA) und können oberflächlich oder sehr detailliert sein (s. Variante 2.1 und Variante 2.2 in Kapitel 6.4). Um eine möglichst umfangreiche und möglichst anbieter- und versionsunabhängige Antwort zu erhalten, stellen wir die Frage mehrfach und ohne vorausgehende Historie – an dasselbe LLM sowie an verschiedene LLMs und Versionen. Im Rahmen des PoC haben wir die Prompts so optimiert, dass sie jeweils dasselbe Ausgabeformat erzeugen.

In Stufe 1.2 prüfen wir die Daten auf ihre Gültigkeit. Das muss ein Profi der jeweiligen (Security-)Domäne tun, um Fehler der KI zu erkennen.

In Stufe 1.3 aggregieren wir die n-fach angefragten Antwort-Datensätze. Zum Abschluss der Daten-Evaluation führen wir KI-unterstützt eine Risikobewertung durch, um zu entscheiden, ob spezielle Security-Maßnahmen der Zugriffskontrolle innerhalb einer ZTA gerechtfertigt bzw. sinnvoll sind.

In Stufe 2 erzeugen wir aus den aggregierten Daten der Stufe 1.3 Anforderungen, die die Lösungsvorschläge direkt mit den durch die KI zu untersuchenden Security-Standards verknüpfen. Diese Anforderungen sind unabhängig von der ZTA-Architektur, also laufzeitagnostisch; wir erzeugen noch keinen ausführbaren Code (z. B. C++, Python, XACML), sondern menschlich lesbare Anforderungen im Format eines (Security-)Controls.

In Stufe 3 erzeugen wir aus den Anforderungen der Stufe 2 schließlich ausführbares Coding, in verschiedenen Varianten (C++, XACML, Python ...), die von der jeweiligen ZTA-Architektur abhängen.

In Stufe 4 beurteilt ein ‚Human in the Loop' das Ergebnis abschließend und unterzieht es einem Test, um es final produktiv einzusetzen. Auch hier braucht es einen Spezialisten der jeweiligen (Security-)Domäne, um die Qualität geeignet zu sichern.

Der beschriebene mehrstufige Ansatz löst mehrere generelle Probleme (wie oben bereits angedeutet (Kapitel 10 / Anhang 3, Attention Erosion)) bzw. hilft, mehrere der genannten Erfolgskriterien zu erfüllen:

- **Beitrag zur Einhaltung des AI-Acts der EU**: Jeder Schritt kann separat durchgeführt werden, alle Zwischenergebnisse der einzelnen Schritte können geloggt werden, inkl. entsprechender Quelleninformationen. Ein ‚Human in the loop' kann zusätzlich jeweils eine Ergebnisbeurteilung durchführen, und das Entstehen des Ergebnisses besser nachvollziehen.

➤ **Wichtige Erkenntnis**: Um die Nachvollziehbarkeit und Plausibilität der vom LLM bereitgestellten Ergebnisse zu ermöglichen, ist ein passendes Wissen über die jeweilige Domäne seitens des ‚Human in the loop' unerlässlich.

- **Unterdrückung von „Silent Truncation"-Effekten** (Kapitel 10 / Anhang 3 (Attention Erosion)), **Optimierung der Modell-Interaktion und Bias-Vermeidung**: Die Aufteilung in distinkte Arbeitsschritte verkürzt die einzelnen Prompts erheblich. Hierdurch wird der Effekt der „Silent Truncation" (stillschweigendes Abschneiden von Informationen) wirksam verhindert oder zumindest stark abgemildert.

- **Ausschließen von KI-Fehlern wegen des fehlenden „Langzeitgedächtnisses"**: Die gleiche Methodik, die Länge des Prompts erheblich zu reduzieren und stattdessen die Unterteilung in mehrere, kleinere Arbeitsschritte vorzunehmen, überwindet das Problem, weil kein „Langzeitgedächtnis" benötigt wird.

Zudem unterbindet die Trennung in voneinander unabhängige Anfragen eine unerwünschte Kontext-Erzeugung, die andernfalls als Bias wirken könnte. Dies stellt sicher, dass das LLM nicht dazu neigt, initial gemachte Angaben oder Bedingungen im weiteren Chatverlauf aufgrund interner Optimierungsprozesse zu vernachlässigen. Methodisch ist der Output eines vorangegangenen Schrittes dabei stets als Input für den jeweils nachfolgenden Schritt definiert. Eine detaillierte Darstellung dieses Ablaufs findet sich in Abbildung 2.

Nebenbedingungen zur Erzeugung von Prompts für das LLM

Um die weiteren genannten Erfolgskriterien zu erfüllen, muss man bei der Erzeugung der Prompts besonders auf bestimmte Formulierungen achten. Im Folgenden diskutieren wir die Nebenbedingungen dafür.

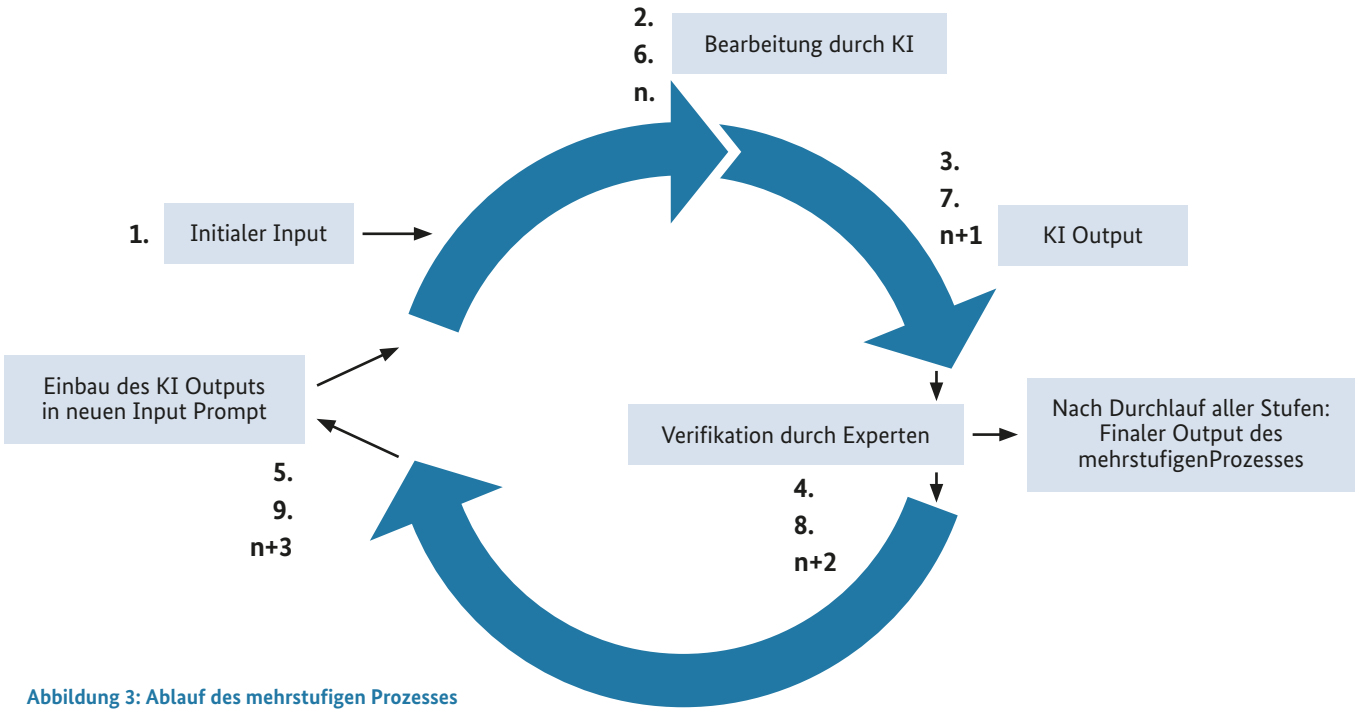


Abbildung 3: Ablauf des mehrstufigen Prozesses

- **Reproduzierbarkeit von maschinenlesbaren Ausgabeergebnissen** (JSON/XML Formate):
  - Die Ausgabestruktur ist speziell zu beschreiben. Welche (JSON/XML/...) Attribute, semantische Erklärung der Attribute, ggf. Längenbegrenzung
  - Hilfreich ist eine gefüllte Beispielstruktur im Prompt
  - Es ist speziell als wichtig zu deklarieren, dass weitere Kommentare oberhalb oder unterhalb der gewünschten (JSON/XML/...) Struktur nicht erlaubt sind (siehe auch Kapitel 9, bzw. die dortigen Links zu den Untersuchungen im zugehörigen GitHub)
  - Hilfreich ist ein zusätzliches Ausweichfeld innerhalb der Struktur, in der das LLM weitere Informationen im Kontext der eigentlichen Frage bereitstellen kann, und daher nicht auf Kommentare außerhalb der Struktur angewiesen ist

  Wer diese Nebenbedingungen berücksichtigt, erzeugt die benötigten Ausgabeformate hoch reproduzierbar.

- **Resilienz-Erhöhung trotz unterschiedlicher Antworten pro unabhängiger, gleicher Anfrage**: Es ist möglich, die gleiche Anfrage kontextfrei mehrfach zu stellen, und den jeweils semantisch ähnlichen, jedoch in Details abweichenden Output nach positiver Bewertung hinterher aggregieren zu lassen, und sich dabei speziell zunutze zu machen, dass die Antworten in Details weitere relevante Informationen enthalten. Bei der Datenaggregation in Schritt 3 ist im Prompt speziell auf geeignete Stichwörter zu achten, um die Daten nicht oberflächlich zusammenzufassen, sondern das Gesamtergebnis **vollständig** zu aggregieren:
  - Explizite Frage nach Gemeinsamkeiten in den Antworten
  - Explizite Frage nach Abweichungen in den Antworten und speziell die Auflistung dieser Abweichungen anfordern
  - Explizite Direktive, die Aggregation „verbatim“ oder „greedy“, bzw. „temperature=0“ durchzuführen (System-Anweisung).
  - Für die Maschinenlesbarkeit eine geeignete Struktur zur Verfügung zu stellen, die entsprechende Attribute für Gemeinsamkeiten und Abweichungen enthält

- **Umgehung von Varianzen aufgrund des „Token-Drop“-Effektes und statistischer Wahl des Such-Startpunktes:** Die aktuell nicht verhinderbare Ausgabe von statistisch varianten Antworten durch den „Token-Drop“-Effekt und statistische Varianzen kann ebenfalls durch mehrfache, kontextfreie, gleiche Anfragen umgangen werden. Da unbekannt ist, bei welchen LLMs der Effekt unter welchen genauen Umständen auftritt, empfiehlt es sich, die Anfrage sowohl mehrfach an ein LLM zu stellen, als auch mehrfach an andere LLMs zu stellen. Um die „Abweichungen“ einzuschränken, empfehlen die Autoren, die Anfrage an mindestens drei LLMs mit je mindestens 3 kontextunabhängigen Anfragen zu stellen, und diese Ergebnisse dann zu aggregieren (s.o. und Kapitel 9 / Anhang 2).

- **Minimierung von Halluzinationen:**
  - Die ‚Erfindung‘ von Antworten ist explizit zu verbieten, z. B. über die Information, wie die Outputstruktur konkret auszufüllen ist, wenn die Antwort nicht

gegeben werden kann (z. B. weil Informationen über das angefragte Objekt, auf das zugegriffen werden soll, nicht bekannt sind, Kapitel 10 / Anhang 3 (Attention Erosion, Halluzinationen, Silent Arbitration))

➤ **Hinweis:** Aufgrund der Architektur von LLMs können Halluzinationen nicht vollständig ausgeschlossen werden, daher wird grundsätzlich die Überprüfung durch den ‚Human in the Loop' benötigt.

- **Vermeidung eines „Lock-in-Effektes"** bei der Auswahl von LLMs, im Zusammenhang mit sich verändernden geopolitischen Randbedingungen: Im Rahmen dieses PoCs wurden bis zu 7 verschiedene LLMs (siehe Anhang 2) mit identischen Prompts getestet, ob die Antworten die oben genannten Kriterien erfüllen. Ergebnis: Je expliziter die Antwort (z. B. Format, Verbot von Halluzinationen, Verwendung bestimmter Stichwörter ...) eingefordert wird, desto reproduzierbarer wird der Output. Der Inhalt des Outputs ist selbstverständlich vom Trainings-/Lernumfang des jeweiligen Modells oder auch der proprietären Algorithmik abhängig.

Erfüllungsgrad der Erfolgskriterien

Mit den oben genannten Maßnahmen lassen sich die genannten Erfolgskriterien erfüllen. Um die Zwischenergebnisse grundsätzlich oder stichprobenartig zu prüfen, braucht es einen **„Human in the Loop" mit domänenspezifischem Wissen**.

Die Ergebnisse der gesamten Untersuchung, also Prompts, Input und Output, stellen wir in Kurzform in Kapitel 9 dar und ausführlich über die dortigen Links zum entsprechenden GitHub bereit.

6.4 Detaillierte Beschreibung von Experiment 2

6.4.1 Variante 1: Generierung von ABAC-Regeln für den Zugriff auf ein SAP-System

Ziel der Variante ist, aus einer Anfrage an ein LLM zur Beurteilung der Security-Relevanz bzw. -Kritikalität eines oder mehrerer SAP-Objekte (z. B. sogenannte Transaktionen, Reports, Funktionsbausteine, Services ...) in einem SAP-S/4H-System final ausführbares Coding für die attributbasierte Zugriffskontrolle zu erzeugen, das in einer ZTA laufen kann. Die ZTA-Maschine kann beliebig implementiert sein und z. B. XACML (eine XML-basierte Open-Source-Policy-Sprache), C++ oder Python ausführen.

Den SAP-Use-Case wählten wir zur Vorbereitung, weil hierfür domänenspezifisches Wissen zur Bewertung der Antworten zur Verfügung stand.

➤ **Ergebnis:** Wir konnten alle Schritte im Rahmen des PoC erfolgreich ausführen und sie dank des vorhandenen domänenspezifischen Wissens verifizieren.

Auszüge aus den Prompts (Input/Output) finden Sie im Anhang (Kapitel 9); die gesamte Versuchsreihe verlinken und veröffentlichen wir in Kapitel 9 über ein zugängliches GitHub-Repository.

6.4.2 Variante 2: Generierung von ABAC-Regeln für den Zugriff auf einen Industrie 4.0-Data-/Collaboration Space

Die Variante beschreibt ein Unternehmen in der EU, das Robotics-Maschinen herstellt, Zulieferer aus verschiedenen Ländern hat und Ingenieuren sowie Software-Entwicklern aus weiteren Ländern Zugriff auf eine gemeinsame Entwicklungsplattform gibt. Die Zugriffsberechtigungen auf die Plattform muss es streng überwachen und regelmäßig bzw. kurzfristig anpassen. Gemäß seinen eigenen Policies muss das Unternehmen laufend Änderungen von Import-/Export-Regularien, Data-Privacy-Regularien, Regularien für den Datenaustausch und für den Datenzugriff beobachten, um Risiken abzuschätzen und zu entscheiden, ob es Zugriffsregeln ändert. Diese geänderten Zugriffsregeln soll es final als ausführbares Coding bereitstellen, wie in Variante 1.

Ein zusätzliches Ergebnis ist, dass sich dieselben Ausgangsdaten, die wir in Stufe 3 (s. Abbildung 1) aggregieren, auch als Grundlage für einen Bericht an das Management und die Rechtsabteilung nutzen lassen. Mit den Daten kann man Risiken abschätzen, etwa bei neuen Ein-/Ausfuhr-Regularien oder Regularien zum Datenaustausch ...

➤ **Ergebnis:** In einem Transfer aus Variante 1 konnten wir alle Schritte im Rahmen des PoC erfolgreich ausführen, inklusive der Erzeugung von Coding für eine ZTA-Laufzeit.

Zusätzlich konnten wir aus den vorliegenden Daten allein durch eine geänderte Frage nicht nur die Anforderungen an die ZTA erzeugen, sondern auch die an das Management und die Rechtsabteilung (s. Kapitel 9).

Auszüge aus den Prompts (Input/Output) finden Sie im Anhang (Kapitel 9); die gesamte Versuchsreihe veröffentlichen wir in einem zugänglichen Repository.

6.5 Zusätzliche, generelle Erkenntnisse aus dem PoC

Aus dem diskutierten PoC ergaben sich folgende weitere Nebenergebnisse, die wir in diesem Kapitel erläutern. Es geht vor allem darum, die im PoC dargelegte Vorgehensweise nachträglich zu erklären und zu begründen.

Die folgenden Argumente beziehen sich ausschließlich auf sachliche bzw. technische Domänen (z. B. Industrie 4.0, Security, ZTA ...) und nicht auf Anfragen, bei denen Diskriminierung o. Ä. droht.

Die Nebenerkenntnisse sind:

- LLM-Antworten sind keine logischen Entscheidungen/Antworten, sondern entsprechen einem Approximationsansatz, der je nach Abstand des LLM-Wissens von der sogenannten ‚Wahrheit' massiv von einer erwarteten logischen Antwort abweichen kann! (Kapitel 10 / Anhang 3 (Semantik))

- Die Stichproben der bereitgestellten Daten unterliegen trotz aller Anforderungen im Prompt, mit ‚greedy' bzw. ‚temperature=0' zu arbeiten, einer semantischen Varianz, die durch entsprechende Ähnlichkeits-Approximation in Vektorräumen und die genannte Seed-Varianz (Kapitel 10 / Anhang 3 (Attention Erosion)) hervorgerufen wird.
  ➤ **Hinweis**: Nicht alle LLMs beachten diese Instruktionen vollständig.

- Ein quellenbasiertes Biasing (z. B. bedingt durch unterschiedliche Datenprovenienz) bei Antworten verschiedener LLMs ist grundsätzlich vorhanden, und trägt zur Varianz von Antworten bei, aus denen die Datenaggregation erzeugt werden kann.

- Als Mitigationsstrategie bietet sich eine Erhöhung der empirischen Streuung durch wiederholte Anfragen an, mit dem Ziel, eine umfassendere Ergebnismenge zu erzielen. Eine Verzerrung der Aussagen durch Überbewertung einzelner Aussagen wird entsprechend weniger wahrscheinlich.

- Auch wenn die einzelnen Antworten stark von den mehrheitlich generierten Ergebnissen abweichen, ergibt sich aus der Erhöhung der Zahl der Anfragen ein vollständigeres Bild

- Massive Abweichungen können durch Domänen-Fachleute zusätzlich aus der Ergebnismenge herausgenommen werden

- Wie im Allgemeinen bekannt: „False-Negative"-Verzerrungen (d. h. existente wichtige Aspekte, die in keinem Datum genannt wurden) können durch die im PoC genannten Ansätze nicht erkannt werden. Generell können „False-Negative"-Verzerrungen nur indirekt durch Indikatoren außerhalb dieser Betrachtung erkannt werden.

Weil eine der Hauptstärken der KI darin liegt, die semantischen Ähnlichkeiten der einzelnen Ergebnisse zu konsolidieren, kann man die genannte semantische Varianz in Kauf nehmen oder sogar erwünschen, um ein umfassenderes Ergebnis zu erzielen.

Anmerkung: Angesichts der Erkenntnisse aus dem PoC ist eine zweite Versuchsreihe sinnvoll, die für alle Schritte sämtliche gewonnenen Erkenntnisse berücksichtigt. Diese künftige Versuchsreihe gehört nicht zu dieser Dokumentation.

VII. Ausblick und Handlungsempfehlungen

Dieses Kapitel fasst die zentralen Erkenntnisse des gesamten Dokuments zusammen. Es gibt einen Ausblick auf die voraussichtlichen Entwicklungen im KI-Bereich und leitet konkrete Handlungsempfehlungen ab. Der Fokus liegt bewusst auf kleinen und mittleren Unternehmen (KMU), die oft weder spezialisierte KI-Abteilungen noch große Rechtsabteilungen haben.

Die Empfehlungen zielen daher darauf, die maximal mögliche Compliance zu erreichen – bei der internen Steuerung, der Anpassung (Fine-Tuning) und der Einsatzumgebung der KI-Systeme – und die Anforderungen an Dokumentation, Qualitätssicherung und menschliche Aufsicht bestmöglich zu erfüllen. Eine Garantie, alle Anforderungen an die Datenherkunftstransparenz von Basismodellen Dritter vollständig zu erfüllen, können wir jedoch nicht geben.

7.1 Zusammenfassung

Aufbauend auf dem 2024 von der Plattform Industrie 4.0 veröffentlichten Diskussionspapier „Zero Trust in der Industrie 4.0 – Perspektiven durch den Einsatz von Machine Learning (ML)“ trägt dieses Update für 2026 der seither rasanten Entwicklung Rechnung.

Dieses Kapitel fasst die wichtigsten Aspekte des Dokuments zusammen. Die Entwicklungen lassen sich in vier Kernbereichen bündeln.

1. **Hohe Geschwindigkeit bei der Weiterentwicklung von LLMs**

   - **Veränderung des KI-Marktes durch neue Marktteilnehmer:** Mit leistungsstarken, frei verfügbaren Open-Source-Modellen, vor allem aus China (z. B. DeepSeek), hat sich die Marktdynamik verändert. Es gibt nun Alternativen zu den geschlossenen Systemen der US-Konzerne; je nach Bedarf laufen sie auf eigener Hardware und ermöglichen so volle Datenhoheit.
   - **Spezialisierte KI-Sicherheitsmodelle:** Technologiekonzerne wie Google (mit Sec-Gemini v1) und Microsoft entwickeln auf Security spezialisierte, kommerzielle KI-Systeme. Diese Modelle versprechen, Sicherheitsteams (SecOps) bei der Vorfallanalyse und Bedrohungsabwehr deutlich effizienter zu machen.
   - **KI-durchgeführte Schwachstellenfindung mit doppelter Verwendung:** Anfang Mai 2026 ergänzte der Hersteller Anthropic seine KI-Software Claude um ein drittes System namens „Mythos“. Es spürt Schwachstellen in Systemsoftware auf und liefert zusätzlich die Angriffsverfahren mit. Mythos könnte die Systeminfrastruktur härten, ließe sich aber ebenso als Angriffswerkzeug missbrauchen. Siehe hierzu Kapitel 1 und eine Mitteilung des BSI [10].

2. **KI und Zero Trust im Industrie 4.0-Kontext:** Die Zero Trust Architektur (ZTA) stellt ein validiertes Sicherheitskonzept nach dem Prinzip "niemals vertrauen, immer verifizieren" dar. Eine Anreicherung von ZTA-Security-Prüfungen mit Hilfe von LLMs ist heute begrenzt möglich. Insbesondere reicht die Performance von LLMs nicht aus, um Echtzeitprüfungen innerhalb einer Zugriffskontrolle durchführen zu können. Allerdings sind LLMs gut geeignet, um komplexe Problemstellungen im Security-/ZTA-Umfeld zu bearbeiten, wie die Erzeugung von lauffähigem Coding zur Zugriffssteuerung, das aus grundlegenden Security-Anforderungen generiert werden kann (s. Kapitel 5). Es wird als Voraussetzung für die Prüfung der Gültigkeit der Informationen grundsätzlich domänenspezifisches Fachwissen benötigt. Siehe auch (Kapitel 9 / Anhang 2, Abstract)

3. **Geopolitische Rahmenbedingungen:** Neuere geopolitische Ereignisse führten u. a. zum Umdenken der Politik in der EU, mit dem Ziel, globale Abhängigkeiten zu verringern. Dies gilt auch für die digitale Souveränität, insbesondere für die Abhängigkeit der EU von Hyperscalern und KI-Anbietern außerhalb der EU. Verschiedene Förderprogramme wurden daher gestartet. Um die Abhängigkeiten zu einem KI-Anbieter zu verringern, besteht die Möglichkeit, die Anfragen an LLMs möglichst LLM-agnostisch zu gestalten. Beispiele hierfür befinden sich in Kapitel 5 und in Kapitel 9 (Anhang 2).

4. **Regulatorische Rahmenbedingungen:** Der EU AI Act setzt als Gesetz verbindliche Regeln für den Einsatz von KI, insbesondere für Hochrisiko-Anwendungen, wie z. B. im Sicherheitsbereich. Für Industrie 4.0-Unternehmen ist die korrekte Einstufung ihrer KI-Systeme von entscheidender Bedeutung,

da viele autonome Steuerungs- und Sicherheitssysteme voraussichtlich als **Hochrisiko-KI** gelten werden. Dies betrifft insbesondere KI in kritischen Infrastrukturen oder als Sicherheitskomponente.

Die EU hat ausdrücklich eine „Warnung vor Überklassifizierung" ausgesprochen, um echte Risiken für Gesundheit, Sicherheit und Grundrechte, wie sie der AI Act klassifiziert, nicht zu verdünnen (s. Kapitel 5.1).

Der durchgeführte Proof of Concept (PoC) ordnet die oben genannten Punkte praxisnah ein: LLMs für kritische Echtzeit-Entscheidungen einzusetzen, ist wegen der Latenzzeiten heute noch nicht praktikabel. Geeignet sind LLMs jedoch für **asynchrone, unterstützende Prozesse** – etwa um Sicherheitsregeln zu erzeugen, zu analysieren und zu dokumentieren.

Wesentliches Ergebnis des PoC: Erst eine mehrstufige, iterative Vorgehensweise – mit separaten Prompts, kontextfreien Mehrfachanfragen an verschiedene LLMs und anschließender Aggregation – erhöht die Wahrscheinlichkeit stark, dass man die definierten Erfolgskriterien erfüllt. Ein „Human in the Loop" mit domänenspezifischem Wissen ist dabei unverzichtbar.

7.2 Ausblick auf kommende Entwicklungen

Die rasante Entwicklung setzt sich fort, und LLMs werden in der Cybersicherheit wichtiger. Für die nähere Zukunft zeichnen sich folgende Trends ab:

1. **Hybride LLM-Architekturen:** Etablierung von LLM-Architekturen, die deren Erklärungsfähigkeit (Reasoning) von deren Wissenshaltung entkoppeln und separate, schnell durchsuchbare Persistenzen (HybridRAG-, Vektor-, Graph-, Keyword- basiert) verwenden.

2. **Weitere Fortsetzung der Entwicklung von Profi-KI-Systemen:** Durch die bereits erfolgte Bereitstellung von „Small-LMs" (Kapitel 10 / Anhang 3 (SLM)) und durch „Mixture of Experts" (MoE)-Architekturen (Kapitel 10 / Anhang 3 (MoE)) innerhalb von LLM-Architekturen entstehen Profi-KI-Systeme, die segmentierter und ressourcenschonender (GPU/TPU) spezifische Aufgaben, wie z. B. die Validierung von Firewall-Regeln oder die Analyse von Malware-Berichten, übernehmen können. Die heutigen Profi-Architekturen führen jedoch durch eine massive Optimierung der verfügbaren Ressourcen zu weiterhin ungenauen und varianten Antworten (s. „Token Drop", „Lost in the Middle"-Effekte (Kapitel 10 / Anhang 3 (Attention Erosion)).

3. **Weitere Erhöhung des Automatisierungsgrads:** Heute etabliert sich das LLM als „Co-Pilot", der menschliche Fachleute durch die Automatisierung wissensintensiver Aufgaben unterstützt. Durch die vermehrte Nutzung von „Agentic AI" können LLMs eigenständig externe Bedrohungsdaten oder neue Regularien überwachen und ggf. proaktiv Vorschläge zur Anpassung von Sicherheitsrichtlinien unterbreiten, bevor eine Schwachstelle ausgenutzt werden kann. Die finale Entscheidung und Freigabe bleibt jedoch beim Menschen.

4. **Entwicklung Richtung Allgemeiner Künstlicher Intelligenz/AGI:** Demis Hassabis (Google DeepMind) rechnet mit 5–10 Jahren bis zum Beginn von AGI, Dario Amodei (Anthropic) mit 12–18 Monaten.

5. **Bereitstellung lokaler KI-Angebote durch das europäische Ökosystem:** KMU können in der EU Beratungsdienstleistungen in Anspruch nehmen. Die European Digital Innovation Hubs (EDIHs) werden als lokale, geförderte Anlaufstellen fungieren, die KMU den Zugang zu Expertise und „Test before Invest"-Möglichkeiten bieten und so die größte Hürde – den Mangel an Fachwissen – abbauen. Weitere Beratungsangebote können künftig auch direkt vom Hersteller (z. B. Anthropic) kommen.

7.3 Handlungsempfehlungen für KMU

Kleinen und mittleren Unternehmen empfehlen wir, die sich bietenden Möglichkeiten strategisch und schrittweise zu nutzen. Die folgenden konsolidierten Handlungsempfehlungen sollen die Umsetzung unterstützen.

1. **Einführung/Erweiterung einer Zero Trust Architektur:** Bevor Sie in Security-KI investieren, schaffen Sie die Basis. Eine solide ZTA ist der größte einzelne Sicherheitsgewinn, den Sie heute erzielen können – auch ohne KI. Die Zero Trust Architektur ist keine rein technische, sondern eine strategische Entscheidung. Beginnen Sie mit grundlegenden Maßnahmen wie z. B. einer Multi-Faktor-Authentifizierung (MFA) für alle Zugänge und/oder der Segmentierung Ihres Netzwerks.

2. **Gezielten Wissensaufbau betreiben und Verantwortlichkeiten klären:** Domänen-wissen ist bei der Ver-

wendung von KI ein kritischer Erfolgsfaktor. Schulen Sie Ihre vorhandenen Fachleute (IT-Admin, Prozessingenieurin oder -ingenieur) darin, LLMs als Werkzeug zu nutzen und deren Ergebnisse kritisch zu validieren.

3. **Vendor-Lock-in-Effekte vermeiden:** Verwenden Sie LLM-agnostische Anfragen/Prompts, um möglichst einfach von einem LLM auf ein anderes wechseln zu können (z. B. bei Versionsupdates, oder bei Veränderungen der geopolitischen Lage). Beispiele zur Erzeugung LLM-agnostischer Anfragen befinden sich in Kapitel 9 bzw. im verlinkten GitHub. Evaluieren Sie den Betrieb von Open-Source-Modellen auf eigener Hardware („On-Premise"). Der Erfolg hängt im Wesentlichen vom eigenen Ressourcenbedarf und der Größe/Architektur der eigenen Hardware ab. Dies gibt Ihnen jedoch die volle Kontrolle über Ihre sensiblen Daten (DSGVO-Konformität) und trägt ebenfalls zur Unabhängigkeit von KI-Angeboten bei. Siehe hierzu auch Kapitel 3.

4. **Zerlegung komplexer Aufgaben:** Der PoC hat gezeigt: Der Schlüssel liegt in der Zerlegung der Aufgaben in kleine, überprüfbare Schritte, um Fehler zu reduzieren. Ferner schafft dies die für den AI-Act geforderte Transparenz. Zur Automatisierung der Ausführung von Einzelschritten gibt es u. a. Open Source Software, wie das ‚Model Context Protocol'/MCP, welches die LLM-Abfragen orchestrieren kann.

5. **Qualitätssicherungsmaßnahmen:** Methoden zur Qualitätssicherung der Ausgabe von LLMs sind dringend zu empfehlen, da durch Änderungen von LLM-Anbietern und Versionen massive Änderungen bei der Datenausgabe entstehen können. Entsprechende Tests sind jeweils bei Änderung des LLMs ausführlich durchzuführen.
6. **Regulierung (EU AI Act):**
   a. Human-in-the-Loop: Setzen Sie KI als Assistenzsystem ein. Die finale Entscheidung in kritischen Prozessen muss bei einem qualifizierten Menschen liegen.
   b. Dokumentation: Beginnen Sie damit, den Einsatzzweck, die Datenquellen und die Verantwortlichkeiten für Ihre KI-Systeme zu dokumentieren. Dies ist eine zentrale Anforderung und schützt Ihr Unternehmen.
   c. Logging von KI-Ergebnissen: Im Zuge der Zerlegung komplexer Aufgaben in kleine Schritte sollte jeder einzelne Schritt bzgl. der Anfrage und Antwort inkl. Meta Daten (Datum, verwendetes LLM und Version, mögliche genannte weitere Quellen der Herkunft der Daten, falls vorhanden) geloggt werden. Solange keine weiteren Voraussetzungen zu erfüllen sind, bestimmen Sie die Vorhaltezeiten für die Log-Daten in einer Unternehmens-Policy, um sie danach zu löschen. Typische Aufbewahrungszeiten reichen von 14 Tagen bis zu über einem Jahr.
   d. Fristen beachten: Beginnen Sie jetzt mit der Schulung Ihrer Mitarbeitenden zur menschlichen Aufsicht (Pflicht seit Feb. 2025) und der Kennzeichnung von KI-Interaktionen (Pflicht seit Aug. 2026).
   e. Datenschutz: Schicken Sie keine sensiblen Unternehmensdaten an öffentliche Cloud-LLMs. Nutzen Sie die KI, um allgemeines Wissen abzufragen, und wenden Sie es intern an.

7.4 Weitere Bemerkungen:

Im Anhang (Kapitel 8 und im dort verlinkten GitHub) finden Sie geeignete Muster für Anforderungskataloge bzw. Controls, um eine KI-unterstützte ZTA im Bereich Industrie 4.0 zu etablieren. Dort stellen wir Muster von Unternehmens-Policies dar: für eine KI-unterstützte Data Governance sowie für die Verwendung von KI in ZTA und Industrie 4.0. Diese Muster passen Sie bei Verwendung unternehmensspezifisch an.

Generell verantworten verschiedenste Unternehmensbereiche – von der Geschäftsführung bis zu den Mitarbeitern in den Fachabteilungen – in ihren jeweiligen Rollen, KI umzusetzen und den AI Act zu erfüllen.

VIII. Anhang 1: Detaillierte Umsetzungs- und Evaluierungsstruktur (Referenz-Repository)

Dieses Kapitel bezieht sich auf den folgenden Link im GitHub:

8.1 Compliance-Dokumentation ZTA+KI

8.1.1 Thema

Auditfähige Compliance-Vorlagen für den Einsatz von Zero Trust Architecture (ZTA) mit KI-Assistenz im Industrie 4.0-Kontext.

8.1.2 Motivation

Industrie 4.0-Umgebungen unterliegen wachsenden regulatorischen Anforderungen durch EU AI Act, ISO/IEC 42001, NIST SP 800-207, IEC 62443 und ISO 27001. Eine zertifizierbare Compliance-Dokumentation muss gleichzeitig C-Level, Legal, CISO, Auditoren und Implementierer adressieren. Dieser Ordner stellt eine vollständige, auditreife Template-Version bereit, die formale RACI-Strukturen, Risiko-Matrizen, KPIs, Lifecycle-Abdeckung, Review-Prozesse und ein granulares SoA-Mapping vereint.

8.1.3 Hauptbotschaften

- Die Dokumente sind als zertifizierbare Vorlage konzipiert (nahe 10/10 Audit-Reife).
- Governance, Policy, Anforderungskatalog und SoA sind konsistent aufeinander abgestimmt.
- Das dreistufige Modul-System (ZTA Teil 1–3) bildet den Weg vom generischen Anforderungskatalog bis zur granularen attributbasierten Verifikation ab.
- Alle Dokumente enthalten Evidenz-Nachweis-Spalten und explizites Normen-Mapping (ISO 42001 Annex A, IEC 62443 SRs).
- Das SoA-Dokument deckt alle 38 ISO/IEC 42001 Annex-A-Controls vollständig ab und dient als Kernstück für Zertifizierung und EU-AI-Act-Conformity-Assessment.

8.1.4 Dokumente im Ordner

- Data-_und_AI-Governance.md
- Policy.md
- Modul_ZTA_Teil_1.md
- Modul_ZTA_Teil_2.md
- Modul_ZTA_Teil_3.md
- SoA.md
- Scope.md ← Architektur und Ausarbeitungsstand der Modul-Teile 1–3
- Abstract.md ← übergeordnetes Abstract des Anhangs 1

8.2 Scope-Hinweis

Die Modul-Teile 1–3 beschreiben einen vollständigen Verifikationsbau über 15 Anforderungen, mehrere hundert Prüfschritte und Verifikationspunkte. Aktuell ist ein exemplarischer Pfad (ZTA-01 → ZTA-01-001 → Verifikationspunkte) vollständig und inhaltlich belastbar ausgearbeitet. Dieser Pfad ist als verbindliche Qualitätsvorlage für alle weiteren Ausarbeitungen zu verstehen. Details zur Gesamtstruktur und zum Ausarbeitungsstand siehe Scope.md.

8.3 Empfohlene Lesereihenfolge

1. Abstract.md (Zusammenfassung und Einordnung)
2. Data-_und_AI-Governance.md (Rahmenwerk, Rollen, Risiko – für C-Level und Legal)
3. Policy.md (operative Grundsätze – für CISO und Fachabteilungen)
4. Modul_ZTA_Teil_1.md (Anforderungskatalog ZTA-01 bis ZTA-15 – für Technikerinnen und Techniker sowie Auditorinnen und Auditoren)
5. Modul_ZTA_Teil_2.md (Prüfschritte exemplarisch für ZTA-01 – siehe Scope.md für Gesamtbild)
6. Modul_ZTA_Teil_3.md (Granulare Verifikationspunkte exemplarisch für ZTA-01-001 – siehe Scope.md)
7. SoA.md (vollständiges Controls-Mapping – für Auditoren und Zertifizierung)

8.4 Verbindung zum Gesamtprojekt ZTA+KI_2026

Dieser Ordner liefert die normative und operative Compliance-Grundlage für alle Architektur-, Sicherheits- und Governanceentscheidungen im Projekt. Er ergänzt Anhang 3 (KI-Modelle), der die technische und theoretische Basis bereitstellt.

IX. Anhang 2: Forschungsergebnisse

Dieser Anhang bezieht sich auf folgenden Link im GitHub:

9.1 Proof of Concept: LLM-gestützte ABAC-Regelgenerierung für Zero Trust Architectures

9.1.1 Thema

Machbarkeitsnachweis für den Einsatz von Large Language Models (LLMs) zur automatisierten Generierung von Attribute-Based Access Control (ABAC)-Regeln in Zero Trust Architectures (ZTA) – exemplarisch für zwei Sicherheitsdomänen.

9.1.2 Motivation

Die Frage, ob heutige General Purpose AI bzw. LLMs zuverlässig zur Verwaltung von ABAC-Regeln in sicherheitskritischen ZTA-Umgebungen eingesetzt werden können, ist nicht trivial. LLMs bringen strukturelle Hindernisse mit – von Halluzinationen über Silent Truncation bis zu Seed-Varianz –, die in Sicherheitskontexten nicht tolerierbar sind. Dieser PoC untersucht systematisch, ob und wie diese Hindernisse durch einen iterativen, LLM-agnostischen Ansatz überwunden werden können, und liefert dazu konkrete, reproduzierbare Ergebnisse für zwei Anwendungsdomänen.

9.1.3 Haupt-Botschaften

- LLM-typische Hindernisse (Halluzinationen, Silent Truncation, Seed-Varianz u. a.) sind durch einen iterativen, schrittweisen Ansatz mit kleinen Prompts beherrschbar.
- Der Ansatz ist LLM-agnostisch: Prompts wurden gegen GPT, Gemini, Grok, Deep-Seek, Mistral, Amazon Nova und Claude getestet.
- Traceability-Anforderungen (EU AI Act) werden durch lückenlose Protokollierung jedes Schritts erfüllt.
- Für beide Use Cases (SAP, Industrie 4.0) konnte ausführbarer ZTA-Code generiert werden (XACML, C++, Python, ABAP).
- Ein Domain Expert als „Human in the Loop" bleibt in jedem Schritt zwingend erforderlich.
- Alle Ergebnisse sind als Näherungslösung zu verstehen (Kapitel 10/Anhang 3 (Semantik, Approximation)).

9.1.4 Dokumente und Ordner

- 1 SAP/ ← Use Case 1: ABAC-Regelgenerierung für SAP S/4HANA
- 2 I40/ ← Use Case 2: ABAC-Regelgenerierung für Industrie 4.0-Kollaborationsplattform
- Scope.md ← Iterativer Ansatz, Schrittstruktur und LLM-Hindernisse im Detail
- Abstract.md ← Übergeordnetes Abstract des Anhangs 2

9.2 Scope-Hinweis

Der PoC folgt einem strukturierten 4-Schritte-Ansatz (Datenerhebung → Validierung → Aggregation → Anforderungen → Code). Jeder Schritt ist durch Prompts und LLM-Antworten dokumentiert. Die vollständige Methodik, die Hindernisse und deren Lösungsstrategien sind in Scope.md beschrieben.

9.3 Empfohlene Lesereihenfolge

1. Abstract.md (Zusammenfassung und Einordnung)
2. Scope.md (Methodik, Hindernisse, Lösungsstrategien)
3. 1 SAP/README.md (Use Case SAP – Beschreibung und Ergebnisse)
4. 2 I40/README.md (Use Case I40 – Beschreibung und Ergebnisse)
5. Einzelne Step-Ordner nach Bedarf (Prompts und LLM-Antworten)

9.4 Verbindung zum Gesamtprojekt ZTA+KI_2026

Dieser Anhang liefert den empirischen Machbarkeitsnachweis für das Gesamtprojekt. Er zeigt, dass LLM-gestützte ZTA-Compliance unter kontrollierten Bedingungen funktioniert, und begründet damit die methodischen Entscheidungen in Anhang 1 (Compliance-Dokumentation) und Anhang 3 (KI-Modelle).

X. Anhang 3: Hinweise zu LLM-Architekturen und Eigenschaften

Dieser Anhang bezieht sich auf folgenden Link im Github.

10.1 KI-Modelle

10.1.1 Thema

Funktionsweise, Grenzen und systemische Rolle von KI-Modellen im Kontext von Zero Trust Architecture (ZTA) und Industrie 4.0.

10.1.2 Motivation

Industrie 4.0-Systeme sind verteilt, dynamisch und sicherheitskritisch. ZTA verlangt explizite Verifikation und kontextbasierte Vertrauensentscheidungen. KI-Modelle übernehmen Mustererkennung und Entscheidungsunterstützung, sind jedoch reine Funktionsapproximationen ohne echtes Verständnis. Dieser Ordner bündelt alle notwendigen Perspektiven, damit die Grenzen klar werden.

10.1.3 Hauptbotschaften

- KI-Modelle sind Funktionsapproximationen (Mustererkennung, Wahrscheinlichkeiten – keine Bedeutung, kein Verständnis).
- Kritisch ist nicht das einzelne Modell, sondern das Zusammenspiel in verteilten Systemen.
- Semantik muss explizit ergänzt werden; Vertrauen kann nicht allein von KI getragen werden.
- ZTA-Anforderungen (Verifikation, Nachvollziehbarkeit, Kontext) übersteigen die natürlichen Fähigkeiten reiner KI.
- Das neue KI-Modelle-Abstract fasst die sechs Unter-Abstracts (Semantik, Attention Erosion, SLM, MoE, KI-Datenblatt, KI-Fingerprint) zusammen und stellt den übergeordneten wissenschaftlichen Rahmen dar.

10.1.4 Dokumente im Ordner

- Attention Erosion
- KI-Datenblatt
- KI-Fingerprint
- MoE
- SLM
- Semantik
- ZTA-Kaskade.md
- Abstract.md
  (übergeordnetes Abstract der KI-Modelle)

10.2 Empfohlene Lesereihenfolge

1. KI-Modelle-Abstract.md (Zusammenfassung aller Perspektiven)
2. Semantik (Grundlagen & Grenzen)
3. Attention Erosion (Modellverhalten im Betrieb)
4. KI-Datenblatt + KI-Fingerprint
   (konkrete Modelle & Dokumentation)
5. SLM + MoE (Architekturvarianten)
6. ZTA-Kaskade.md (Systemdynamik)

10.3 Verbindung zum Gesamtprojekt ZTA+KI_2026

Dieser Ordner liefert die technische und theoretische Basis für alle weiteren Compliance-, Architektur- und Sicherheitsentscheidungen im Projekt.

XI. Anhang 4: Maßnahmen Europas und Deutschlands zu digitaler Souveränität.

Die EU-Kommission hat seit 2021 mit dem Digital Europe Programm (DEP) und 8,1 Mrd. Euro ein großes Infrastruktur-Paket geschnürt, zunächst bis 2027, um die digitale Souveränität zu stärken [11]. Sie nennt dazu folgende Ziele: a) die EU zu einem Zentrum von Weltrang für KI ausbauen; b) die Resilienz gegenüber potenziellen Risiken erhöhen; c) KI-Souveränität erreichen, im Sinne von „KI Made in EU", technologisch wie ökonomisch. Mit diesen Zielen adressiert die EU KI-Anwendungen in sehr vielen Märkten: Automotive, Mobilität, Robotik, Gesundheitsversorgung, IT-Sicherheit, Landwirtschaft und öffentliche Verwaltung, aber auch Industrie 4.0. Für die europäische Industrie sind aus EU-Sicht drei Aspekte wirtschafts- bzw. umsatzrelevant: a) KI in Europa entwickeln, um den EU-Binnenmarkt zu stärken und die Abhängigkeit von ausländischen Lösungen zu verringern; b) KI in Europa vermarkten – eigene Fähigkeiten aufbauen, um den EU-Binnenmarkt abzudecken; c) KI-Lösungen in die Welt exportieren – erweiterte Geschäftsmodelle, etwa wenn der Maschinenbau vom Hersteller zum Dienstleister wird. Das mehrjährige EU-Finanzierungsprogramm legt vier Schlüsselbereiche fest und budgetiert sie: Supercomputer (High Performance Computing; HPC), künstliche Intelligenz, Cybersecurity und digitale Kompetenz. Eine Maßnahme ist der Aufbau eines Netzwerks europäischer digitaler Innovationszentren (englisch: European Digital Innovation Hubs, kurz EDIH [11]). Zum Zeitpunkt dieses Berichts zählt man in Europa 227 Innovation Hubs, davon 16 in Deutschland [12]. Die deutschen EDIH-Zentren stehen unter der Schirmherrschaft des Bundesministeriums für Wirtschaft und Energie (vormals BMWK). Mit dem Programm EDIH will die EU u. a. die Fähigkeiten im Bereich KI in Europa stärken und die Akteure (z. B. KI-Experten) stärker vernetzen. Um Computer und KI weiterzuentwickeln, startete die EU 2025 den Aufruf zu KI-Gigafactories; daraufhin gingen 76 Interessenbekundungen an 60 Standorten in 16 Mitgliedsländern ein [13] [14]. Das KI-Gigafactory-Konzept baut auf der KI-Fabriken-Initiative auf und nutzt Europas hochmodernes EuroHPC-Supercomputing-Netzwerk [15], um das ambitionierte Ziel zu verwirklichen, der weltweit führende KI-Kontinent zu werden. KI-Gigafactories sollen groß angelegte KI-Rechen- und Datenspeicherzentren werden, die speziell dafür entwickelt sind, KI-Modelle und -Anwendungen der nächsten Generation im Hyperskalenbereich zu entwickeln, zu trainieren und bereitzustellen. Durch enorme Rechenleistung, energieeffiziente Rechenzentren und KI-gesteuerte Automatisierung sollen sie neue Maßstäbe für das Training, die Inferenz und den Einsatz von KI-Modellen setzen. Ende 2025 startete die EU eine weitere Initiative namens IPCEI AI (Important Project of Common European Interest for Artificial Intelligence). Mit dem Aufruf, der in Deutschland vom BMWE erfolgte, wollen die EU-KOM und die Bundesregierung ein leistungsfähiges und innovatives europäisches Ökosystem aufbauen bzw. befördern. Im Fokus stehen u. a. industrielle Anwendungen von KI wie Produktion und Robotik. Um zusätzlich die Mikroelektronik zu fördern – etwa Mikroprozessoren mit KI-Funktion –, startete im November 2025 eine parallele Initiative namens IPCEI AST (Advanced Semiconductor Technologies). Das Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) beschloss 2018 (damals noch als BMBF) das Projekt „100 neue Hochschul-Professuren für die KI-Technologie" und startete es 2022; es soll Ausbildung und Forschung in Deutschland stärken. Diese Stellen wurden bis Anfang 2025 besetzt [16]. Das Deutsche Forschungszentrum für künstliche Intelligenz (DFKI) hat bis zum Erscheinen dieses Dokuments ein Netzwerk von fünf deutschen Kompetenzzentren für die KI-Forschung aufgebaut: BIFOLD, MCML, LAMARR, TUE.AI Center und ScaDs.AI.

Diese Auflistung der EU-Maßnahmen in den Bereichen Supercomputing und KI ist nur ein Ausschnitt und erhebt keinen Anspruch auf Vollständigkeit. Der Anhang soll zeigen, dass Europa für die digitale Souveränität in den letzten Jahren viele politische bzw. wirtschaftspolitische Initiativen gestartet und mit erheblichen finanziellen Ressourcen ausgestattet hat, die in den nächsten Jahren Wirkung zeigen werden. Wie hoch die technologische Souveränität im EU-Binnenmarkt ausfällt, lässt sich erst in einigen Jahren beantworten. Dabei ist eine differenzierte Betrachtung nach „Typ" der KI – z. B. konventionelle KI, Anomalie-KI, LLM-KI, Assistant-KI – und nach Anwendungsdomäne wie Industrie 4.0, Medizintechnik und Automobil angezeigt.

Die Autorinnen und Autoren dieses Dokuments bezweifeln, dass Europa mit den beiden Industrienationen China und USA bei KI mithalten wird. Wirtschaftliche Interessen und militärische Macht führen seit mehreren Jahren zu sehr hohen Ressourceneinsätzen in diesen Ländern, die mit technologischen Infrastrukturen (KI-Chips und Rechenzentren) ergänzt werden.

XII. Verweise

[1] Plattform Industrie 4.0, „Whitepaper "Zero Trust and Industrie 4.0" (2024/2025,“ 04 2024. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/202404_Zero_Trust.pdf. [Zugriff am 03 02 2026].

[2] E. Burzstein und M. Tishchenko, „https://blog.google/,“ Gooogle, 01 04 2025. [Online]. Available: https://blog.google/security/google-launches-sec-gemini-v1-new/. [Zugriff am 01 05 2025].

[3] Anthropic, „https://www.anthropic.com,“ Anthropic, 07 04 2026. [Online]. Available: https://www.anthropic.com/glasswing. [Zugriff am 23 07 2026].

[4] ComputerWeekly, „https://www.computerweekly.com/,“ ComputerWeekly, 31 03 2025. [Online]. Available: https://www.computerweekly.com/de/tipp/Top-8-Unternehmen-fuer-KI-Hardware. [Zugriff am 21 04 2025].

[5] Deutschlandfunk, „https://www.deutschlandfunk.de/,“ Deutschlandfunk, 16 06 2026. [Online]. Available: https://www.deutschlandfunk.de/warum-die-sperrung-der-ki-modelle-von-anthropic-so-brisant-ist-100.html. [Zugriff am 21 07 2026].

[6] grandviewresearch.com(, „Global Zero Trust Architecture Marke (Grand View Research, 2025),“ [Online]. Available: https://www.grandviewresearch.com/industry-analysis/zero-trust-architecture-market-report. [Zugriff am 03 02 2026].

[7] mordorintelligence.com, „Zero Trust Security Market (Mordor Intelligence, 2026),“ [Online]. Available: https://www.mordorintelligence.com/industry-reports/zero-trust-security-market. [Zugriff am 03 02 2026].

[8] European Commission, „EU AI-Act - Regulation - EU - 2024/1689 - EN - EUR-Lex,“ 13 06 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. [Zugriff am 03 02 2026].

[9] Plattform Industrie 4.0, „https://www.plattform-i40.de/,“ Plattform Industrie 4.0, 15 04 2024. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/202404_Zero_Trust.html. [Zugriff am 01 02 2025].

[10] C. Plattner, „ttps://www.bsi.bund.de/,“ BSI - Bundesamt für Sicherheit in der Informationstechnik, 08 05 2026. [Online]. Available: https://www.bsi.bund.de/DE/Service-Navi/Presse/Alle-Meldungen-News/Blog/KI-Modelle_neue_Zeitrechnung_260508.html. [Zugriff am 01 06 2026].

[11] Bundesministerium für ¬Forschung, Technologie und Raumfahrt, „Digital Europe Programme (DIGITAL),“ [Online]. Available: https://www.nks-dit.de/weitere-foerdermoeglichkeiten/digitaleurope. [Zugriff am 03 02 2026].

[12] Bundesministeriums für Digitales und Staatsmodernisierung (BMDS), „Summit on European Digital Sovereignty,“ [Online]. Available: https://www.de.digital/DIGITAL/Redaktion/DE/Dossier/european-digital-innovation-hubs.html. [Zugriff am 03 02 2026].

[13] Europäische Union, „https://digital-strategy.ec.europa.eu/,“ Europäische Union, 30 06 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/de/news/overwhelming-response-76-respondents-express-interest-european-ai-gigafactories-initiative. [Zugriff am 01 07 2025].

[14] EuroHPC, „The European High Performance Computing Joint Undertaking (EuroHPC JU) - HammerHAI,“ [Online]. Available: https://eurohpc-ju.europa.eu/ai-factories/germany_en. [Zugriff am 03 02 2026].

[15] EuroHPC, „HammerHAI (Hybrid and Advanced Machine Learning Platform for Manufacturing, Engineering, and Research), Stuttgart and the JUPITER AI Factory (JAIF), Juelich are based in Germany.,“ [Online]. Available: https://eurohpc-ju.europa.eu/ai-factories/germany_en. [Zugriff am 03 02 2026].

[16] Lernende Systeme, „Köpfe für KI: 100 zusätzliche Professuren in Deutschland geschaffen," [Online]. Available: https://www.plattform-lernende-systeme.de/aktuelles-newsreader/koepfe-fuer-ki-100-zusaetzliche-professuren-in-deutschland-geschaffen.html . [Zugriff am 03 02 2026].

[17] Plattform Industrie 4.0, „https://www.plattform-i40.de," Plattform Industrie 4.0, 06 12 2018. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/zugriffssteuerung-industrie40-komponenten.html. [Zugriff am 01 02 2025].

[18] KI Gemini 2.5 Pro, „KI Erklärung zum Begriff "Silent Truncation"," 16.12.2025. [Online].

[19] Statista GmbH, „ Markvolumen für Künstliche Intelligenz weltweit im Jahr 2021 und 2022 mit einer Prognose bis 2030," [Online]. Available: https://de.statista.com/statistik/daten/studie/1405265/umfrage/kuenstliche-intelligenz-marktvolumen/. [Zugriff am 03 02 2026].

[20] Plattform Industrie 4.0, „https://www.plattform-i40.de/," Plattform Industrie 4.0, 15 04 2024. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/202404_Zero_Trust.html. [Zugriff am 01 02 2025].

[21] Global Market Insights Inc, „Zero Trust Architecture Marke (Global Market Insights, 2025)," [Online]. Available: https://www.gminsights.com/industry-analysis/zero-trust-architecture-market. [Zugriff am 03 02 2026].

XIII. Glossar

Begriff (DE)	EN	Abk.	Definition
Agentic AI	Agentic AI	–	Autonome KI-Systeme, die Ziele eigenständig planen, entscheiden und in iterativen Perception-Reasoning-Action-Loops handeln.
Automatische Extraktion strukturierter Graphen	Automatic Structured Graph Extraction	–	Automatisierte Erzeugung von Knowledge Graphs aus unstrukturierten oder multimodalen Quelldaten.
General Purpose AI	General Purpose AI	**GPAI**	Breit einsetzbare, nicht aufgabenspezifisch optimierte KI-Modelle; im Vergleich zu spezialisierten Modellen meist langsamer in der Inferenz.
Generative Pre-trained Transformer	Generative Pre-trained Transformer	**GPT**	Transformer-basierte generative Sprachmodellfamilie, die durch Pre-Training auf großen Textmengen lernt.
Graph Neural Network	Graph Neural Network	**GNN**	Neuronales Netz für Graph-Strukturen (Knoten/Kanten), optimiert für Mustererkennung in vernetzten Daten, z. B. Anomalieerkennung in OT/IT.
Graphbasierte Traversierung	Graph-based Traversal	–	Navigation durch Knoten und Kanten eines Graphen zur Beantwortung relationaler bzw. Multi-Hop-Abfragen.
Hybrid Retrieval / HybridRAG	Hybrid Retrieval / HybridRAG	–	Retrieval-Augmented Generation, die Vektor-, Graph- und Keyword-Retrieval kombiniert, um LLMs präziseren Kontext zu liefern.
Industrial Control Systems	Industrial Control Systems	**ICS**	Sammelbegriff für Systeme zur Steuerung von Industrieanlagen und kritischen Infrastrukturen.
Inhärent probabilistische Modelle	Inherently Probabilistic Models	–	Stochastische Modelle (z. B. LLMs) mit systembedingter Ergebnisunsicherheit und Halluzinationsrisiko.
Klassische Keyword-Suche	Classical Keyword Search / Sparse Retrieval	–	Textbasierte Suche über Wortübereinstimmungen, z. B. nach dem BM25-Verfahren.
Knowledge Graph (temporal-aware)	Temporal-aware Knowledge Graph	–	Wissensgraph mit semantischen Relationen und zusätzlicher zeitlicher Modellierung (Gültigkeitszeiträume, Ereignisfolgen).
Large Language Model	Large Language Model	**LLM**	Transformer-basiertes Modell mit Milliarden Parametern, das auf Text vortrainiert ist und zu Generierung, Verständnis und Reasoning fähig ist; die generative/reasoning-fähige Kernkomponente hybrider Architekturen.
Machine Learning	Machine Learning	**ML**	Überbegriff für Algorithmen, die aus Daten lernen, statt explizit programmiert zu werden – von klassischen Verfahren bis Deep Learning.
Mixture of Experts	Mixture of Experts	**MoE**	LLM-Architektur, bei der pro Anfrage nur ein Teil der „Experten"-Teilmodelle aktiviert wird, um die Inferenz zu beschleunigen.
Multi-Faktor-Authentifizierung	Multi-Factor Authentication	**MFA**	Verfahren, bei dem eine Identität über mindestens zwei unabhängige Faktoren nachgewiesen werden muss.
Multi-Head-Attention-Mechanismus	Multi-Head Attention Mechanism	–	Kernmechanismus der Transformer-Architektur: parallele Aufmerksamkeitsberechnung über mehrere „Köpfe" zur Erfassung unterschiedlicher Kontextbeziehungen.

Begriff (DE)	EN	Abk.	Definition
Multi-Hop-Schlussfolgerungen	Multi-Hop Reasoning	–	Mehrstufige Ableitung einer Antwort über mehrere verkettete Relationen hinweg.
Operational Technology	Operational Technology	**OT**	Hardware und Software zur Überwachung und Steuerung physischer Prozesse.
Policy Administrator Point	Policy Administrator Point	**PAP**	ZTA-Komponente, die Sicherheitsrichtlinien verwaltet und auditiert.
Policy Decision Point	Policy Decision Point	**PDP**	ZTA-Komponente, die anhand von Kontextdaten über Zugriffsanfragen entscheidet.
Policy Enforcement Point	Policy Enforcement Point	**PEP**	ZTA-Komponente, die eine Zugriffsentscheidung unmittelbar technisch durchsetzt.
Precision	Precision	–	Genauigkeitsmaß: Anteil tatsächlich relevanter Elemente an allen gefundenen Elementen (TP / (TP + FP)).
Reasoning	Reasoning	–	Fähigkeit zu logischen, kausalen, mehrstufigen Schlussfolgerungen, bei LLMs häufig chain-of-thought- oder RL-gestützt.
Recall	Recall	–	Vollständigkeitsmaß: Anteil gefundener relevanter Elemente an allen tatsächlich relevanten Elementen (TP / (TP + FN)).
Rein vektorielle Modelle	Purely Vector-based / Embedding-only Models	–	Modelle, die ausschließlich Embedding-basierte Ähnlichkeitssuche nutzen, ohne Graph- oder Regelkomponente.
Security Operations Center	Security Operations Center	**SOC**	Zentrale Stelle zur Überwachung von und Reaktion auf Sicherheitsvorfälle.
Silent Truncation	Silent Truncation	–	Stilles Abschneiden von Eingabetext bei Überschreitung des Kontextfensters eines LLM, ohne Fehlermeldung.
Small LLM	Small Large Language Model	–	Kleineres, aufgabenspezialisiertes LLM (oft MoE-basiert), das schneller reagiert als ein GPAI-Modell.
Speicherprogrammierbare Steuerung	Programmable Logic Controller	**SPS**	Rechnersystem zur Steuerung und Regelung von Maschinen und Prozessen.
Tiny-LLM	Tiny / Micro / Nano / Lite LLM	–	Stark quantisiertes und destilliertes LLM für extrem schnelle Edge-Inferenz.
Token-basierte Modelle	Token-based Models	–	Modelle, die Text nur als Token-Sequenz verarbeiten, ohne explizite Graph- oder Strukturkomponente.
Vektor-Retrieval	Vector Retrieval / Dense Retrieval	–	Semantische Suche über Embeddings und Ähnlichkeitsmaße, z. B. Cosine-Similarity.
Wissensspeicherung	Knowledge Storage / Persistent Knowledge Base	–	Externe, persistente Ablage von Wissen, z. B. in Knowledge Graphs oder Vector Stores.
Zero Trust Architektur	Zero Trust Architecture	**ZTA**	Sicherheitsarchitektur, die keinem Nutzenden, Gerät oder keiner Anwendung standardmäßig vertraut; setzt die Zero-Trust-Prinzipien technisch um.
Zero-Trust-Prinzipien	Zero Trust Principles	–	Grundsatz „niemals vertrauen, immer überprüfen": kontinuierliche Verifikation von Identität, Gerät, Kontext und Verhalten statt implizitem Perimeter-Vertrauen; Basis der ZTA.

AUTORINNEN UND AUTOREN
Thomas Walloschke (secon trust consult) | Dr. Michael Schmitt (SAP SE) | Dr. Bernd Kosch (Industrie KI GmbH)
Michael Jochem (Robert Bosch GmbH) | Dr. Detlef Houdeau (Infineon Technologies AG)

DIES IST EINE PUBLIKATION DER UAG „KI FÜR SICHERE INDUSTRIELLE PROZESSE“ DER AG „SICHERE KOLLABORATIVER ÖKOSYSTEME“ DER PLATTFORM INDUSTRIE 4.0.