Bei diesem Beitrag handelt es sich um einen wissenschaftlich begutachteten und freigegebenen Fachaufsatz („reviewed paper“).

doi.org/10.37544/1436-4980-2025-09-21

Methodik für eine risikobasierte autonome Cyberabwehr mit LLM und RAG

Autonome Cyberabwehr

M. Geyer, J. Schwab

ZUSAMMENFASSUNG Dieser Beitrag skizziert eine Methodik für autonome Cyberabwehr in Industrie-4.0-Umgebungen, die organisatorische Risiken mit generativer KI verbindet. Durch die Kombination von LLMs, RAG und standardisierten SOPs entsteht ein strukturierter Entscheidungsprozess, der Echtzeitreaktionen und Auditierbarkeit ermöglicht. Die konzeptionellen Grundlagen, technischen Herausforderungen und Potenziale werden diskutiert. Eine prototypische Umsetzung wurde separat evaluiert.

STICHWÖRTER

Industrie 4.0, Künstliche Intelligenz (KI), Messen/Steuern/Regeln

Autonomous cyber defence – Methodology for risk-based autonomous cyber defence with LLM and RAG

ABSTRACT This article presents a methodology for autonomous cyber defence in Industry 4.0 environments, linking organisational risk management with generative AI. Combining LLMs, RAG, and standardised SOPs creates, a structured decision-making process that enables real-time response and traceable mitigation. Conceptual foundations, technical implications, and practical potential are discussed. A prototype implementation was evaluated separately.

1 Einleitung

Der zunehmende Vernetzungsgrad industrieller Fertigungsumgebungen – insbesondere in hochautomatisierten Gigafabriken – schafft neue Angriffspunkte für Cyberbedrohungen. In modernen Produktionslinien kommunizieren hunderte cyber-physischer Systeme (CPPS) über ein konvergentes IT/OT-Backbone, das hohe Anforderungen an Sicherheit, Verfügbarkeit und Reaktionsgeschwindigkeit stellt [1, 2].

Zugleich steigt der Schaden durch Cyberangriffe auf deutsche Industrieunternehmen kontinuierlich – allein für 2024 wird er auf über 260 Milliarden Euro geschätzt [3]. Fälle wie Stuxnet, Triton/Trisis, Colonial Pipeline oder LockerGoga zeigen exemplarisch, dass gezielte Angriffe nicht nur Informationssicherheit, sondern auch Produktionssicherheit und in kritischen Fällen sogar die physische Sicherheit von Menschen gefährden können [4–7].

Bestehende Sicherheitslösungen wie Firewalls, IDS/IPS (Intrusion Detection Systems / Intrusion Prevention Systems) oder Endpoint Detection and Response (EDR) sind primär auf Alarmierung ausgelegt und setzen nur begrenzte automatisierte Gegenmaßnahmen um [8]. Die Reaktionszeiten menschlicher Operatoren sind den Anforderungen industrieller Taktzeiten häufig nicht gewachsen. Vor diesem Hintergrund gewinnen autonome, KI-gestützte Abwehrsysteme zunehmend an Bedeutung. Im Fokus stehen Architekturen, die organisatorische Risikobewertungen (etwa gemäß ISO/IEC 27001) mit operativen Entscheidungs- und Handlungsfähigkeiten verknüpfen können [9].

Dieser Beitrag stellt einen methodischen Rahmen vor, der generative KI, insbesondere große Sprachmodelle (LLMs), mit dem Prinzip der Retrieval-Augmented Generation (RAG) und maschinenlesbaren Sicherheitsrichtlinien kombiniert. Ziel ist es, sicherheitsrelevante Ereignisse automatisiert auf Risiken abzubilden, passende Maßnahmen auszuwählen und über standardisierte Schnittstellen umzusetzen, unter Berücksichtigung von Auditierbarkeit, Normkonformität und Erklärbarkeit. Die konzeptionelle Basis sowie zentrale Herausforderungen und Lösungsansätze werden im Folgenden dargestellt. Eine exemplarische Umsetzung der Methodik wurde separat untersucht und ist nicht Gegenstand dieses Beitrags.

2 Verwandte Arbeiten

Die zunehmende Komplexität industrieller IT/OT (Information Technology / Operational Technology)-Architekturen spiegelt sich in den Anforderungen an Cybersicherheitslösungen wider. Verschiedene etablierte Rahmenwerke – etwa ISO/IEC 27001, IEC 62443 oder die NIST SP 800-Serie – betonen die Bedeutung mehrschichtiger, risikobasierter Schutzmechanismen für OT-Umgebungen. In der industriellen Praxis zeigen Studien aber, dass viele Produktionsstätten weiterhin auf reaktive und manuell gesteuerte Sicherheitsprozesse setzen, etwa durch passive Intrusion Detection oder starre SOC-Playbooks [10]. Die „Enisa” fordert daher explizit Echtzeit-Reaktionen, die sich an Risikobewertungen orientieren und in automatisierte Abläufe überführt werden können [11].

Gleichzeitig gewinnen große Sprachmodelle (LLMs) in der Cybersicherheitsforschung an Bedeutung. Sie wurden bereits erfolgreich für Log-Analyse, Incident Triage und teilweise automatisierte Systemadministration eingesetzt [12, 13]. Aufgrund ihrer Fähigkeit zur probabilistischen Schlussfolgerung und zur Generierung adaptiver Handlungspläne gelten LLMs als geeignet für dynamische Entscheidungsprozesse [14, 15]. Allerdings verbleibt in nahezu allen bisherigen Arbeiten der Mensch im Entschei-

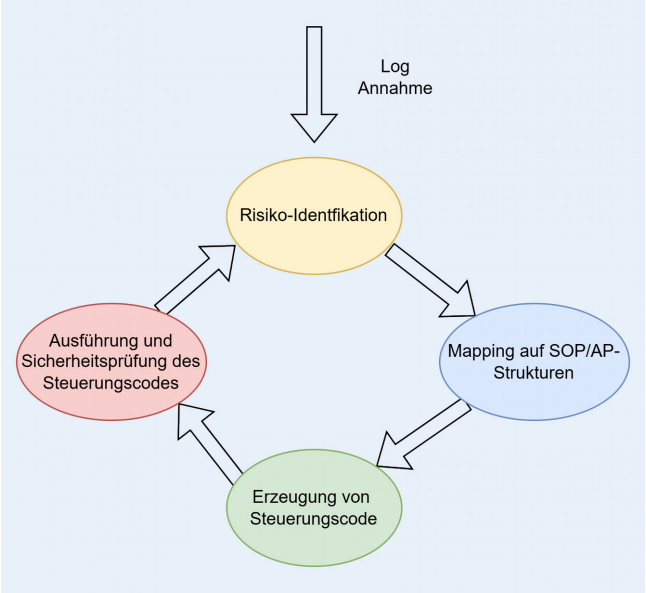

Bild 1 Konzeptionelle Darstellung des Entscheidungsregelkreises für autonome Cyberabwehr. *Grafik: Fraunhofer IPA*

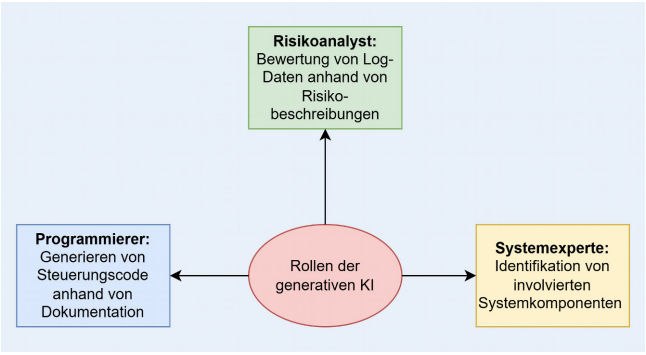

Bild 2 Konzeptionelle Darstellung der LLM (Large Language Model)-Rollen. *Grafik: Fraunhofer IPA*

dungs- oder Reaktionspfad, vor allem bei Maßnahmenumsetzung oder Policy-Abgleich.

Zur Reduktion „Halluzinationen“ und zur verbesserten Kontextintegration wurde die Retrieval-Augmented Generation (RAG) entwickelt. Dabei werden strukturierte oder semi-strukturierte Wissenselemente (wie API-Spezifikationen oder Richtlinienausschnitte) in den Promptfluss eingebettet [16]. Neuere Varianten wie Graph-RAG oder agentische RAG-Architekturen erlauben sogar Tool-Nutzung und komplexe Entscheidungsbäume [17], wurden jedoch bislang kaum auf sicherheitskritische Produktionskontexte übertragen. Mit dem Model Context Protocol (MCP) existiert zudem ein vielversprechender Standard zur Werkzeugintegration in heterogenen Infrastrukturen [18].

In bisherigen Forschungsarbeiten wurden die nachfolgenden Teilaspekte isoliert betrachtet und in keinen gemeinsamen Gesamtkontext überführt:

- die Formalisierung von Risikomodellen,
- die Nutzung von LLMs für Entscheidungsfindung, oder
- die technische Umsetzung von Maßnahmen (zum Beispiel via API (Application Programming Interface)-Tooling).

Diesen Gedankengang vollendend erschließt sich folgende Forschungslücke: Bislang fehlt ein integriertes Konzept, das alle drei Aspekte (risikobasierte Steuerung, generative Entscheidungslogik und auditable Maßnahmenumsetzung) in einem konsistenten methodischen Rahmen vereint. Genau diese Lücke adressiert der vorliegende Beitrag mit einem generischen Architektur- und Methodenvorschlag zur risikoorientierten KI-basierten Cyberabwehr.

3 Methodischer Rahmen zur risikobasierten Entscheidungsautomatisierung

Die methodische Grundlage des vorgestellten Ansatzes ist ein geschlossener Entscheidungs- und Handlungszyklus, der sicherheitsrelevante Ereignisse aus IT-/OT-Systemen mit strukturierten Risikobewertungen verknüpft und darauf aufbauend automatisierte Maßnahmen einleitet. Ziel ist es, generative KI nicht nur zur Analyse oder Beratung zu verwenden, sondern sie normgeleitet zur operativen Entscheidungsunterstützung und Ausführung zu befähigen. Eine Darstellung findet sich in **Bild 1**.

Im Zentrum steht ein mehrstufiger Prompt-Flow, der folgende Schritte umfasst:

1. Risiko Identifikation:
   Sicherheitsrelevante Logdaten oder SIEM (Security Information and Event Management)-Warnmeldungen werden über standardisierte Schnittstellen entgegengenommen. Die Daten dienen als Ausgangspunkt für eine risikobasierte Interpretation, wobei eine vordefinierte Risikomatrix (zum Beispiel ISO/IEC 27001-konform) als Grundlage dient.
2. Mapping auf SOP/AP-Strukturen:
   Das System identifiziert auf Basis des erkannten Risikos die zugehörige Standardarbeitsanweisung (SOP) und den Aktionsplan (AP). Diese Dokumente sind maschinenlesbar kodiert und beschreiben sowohl technische als auch organisatorische Maßnahmen.
3. Erzeugung von Steuerungscode:
   Der LLM-basierten Entscheidungsinstanz wird die vollständige Aufgabenbeschreibung samt technischer Dokumentation (zum Beispiel API-Spezifikationen via OpenAPI oder Freitext) zur Verfügung gestellt. Daraus erzeugt sie ausführbaren Steuerungscode, der konkrete Abwehrmaßnahmen über kontrollierte Schnittstellen einleitet.
4. Ausführung und Sicherheitsprüfung des Steuerungscodes:
   Vor der Ausführung erfolgt eine policy-basierte Prüfung des generierten Codes. Nur Maßnahmen, die vordefinierte, erlaubte Operationen abbilden (wie das Blockieren von IP-Adressen oder das Quarantänisieren von Systemkomponenten), werden automatisiert ausgeführt.

Zur Erhöhung der Robustheit gegenüber Halluzinationen und Prompt-Injections wird der gesamte Prompt-Flow durch Retrieval-Augmented Generation (RAG) unterstützt: Relevante Policies, API-Beschreibungen und Systemzustände werden bei Bedarf aus einer Wissensbasis eingebunden. Eine Darstellung der LLM-Rollen findet sich in **Bild 2**.

Ein solches Verfahren erfordert eine api-zentrierte Infrastruktur mit klar definierter Aktorenschnittstelle, eine standardisierte Darstellung von Sicherheitsrichtlinien und SOPs und die Fähigkeit, vertrauenswürdig generierten Code kontrolliert auszuführen und zu dokumentieren.

Ein prototypischer Aufbau, in welchem diese Methodik unter realistischen Bedingungen getestet wurde, ist im Rahmen eines separaten Projekts dokumentiert [19]. Die nachfolgenden Abschnitte analysieren auf Basis dieses methodischen Rahmens zentrale Anforderungen, Gestaltungsempfehlungen und offene Herausforderungen.

4 Erkenntnisse aus der Anwendung des Methodenkonzepts

Zur Validierung der konzipierten Methodik wurde der Ansatz in einem repräsentativen, isolierten IT/OT-Umfeld exemplarisch umgesetzt und durch kontrollierte Angriffs- und Beobachtungsszenarien bewertet. Dabei konnten die folgenden zentralen Beobachtungen abgeleitet werden, die Rückschlüsse auf die Praxistauglichkeit des Konzepts zulassen.

4.1 Risikoerkennung als zentraler Erfolgsfaktor

Das methodische Herzstück des Ansatzes liegt in der zuverlässigen Ableitung passender Risikoeinstufungen auf Basis dynamischer Ereignisdaten. In den durchgeführten Testsituationen identifizierte das System wiederholt das jeweils zutreffende Risiko und leitete daraus korrekt die passende Kombination aus Standardarbeitsanweisung (SOP) und Aktionsplan (AP) ab. Dies bildet die Grundlage für die Effektivität der nachfolgenden Entscheidungen und Interventionen.

Bei der praktischen Evaluierung wurden 2261 Testläufe durchgeführt. Diese bestanden aus 1360 Portscans, 461 Malwareinfektionen, 300 gutartigen Tests und 140 Tests gegen den Zufall. Von den 2261 Tests waren 137 Testläufe fehlerhaft, da diese in ein Timeout liefen. Dies lag an der fehlenden Parallelisierungsverarbeitungsfähigkeit in der Kommunikation mit dem LLM. Die restlichen 2124 Tests zeigten, dass das LLM die korrekten Risiken identifizierte. Wobei 7 Tests in der nachfolgenden Sektion gezielt behandelt werden [20].

4.2 Geringe Anfälligkeit für Fehlalarme

Im Rahmen realistischer Datenströme (wie SSH-Logins, Heartbeats, Systemlogs) zeigten sich nur vereinzelt Situationen, in denen das System eine menschliche Freigabe anregte. Bemerkenswert ist dabei, dass keinerlei falsche Maßnahmen eingeleitet wurden – stattdessen entschied sich das System in mehrdeutigen Situationen korrekterweise systematisch für Nichtstun oder Eskalation.

Dieses Verhalten wurde speziell bei den 300 gutartigen Tests überprüft. Von den 300 Testläufen waren 7 Datensätze vorhanden, bei denen das LLM einer Überprüfung durch den Menschen einforderte. Des Weiteren waren hier 15 Errors aufgrund der Parallelisierungsthematik vorhanden. Die restlichen 278 Tests zeigten, dass das LLM gutartige Logs erkennt und keine operativen Maßnahmen durchführt. Daher wird festgehalten, dass das System nicht über-reaktiv ist und gefährliche Maßnahmen einleitet, obwohl kein Grund dafür besteht [20].

4.3 Zeitverhalten in Reaktionsketten

Die Methode erwies sich in den getesteten Szenarien als reaktionsschnell: Entscheidungs- und Interventionsketten, von der Risikobewertung bis zur ausgeführten Gegenmaßnahme, lagen im Bereich unter einer Minute bei leichten Bedrohungen (zum Beispiel Port-Scans) und unter zehn Minuten bei aufwendigeren Maßnahmen (zum Beispiel Quarantäne und Neuaufsetzen von Systemen). Dieses Verhalten entspricht den Erfordernissen moderner Incident-Response-Modelle und liegt deutlich unter dem „1–10–60“-Benchmark der Branche. Es sei allerdings angemerkt, dass diese Werte selten in der Praxis erreicht werden [21].

4.4 Zuverlässige Ausführung operativer Maßnahmen

Die generierten Reaktionen, zum Beispiel Firewall-Regeln, VM-Isolationen oder Service-Redeployments, wurden konsistent korrekt umgesetzt. Die Verwendung kontrollierter Schnittstellen und ein vorgelagerter Sicherheitsscan des generierten Codes (Whitelisting zulässiger Operationen) sorgten für ein hohes Maß an Ausführungssicherheit.

4.5 Methodische Relevanz

Die Ergebnisse zeigen, dass der vorgestellte methodische Ansatz, vor allem die risikogeführte Auswahl von Maßnahmen und deren kontextuelle Generierung, geeignet ist, zentrale Schwächen bestehender sicherheitstechnischer Lösungen in IT/OT-Umgebungen zu adressieren. Insbesondere die enge Kopplung von Sicherheitsrichtlinien, Telemetrie, und Aktorik in einem geschlossenen Regelkreis erweist sich als vielversprechendes Gestaltungsprinzip.

5 Diskussion

Die Analyse der experimentellen Ergebnisse zeigt, dass das untersuchte Methodenkonzept wesentliche Anforderungen an eine autonome IT/OT-Abwehr erfüllt: Sicherheitsvorfälle werden risikobasiert erkannt, eingeordnet und mit konkreten technischen Maßnahmen beantwortet – vollständig automatisiert und innerhalb kurzer Zeit. Im Folgenden werden zentrale Befunde kontextualisiert und technische sowie methodische Implikationen diskutiert.

5.1 Einfluss auf die betriebliche Verfügbarkeit

Cybersicherheitsvorfälle beeinträchtigen unmittelbar die Verfügbarkeit industrieller Anlagen. Die im Testsystem gemessenen Reaktionszeiten liegen deutlich unter den branchenüblichen Benchmarks für manuelle SOC-Prozesse. Daraus ergibt sich ein hohes Potenzial, sicherheitsbedingte Ausfallzeiten zu reduzieren und die Gesamtanlageneffektivität (OEE) zu verbessern, vorausgesetzt, die automatisierte Reaktion bleibt technisch korrekt und organisatorisch legitimiert.

5.2 Steuerbarkeit durch strukturierte Entscheidungslogik

Die risikoorientierte Entscheidungslogik basiert auf der strukturierten Verknüpfung von Bedrohungen, Aktionsplänen und Standardarbeitsanweisungen. Dies erlaubt eine gezielte Steuerung der LLM-basierten Entscheidungsfindung. Die explizite Differenzierung zwischen autonom durchführbaren und menschlicher

Freigabe schafft Transparenz und Auditierbarkeit – auch in sicherheitskritischen Szenarien.

5.3 Abhängigkeit vom Modellverhalten

Die Zuverlässigkeit der Entscheidungs- und Umsetzungsschritte hängt maßgeblich vom eingesetzten Modell ab. Kommerzielle Modelle lieferten konsistente Ergebnisse, während kleinere Open-Source-Modelle häufig an Syntaxfehlern oder Missverständnissen der Dokumentation scheiterten. Hinzu kommt, dass einige Modelle sicherheitsbezogene Aufgaben fälschlich als „kritisch" einstuften und deshalb keine Ausgaben generierten – ein Effekt der inhaltlichen Filtermechanismen, die nicht für industrielle Anwendungen optimiert sind. Der beschriebene Entscheidungsrahmen ist prinzipiell modellunabhängig und lässt sich auch auf künftige, lokal feinabgestimmte Modelle übertragen.

5.4 Umgang mit unklaren Vorgaben

Ein auffälliger Aspekt ist die Fähigkeit des Systems, bei unvollständigen oder offen formulierten Anweisungen eigene, technisch plausible Lösungen zu entwickeln. In mehreren Fällen wurde eine Maßnahme korrekt abgeleitet, obwohl die zugehörige SOP keine Details zur Umsetzung enthielt. Dies eröffnet Chancen für die Reaktion auf unbekannte Angriffsszenarien, stellt aber auch Anforderungen an die Spezifikationstiefe der SOPs und die Akzeptanz solcher autonomen Interpretationen im Betrieb.

5.5 Abhängigkeit von Schnittstellen und Automatisierung

Die technische Wirksamkeit des Systems setzt standardisierte Schnittstellen, maschinenlesbare Dokumentation und eine durchgängig automatisierbare Infrastruktur voraus. In vielen industriellen Umgebungen sind diese Voraussetzungen derzeit nur eingeschränkt gegeben. Dies limitiert kurzfristig die Übertragbarkeit auf bestehende Produktionsnetze und verdeutlicht den Bedarf an organisationsweitem Infrastruktur-Redesign zur Integration autonomer Schutzsysteme.

6 Ausblick

Die Proof-of-Concept-Implementierung lieferte überraschend gute Resultate. Da es sich um eine Laborumgebung handelte, in der die Methodik praktisch getestet wurde, ist erst durch eine Überführung in ein operationelles Umfeld eine fundierte Aussage über den Erfolg möglich.

Anhand der durchgeführten Evaluation ergeben sich mehrere Anschlussfragen für künftige Arbeiten:

- Skalierung und Stresstest: Die getestete Architektur soll im Rahmen eines 24-Stunden-Dauerbetriebs unter realitätsnaher Last in einem erweiterten Testnetz mit OpenStack- und Proxmox-Knoten evaluiert werden.
- Abdeckung weiterer Angriffsszenarien: Zukünftige Tests sollen auch Angriffe berücksichtigen, die bislang nicht Teil der Risikomatrix sind, um die kreativen Fähigkeiten der LLM-Komponente zur Gefahrenabwehr besser einschätzen zu können.
- Modellvergleich und Feinabstimmung: Es ist zu prüfen, ob aktuelle Open-Source-Modelle durch Feintuning auf Sicherheitsdaten mit kommerziellen Lösungen gleichziehen können, insbesondere hinsichtlich Konsistenz und Codequalität.
- Technische Weiterentwicklung: Der bisherige Prompt-Flow-Ansatz könnte durch eine agentenbasierte Architektur ergänzt oder ersetzt werden, was die Autonomie weiter erhöhen, aber auch die Komplexität steigern würde.
- Standardisierung und Interoperabilität: Der systematische Einsatz standardisierter Schnittstellen wie dem Model Context Protocol (MCP) soll die Übertragbarkeit auf unterschiedliche Umgebungen erleichtern.
- Erweiterung regulatorischer Abdeckung: Perspektivisch wird angestrebt, auch Anforderungen aus IEC 62443, NIS-2 und vergleichbaren Regelwerken formal in die Entscheidungslogik einzubinden.
- Sicherheitsbewertung: Die Robustheit gegenüber Prompt-Injections und adversarial-Prompts erfordert eigene Testkampagnen unter Red-Team-Bedingungen.
- Governance und Ethik: Letztlich ist zu klären, inwieweit autonome Entscheidungen in "Kritis"-Umgebungen rechtlich zulässig und operativ verantwortbar sind, vor allem im Hinblick auf menschliche Kontrollinstanzen, Zertifizierbarkeit und Fail-Safe-Konzepte.

Damit wird der Weg geebnet für selbstverteidigende Produktionssysteme, in denen Sicherheit kein nachgelagerter Prozess, sondern ein integraler Teil des industriellen Regelkreises ist.

Die vorgestellte Methodik verdeutlicht exemplarisch den Paradigmenwechsel hin zu selbstverteidigenden Systemarchitekturen, in denen Cyberabwehr kein externer Prozess mehr ist, sondern integraler Bestandteil industrieller Intelligenz.

7 Fazit

Die vorliegende Untersuchung zeigt, dass ein lokal betriebener, risikobasierter LLM-Controller in der Lage ist, sicherheitsrelevante Vorfälle in konvergenten IT/OT-Umgebungen automatisiert zu erkennen, einzuordnen und mit wirksamen Gegenmaßnahmen zu beantworten. Die getestete Architektur kombinierte eine ISO-27001-konforme Risikomatrix, eine Retrieval-augmented Generierung sowie API-basierte Aktoren für Live-Reaktionen.

In einem industriellen Testbed konnten alle simulierten Port-Scan- und Malware-Angriffe in unter zehn Minuten erfolgreich eingedämmt werden – bei einer Falsch-Positiv-Rate von lediglich 2,8 % und vollständiger Nachvollziehbarkeit aller Schritte. Die generierten Gegenmaßnahmen waren technisch korrekt, ausführbar und dokumentiert.

Die Ergebnisse zeigen, dass eine risikobewusste Automatisierung mit LLMs das Potenzial hat, die durchschnittliche Reaktionszeit in OT-Netzen drastisch zu verkürzen und gleichzeitig regulatorische Anforderungen an Auditierbarkeit, Nachvollziehbarkeit und Standardkonformität zu erfüllen.

Für künftige Forschungsfelder, speziell im Hinblick auf Governance und Ethik, lassen sich das Potenzial und die Gefahren autonomer KI-Systeme bereits erahnen. Diesen Gedanken folgend eröffnet sich ein Spannungsfeld zwischen proaktiven, vollständig handlungsfähigen Systemen und dem Menschen. Zur Risikoreduktion sind vorausschauende und kompromittierungslose Techniken zu entwickeln und zu implementieren.

FÖRDERHINWEIS

Diese Arbeit wurde gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) für das Projekt „DigiBattPro 4.0" unter dem Förderkennzeichen 03XP0374C [22].

DANKSAGUNG

Die konzeptionellen Grundlagen dieses Beitrags wurden im Rahmen der Masterarbeit von Maximian Geyer an der Universität Stuttgart entwickelt [20]. Die technische Umsetzung und Evaluation des beschriebenen Systems wurde in einem begleitenden Fachartikel dokumentiert, der zur Konferenz *PAIS 2025* eingereicht wurde [19].

LITERATUR

[1] Degen, F.: Lithium-ion battery cell production in Europe: Scenarios for reducing energy consumption and greenhouse gas emissions until 2030. Journal of Industrial Ecology (2023) 3, pp. 964–976

[2] Stock, D.; Bauernhansl, T.; Weyrich, M. et al.: System Architectures for Cyber-Physical Production Systems enabling Self-X and Autonomy. IEEE International Conference on Emerging Technologies and Factory Automation (ETFA) 2020 (2020), pp. 148–155

[3] Bitkom: Financial damage from cybercrime in Germany in 2024 (in billion euros). Internet: www.statista.com/statistics/1360289/financial-damage-cyber-crimes-germany/. Zugriff am 12.08.2025

[4] Beerman, J.; Berent, D.; Falter, Z. et al.: A Review of Colonial Pipeline Ransomware Attack. International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW) 2023 (2023) 23, pp. 8–15

[5] Kushner, D.: The Real Story of Stuxnet. IEEE Spectrum (2013) 50, pp. 48–53

[6] Adamov, A.; Carlsson, A.; Surmacz, T.: An Analysis of LockerGoga Ransomware. IEEE East-West Design & Test Symposium (EWDTS) (2019), pp. 1–5

[7] Mekdad, Y.; Bernieri, G.; Conti, M. et al.: A threat model method for ICS malware: the TRISIS case. ACM International Conference on Computing Frontiers (2021), pp. 221–228

[8] Scarfone, K.; Mell, P.: Guide to Intrusion Detection and Prevention Systems (IDPS). NIST Special Publication 800–94 (2007), doi.org/10.6028/NIST.SP.800–94

[9] DIN EN ISO/IEC 27001: Informationssicherheit, Cybersicherheit und Datenschutz – Informationssicherheitsmanagementsysteme – Anforderungen (ISO/IEC 27001:2022). Deutsche Fassung EN ISO/IEC 27001:2023, Ausgabe 2024–01

[10] Dhirani, L.; Armstrong, E.; Newe, T.: Industrial IoT, Cyber Threats, and Standards Landscape: Evaluation and Roadmap. Sensors 2021 (2021) 21, #3901

[11] European Union Agency for Cyber-Security: ENISA Threat Landscape for Industrial Control Systems 2024. Internet: www.enisa.europa.eu/publications/enisa-threat-landscape-2024 . Zugriff am 12.08.2025

[12] Cao, C.; Wang, F; Lindley, L. et al.: Managing Linux servers with LLM-based AI agents: An empirical evaluation with GPT4. Machine Learning with Applications 17 (2024) #100570, doi.org/10.1016/j.mlwa.2024.100570

[13] Xu, H.; Wang, S.; Li, N. et al.: Large Language Models for Cyber Security: A Systematic Literature Review. arXiv preprint 2405.04760, doi.org/10.48550/arXiv.2405.04760

[14] Liu, O.; Fu, D.; Yogatama, D. et al.: DeLLMa: Decision Making Under Uncertainty with Large Language Models. arXiv preprint 2024, doi.org/10.48550/arXiv.2402.02392

[15] Li, S.; Puig, X.; Paxton, C. et al.: Pre-Trained Language Models for Interactive Decision-Making. arXiv preprint 2022, doi.org/10.48550/arXiv.2202.01771

[16] Gokcimen, T.; Das, B.: A novel system for strengthening security in large language models against hallucination and injection attacks with effective strategies. Alexandria Engineering Journal 123 (2025), pp. 71–90

[17] Wang, W.; Wang, Y.; Joty, S. et al.: RAP-Gen: Retrieval-Augmented Patch Generation with CodeT5 for Automatic Program Repair. ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering (2023) 31, pp. 146–158

[18] Hou, X.; Zhao, Y.; Wang, S. et al.: Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. arXiv preprint 2025, doi.org/10.48550/arXiv.2503.23278

[19] Geyer, M.; Schwab, J.: Risk-Aware Autonomous Defence with Generative AI: Ethical and Accountability Challenges in Cyber-Physical Infrastructures. In: Workshop on AI in Security and Defense (AI4SD) at ECAI 2025, OpenHSU Proceedings, 2025

[20] Geyer, M.: Autonome IT-/OT-Verteidigung in der Intralogistik. Entwicklung eines resilienten KI-basierten Sicherheitssystems für die vernetzte Batteriezellenproduktion. Masterarbeit, Universität Stuttgart, 2025

[21] CrowdStrike: The 1/10/60 Minute Challenge: A Framework for Stopping Breaches Faster. Stand: 2025. Internet: https://www.crowdstrike.com/en-us/resources/crowdcasts/the-1–10–60-minute-challenge-a-framework-for-stopping-breaches-faster/. Zugriff am 12.08.2025

[22] Fraunhofer IPA: DigiBattPro 4.0 – BMBF: Digitalisierungslösungen und Materialentwicklung für die Batterieproduktion. Stand: 2025. Internet: www.ipa.fraunhofer.de/de/referenzprojekte/DigiBattPro40-BMBF.html. Zugriff am 12.08.2025

Maximian Geyer
maximian.geyer@ipa.fraunhofer.de

Jannik Schwab, M.Sc.

Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstr. 12, 70569 Stuttgart
www.ipa.fraunhofer.de