Bachelorarbeit

Autonome KI-Agenten in der Cyberkriminalität
Risiken, Szenarien und Handlungsoptionen für Unternehmen

Abschlussarbeit für die Bachelorprüfung
zur Erlangung des Grades „Bachelor of Science“

von

Mohammadreza Tavakoli Dargani
Matrikelnummer: ******
******

aus

Kerpen

Erstgutachter: Prof. Dr. rer. pol. Matthias Mehrtens
Zweitgutachter: Prof. Dr.-Ing. René Treibert

Hochschule Niederrhein
Campus Mönchengladbach
Fachbereich 08 | Wirtschaftswissenschaften
Studiengang: **Cyber Security Management**

Fachsemester: 7
2025

Inhaltsverzeichnis

1. Abkürzungsverzeichnis

z.B.	zum Beispiel
KI	künstliche Intelligenz
bzw.	beziehungsweise
bspw.	beispielsweise
u. a.	unter anderem
EU AI Act	European Artificial Intelligence Act
ENISA	European Union Agency for Cybersecurity
DSGVO	Datenschutz-Grundverordnung
MLOps	Machine Learning Operations
HLEG	High-Level Expert Group on Artificial Intelligence
LLM	Large Language Model
NLP	Natural Language Processing
DDoS	Distributed Denial of Service
BKA	Bundes Kriminalamt
BSI	Bundesamt für Sicherheit in der Informationstechnik
BEC	Business-Email-Compromise
OSINT	Open-source intelligence
CaaS	Cybercrime as a Service
DBIR	Data Breach Investigations Report
CVE	Common Vulnerabilities and Exposures
DNS	Domain Name Server
RDP	Remote Desktop Protocol
CTF	Capture the flag
OWA	Outlook web app
DMARC	Domain-based Message Authentication, Reporting and Conformance
PoC	Proof of Concept
EASM	External Attack Surface Management

SIEM	Security Information and Event Management
KMU	Kleine und mittlere Unternehmen

2. Abbildungsverzeichnis

3. Einleitung

Wir leben in einer Zeit, in der die Vorstellung der Zukunft aufgrund des Fortschritts der Technologien immer schwieriger wird. Der technologische Fortschritt im Bereich der Informationstechnologie schreitet in einem derart hohen Tempo voran, dass selbst kurzfristige Entwicklungen oft unvorhersehbar bleiben. Ein wesentlicher Treiber dieser Entwicklung ist das exponentielle Wachstum neuer Technologien. Globale Technologiekonzerne wie z.B. Google, Microsoft, Apple oder Samsung sowie zahlreiche aufstrebende Start-Ups bestreben es, mit dieser Entwicklung Schritt zu halten, um Marktanteile zu sichern und Nutzerbindungen zu stärken, um mehr User zu gewinnen oder keine User zu verlieren. Um an der Spitze der Innovation zu bleiben und unseren Wettbewerbsvorteil zu sichern, ist Schnelligkeit entscheidend.

Dabei entsteht ein Spannungsfeld zwischen den Chancen bzw. Vorteilen und den Risiken neuer Technologien. Während auf der einen Seite innovative Anwendungen, Effizienzsteigerungen, Komfortabilität und neue Geschäftsmodelle entstehen, kommen auf der anderen Seite zunehmend sicherheitsrelevante und ethische Fragestellungen in Betracht. Dieses Spannungsfeld bildet den Ausgangspunkt der vorliegenden Arbeit, wobei sie sich auf Risiken, Szenarien und Handlungsoptionen für Unternehmen im Kontext von KI-Agenten konzentriert.

Diese Arbeit konzentriert sich auf autonome KI-Agenten, also Systeme, die in der Lage sind, menschliche Handlungen sowie das Verhalten in zunehmend komplexen Kontexten selbständig zu übernehmen. Erfolgreiche Technologien zeichnen sich i.d.R. dadurch aus, dass sie das menschliche Leben komfortabler gestalten. Künstliche Intelligenz verfolgt dabei genau das Ziel, Aufgaben zu automatisieren, die bislang manuell durch den Menschen erledigt werden mussten, sei es das Verfassen von Texten, das Generieren von Bildern und Audiodateien, das Auswerten und Vergleichen komplexer Inhalte oder das Interpretieren verschiedenster Dateiformate im situativen Zusammenhang und vieles mehr, wobei der Einsatz von autonomen KI-Agenten immer vielfältiger wird.

Gerade weil viele dieser Tätigkeiten mit erheblichem zeitlichem Aufwand verbunden sind, werden solche Systeme zunehmend aus Gründen der Bequemlichkeit sowohl von Privatpersonen als auch innerhalb von Unternehmen eingesetzt. In diesem Zusammenhang ist ein gesellschaftlicher Trend zu beobachten, den der Soziologe Rainer Paris in seinem Werk „Theorie der Bequemlichkeit. Ein Entwurf“ wie folgt beschreibt:

„Die Gewohnheit ist verpönt und ihre sprichwörtliche Macht gefürchtet, wird sie doch verdächtigt, zwischen uns und unserem besseren Ich zu stehen. [...] Während wir ihren Zustand durchaus goutieren, suchen wir sie als Selbst- oder Fremdzuschreibung zu vermeiden: Wir möchten es uns bequem machen, [...].“[1]

Diese Aussage unterstreicht die zunehmende Bereitschaft vieler Menschen, im Alltag wie im Arbeitsleben zugunsten von Effizienz und Zeitersparnis sensible Daten wie z.B. intern vertrauliche Unternehmensdaten oder personenbezogene Daten preiszugeben. Dadurch entsteht eine neue Form der Verwundbarkeit auf organisatorischer Ebene, da Mitarbeiter aus Bequemlichkeit unbewusst sicherheitsrelevante Risiken erzeugen, und auf technischer Ebene, weil innovationsgetriebene Unternehmen unter hohem Zeitdruck arbeiten und dabei nicht selten grundlegende Sicherheitsmechanismen vernachlässigen.

[1] Paris, Rainer (o. D.): Bequemlichkeit und Gewohnheit, [online] https://www.soziopolis.de/bequemlichkeit-und-gewohnheit.html.

Beide Faktoren eröffnen neue Angriffsflächen, die gezielt durch autonome, KI-gestützte Systeme ausgenutzt werden könnten.

3.1. Problemstellung

Mit dem zunehmenden Einsatz von künstlicher Intelligenz in alltäglichen wie auch unternehmerischen Prozessen eröffnen sich nicht nur neue Chancen, sondern auch bislang kaum absehbare Risiken, insbesondere dann, wenn diese Technologie autonom agiert. Autonome KI-Agenten, die in der Lage sind, Informationen zu beschaffen, Entscheidungen zu treffen und eigenständig Aktionen auszuführen, überschreiten eine Schwelle: Sie agieren nicht nur als Werkzeuge menschlicher Steuerung, sondern entwickeln sich zu teilweise eigenständigen Akteuren im digitalen Raum.

Während Unternehmen diese Systeme häufig zur Automatisierung, Effizienzsteigerung und Prozessoptimierung einsetzen, machen sich gleichzeitig auch Cyberkriminelle deren Fähigkeiten zunutze, zum Beispiel für automatisierte Angriffsketten, Social-Engineering-Taktiken oder zur schnellen Ausnutzung öffentlich bekannter Schwachstellen. Aufgrund der Offenheit vieler dieser Tools, der geringen Einstiegshürden und der stetig verbesserten Leistungsfähigkeit entstehen neuartige Bedrohungsszenarien, auf die bestehende Sicherheitsstrategien kaum vorbereitet sind.
Die Problematik verschärft sich dadurch, dass viele Unternehmen weder über die nötige Sensibilisierung noch über geeignete Kontrollmechanismen verfügen, um mit den Risiken autonomer KI-Systeme umzugehen. Organisatorisch sind oftmals keine Prozesse etabliert, die menschliches Fehlverhalten im Umgang mit KI (z. B. unreflektierte Nutzung oder Datenfreigabe) unterstützen oder fördern. Auf technischer Ebene mangelt es an Schutzmechanismen, die KI-basierte Angriffe frühzeitig erkennen oder abwehren können. Gleichzeitig stehen Unternehmen unter Innovationsdruck, was in der Praxis häufig dazu führt, dass Sicherheitsfragen hinter wirtschaftlichen Zielen zurückstehen.

Die zunehmende technologische Bequemlichkeit, der Mangel an fundierter Aufklärung sowie die hohe Dynamik technischer Entwicklungen schaffen eine kritische Angriffsfläche – nicht nur für individuelle Anwender, sondern in besonderem Maße für Unternehmen. Dieses Spannungsfeld verleiht der Thematik sowohl aus sicherheitsstrategischer als auch aus gesellschaftlich-organisatorischer Sicht eine hohe Relevanz und bildet den Ausgangspunkt der vorliegenden Untersuchung.

3.2. Zielsetzung der Arbeit und Forschungsfrage

Ziel dieser Arbeit ist es, die sicherheitsrelevanten Risiken autonomer KI-Agenten aus einer angriffszentrierten Perspektive systematisch zu untersuchen. Dabei soll aufgezeigt werden, wie solche Systeme von Cyberkriminellen eingesetzt werden können, um gezielte und teilweise automatisierte Angriffe auf Unternehmen durchzuführen. Im Fokus steht nicht nur die technische Leistungsfähigkeit dieser Agenten, sondern insbesondere ihre Wirkung auf organisatorische Strukturen, Prozesse und menschliches Verhalten innerhalb von Unternehmen.

Die Arbeit verfolgt einen praxisnahen und interdisziplinären Ansatz: Anhand ausgewählter Angriffsszenarien werden konkrete Schwachstellen analysiert, die durch den Einsatz autonomer KI-Systeme entstehen können – sowohl auf

technischer als auch auf menschlich-organisatorischer Ebene. Darauf aufbauend werden Handlungsoptionen für Unternehmen entwickelt, die zur Risikominimierung beitragen und als Grundlage für strategische Entscheidungen im Umgang mit KI dienen können.

Nicht Ziel dieser Arbeit ist die technische Implementierung oder Bewertung konkreter KI-Modelle auf Code- oder Architekturebene. Ebenso werden keine umfassenden rechtlichen oder regulatorischen Ausarbeitungen vorgenommen; rechtliche Aspekte werden nur soweit berücksichtigt, wie sie für das Verständnis der Sicherheitslage notwendig sind. Auch wirtschaftliche Potenzialanalysen, etwa zur Rentabilität von KI-Systemen, sowie eine detaillierte ethisch-philosophische Diskussion der Technologieentwicklung fallen nicht in den direkten Fokus dieser Arbeit. Die Untersuchung konzentriert sich bewusst auf den sicherheitsstrategischen Blickwinkel aus Unternehmensperspektive.
Aus dieser Zielsetzung ergibt sich folgende zentrale Forschungsfrage:

Wie verändert der Einsatz autonomer KI-Agenten das Bedrohungspotenzial im Cyberraum aus Sicht des Unternehmensschutzes?

Zur weiteren Konkretisierung dieser Fragestellung werden folgende Unterfragen formuliert:

- Welche realistischen Angriffsszenarien mit autonomen KI-Agenten lassen sich derzeit identifizieren?
- Welche organisatorischen und technischen Schwächen werden durch solche Szenarien besonders angreifbar?
- Inwiefern tragen menschliche Faktoren wie Bequemlichkeit, mangelnde Awareness oder unreflektierte Nutzung zur Verwundbarkeit bei?
- Welche konkreten Schutzmaßnahmen und organisatorischen Anpassungen sind erforderlich, um Unternehmen gegen KI-gestützte Angriffe resilienter zu machen?

3.3. Aufbau der Arbeit

Diese Arbeit gliedert sich in insgesamt sieben Abschnitte.

Zunächst wird das Thema eingeführt und die Relevanz der Untersuchung dargelegt. Neben einer allgemeinen Einführung werden die Problemstellung, die Zielsetzung, die zentralen Forschungsfragen sowie der Umfang der Arbeit erläutert.

Im Anschluss werden die theoretischen und begrifflichen Grundlagen behandelt, die für das Verständnis der weiteren Analyse erforderlich sind. Dazu zählen insbesondere die Konzepte der künstlichen Intelligenz und der autonomen Agenten sowie grundlegende Aspekte der Cyberkriminalität und deren Überschneidungen. Ergänzend wird auf relevante rechtliche und ethische Rahmenbedingungen eingegangen.

Darauf folgt eine Auseinandersetzung mit der Rolle autonomer KI-Agenten im Kontext digitaler Bedrohungen. Es werden deren Funktionsweise, technische Möglichkeiten sowie Missbrauchsszenarien dargestellt. Relevante Forschungsansätze und Entwicklungen aus der Praxis werden ebenfalls berücksichtigt.

In einem weiteren Abschnitt werden realistische Angriffsszenarien aus Angreiferperspektive analysiert. Ziel ist es, konkrete Anwendungsmöglichkeiten autonomer KI-Systeme in Cyberangriffen aufzuzeigen und deren Kritikalität zu bewerten.

Daraufhin widmet sich die Arbeit den daraus resultierenden Risiken für Unternehmen. Es werden sowohl technische Schwachstellen als auch organisationale und menschliche Faktoren betrachtet, die durch den Einsatz solcher Systeme besonders angreifbar sind. Eine strukturierte Risikoanalyse ergänzt diesen Teil.

Auf dieser Grundlage werden anschließend Handlungsoptionen und Schutzmaßnahmen abgeleitet. Der Fokus liegt dabei auf organisatorischen Strategien, Awareness-Ansätzen sowie strukturellen und technischen Vorkehrungen zur Verbesserung der Unternehmenssicherheit gegenüber KI-basierten Bedrohungen.

Den Abschluss der Arbeit bildet ein Fazit mit einer Zusammenfassung der zentralen Erkenntnisse, einer Reflexion der Grenzen der Untersuchung sowie einem Ausblick auf mögliche zukünftige Entwicklungen in diesem Themenfeld.

4. Theoretische und technische Grundlagen

Um die sicherheitsrelevanten Herausforderungen autonomer KI-Agenten im Kontext der Cyberkriminalität fundiert untersuchen zu können, ist es zunächst notwendig, die zugrunde liegenden Begriffe und technischen Konzepte zu präzisieren. Dieses Kapitel schafft die erforderliche theoretische und terminologische Basis, um die nachfolgenden Analysen und Bewertungen fachlich einzuordnen.

Im ersten Abschnitt werden die grundlegenden Prinzipien künstlicher Intelligenz sowie die Funktionsweise und charakteristischen Merkmale autonomer KI-Agenten dargestellt. Daran anschließend erfolgt eine systematische Einordnung relevanter Erscheinungsformen der Cyberkriminalität, bevor die Schnittstellen zwischen KI-Systemen und kriminellen digitalen Aktivitäten herausgearbeitet werden. Ergänzend werden zentrale rechtliche, ethische und gesellschaftliche Rahmenbedingungen skizziert, die für ein ganzheitliches Verständnis der Thematik maßgeblich sind.

4.1. Künstliche Intelligenz und autonome Agenten

Künstliche Intelligenz hat sich in den vergangenen Jahren von einem abstrakten technologischen Konzept zu einem zentralen Bestandteil moderner Arbeits- und

Lebenswelten gewandelt. In Unternehmen kommen heute nicht mehr lediglich klassische Chatbots oder regelbasierte Automatisierungslösungen zum Einsatz, sondern zunehmend sogenannte autonome KI-Agenten. Diese stehen für eine neue Generation intelligenter Systeme, die in der Lage sind, Aufgaben eigenständig zu interpretieren, Handlungsstrategien zu entwickeln, geeignete Werkzeuge auszuwählen und Ergebnisse weitgehend unabhängig umzusetzen.

Autonome Agenten sind Softwarekomponenten, die ein vorgegebenes Ziel selbstständig erfassen, in sinnvolle Teilaufgaben strukturieren, relevante Informationen beschaffen, passende Tools nutzen und die Bearbeitung – mit keinem oder nur minimalem menschlichen Eingriff – vollständig durchführen können. Damit verschiebt sich der Schwerpunkt von einer rein reaktiven hin zu einer proaktiven, zielorientierten Aufgabenerfüllung. Unternehmen wie Salesforce, Microsoft, Google und Slack treiben diese Entwicklung seit den Jahren 2024/2025 intensiv voran.

4.2. Abgrenzung zu klassischen KI-Systemen

Im Unterschied zu klassischen KI-Agenten oder Chatbots, die in der Regel auf vordefinierte Eingaben reagieren, handelt ein autonomer Agent proaktiv und adaptiv. Er benötigt keine exakten Befehle, sondern orientiert sich an Zielvorgaben und entwickelt daraus eigene Aktionspläne. Während einfache Chatbots etwa eine FAQ-Antwort ausgeben oder ein Formular generieren, analysiert ein autonomer Agent den Gesamtzusammenhang, erfasst Kontextinformationen, priorisiert Aufgaben und entscheidet, welche Ressourcen oder Systeme er einbindet.

In der Praxis zeigt sich dies beispielsweise in der Weiterentwicklung des Slackbots zu einem „persönlichen Arbeitsagenten“: Dieser ist in der Lage, Unterhaltungen zu analysieren, Projektbriefings zusammenzufassen oder Meetings zu planen – alles im Kontext des jeweiligen Nutzers. Auch Microsoft hat mit dem „Agent Mode“ für den 365-Copilot die Grenze zwischen Assistenz und Autonomie überschritten. Der Agent erstellt nicht nur Berichte, sondern analysiert Datenquellen, bereinigt Datensätze und visualisiert Ergebnisse – ohne dass dafür eine manuelle Steuerung notwendig wäre.

4.2.1. Struktur und Funktionsweise autonomer Agenten

Die grundlegende Architektur eines autonomen Agenten basiert auf dem sogenannten Agenten-Loop – einem iterativen Ablaufmodell, das sich aus den fünf zentralen Phasen zusammensetzt:
(1) Zielverstehen,
(2) Planung,
(3) Ausführung,
(4) Reflexion,
(5) Ergebnisbereitstellung.

Diese Funktionsweise wird in Abbildung 1 („Was ist ein Agent?“) schematisch dargestellt. Der Agent startet mit der Wahrnehmung von Kontextinformationen – sei es durch Spracheingabe, Dokumentenanalyse oder Nutzungsdaten. Anschließend analysiert er mittels logischer Modelle und maschinellen Lernens die Anforderungen, plant die nächsten Schritte und ruft entsprechende Werkzeuge oder Datenbanken ab.

Wichtig ist: Nach jedem Schritt überprüft er das Zwischenresultat und passt gegebenenfalls seinen Plan an, bevor er mit dem nächsten Teilschritt fortfährt.

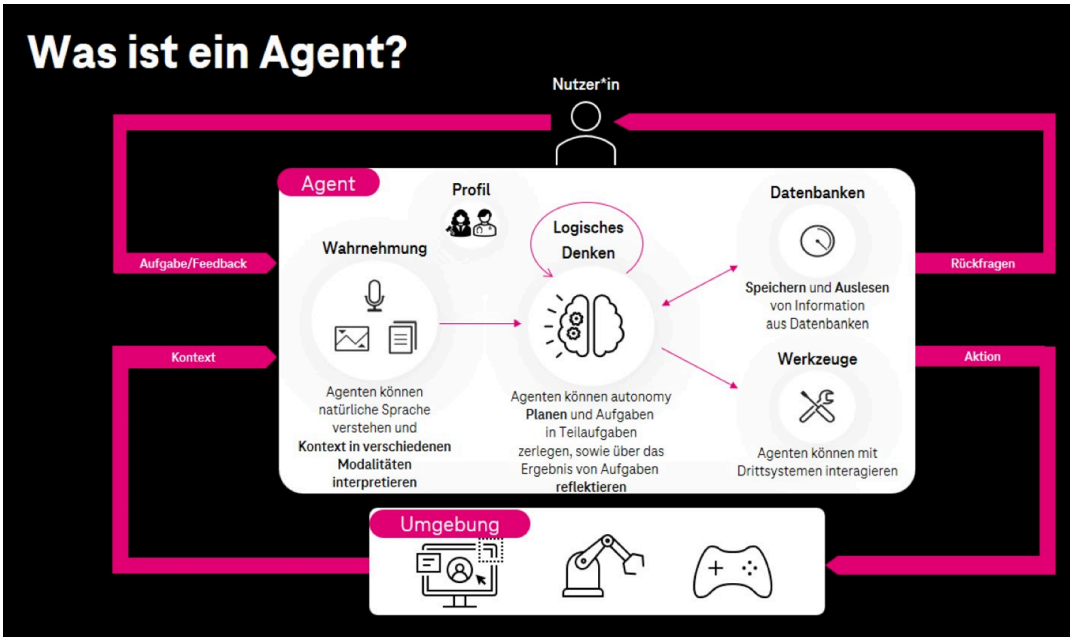

Abbildung 1
Deutsche Telekom MMS GmbH (2026): Autonome KI-Agenten: Selbstlernende KI-Systeme im Einsatz, Deutsche Telekom MMS GmbH, [online] https://www.telekom-mms.com/blog/artikel/detail/autonome-ki-agenten.

Diese iterative, lernfähige und anpassbare Struktur unterscheidet autonome Agenten nicht nur von einfachen Bots, sondern auch von klassischen RPA-Tools[2] oder regelbasierten Workflows. Letztere scheitern, sobald unvorhergesehene Situationen auftreten oder Inputdaten von der Norm abweichen.

4.2.2. Nutzerbedürfnisse und Vertrauensfragen

Mit der wachsenden Autonomie von KI-Agenten steigt auch der Anspruch an ihre Kontrollierbarkeit und Nachvollziehbarkeit. Eine Studie von YouGov (2025) zeigt, dass 45 % der befragten deutschen Nutzer:innen insbesondere darauf Wert legen, jederzeit nachvollziehen zu können, was die KI in ihrem Namen getan hat. 37 % wollen die Möglichkeit behalten, Aufgaben manuell zu übernehmen oder abzubrechen, während 36 % selbst darüber entscheiden möchten, welche Aufgaben die KI überhaupt übernehmen darf.

[2] RPA (Robotic Process Automation = Robotergesteuerte Prozessautomatisierung) ist eine Softwaretechnologie zur schnellen und zuverlässigen Automatisierung digitaliegebauter Aufgaben.

Abbildung 2
Deutsche Telekom MMS GmbH (2026): Autonome KI-Agenten: Selbstlernende KI-Systeme im Einsatz, Deutsche Telekom MMS GmbH, [online] https://www.telekom-mms.com/blog/artikel/detail/autonome-ki-agenten.

Diese Ergebnisse verdeutlichen, dass die Akzeptanz autonomer Systeme nicht allein von ihrer Effizienz, sondern vor allem vom Vertrauen in ihre Entscheidungen abhängt. Transparente Entscheidungslogiken, nachvollziehbare Systemreaktionen und Kontrollmechanismen („Human-in-the-Loop“) sind daher nicht nur gesetzlich gefordert (z. B. durch den EU AI Act), sondern auch aus praktischer Perspektive essenziell. Dies kann aber auf der anderen Seite daran liegen, dass die KI-Technologie noch in ihrer Anfangsphase eingeschätzt wird und nicht alle Ergebnisse der Wahrheit entsprechen oder dem am effizientesten Weg folgen. Diese Ergebnisse werden sich in Zukunft mit hoher Wahrscheinlichkeit ändern, da eine reife KI mehr Vertrauen gewinnt und nach langer Nutzung ein Vertrauen in die Werke und Ergebnisse der KI gebaut wird. Dies wird dazu führen, dass weniger Menschenaktion in die Aufgaben integriert wird und alles mit autonomen KI-Systemen automatisiert vorangeht.

4.2.3. Anwendungsfelder in der Unternehmenspraxis

Autonome Agenten finden bereits heute Anwendung in zahlreichen Unternehmensbereichen. Besonders hohes Potenzial wird ihnen im Kundenservice (57 %), im Backoffice und HR (42 %) sowie im Marketing und in der Vertragsabwicklung zugesprochen, wie eine Erhebung unter 162 IT-Entscheider:innen zeigt.

In der Praxis bedeutet das beispielsweise:

- **Im Kundenservice:** Bearbeitung von Anfragen ohne Wartezeit, automatische Klassifizierung von Tickets, Statusmeldungen an Kund:innen

- **Im IT-Support:** Diagnose technischer Störungen durch Logfile-Analyse, Neustart von Systemdiensten, automatische

Dokumentation

- **Im Personalwesen:** automatisiertes Onboarding neuer Mitarbeitender, Verarbeitung von Reisekosten, digitale Urlaubsfreigaben

- **Im Vertrieb & Marketing:** Planung von Kampagnen, Generierung von Inhalten, Analyse von Kundenverhalten

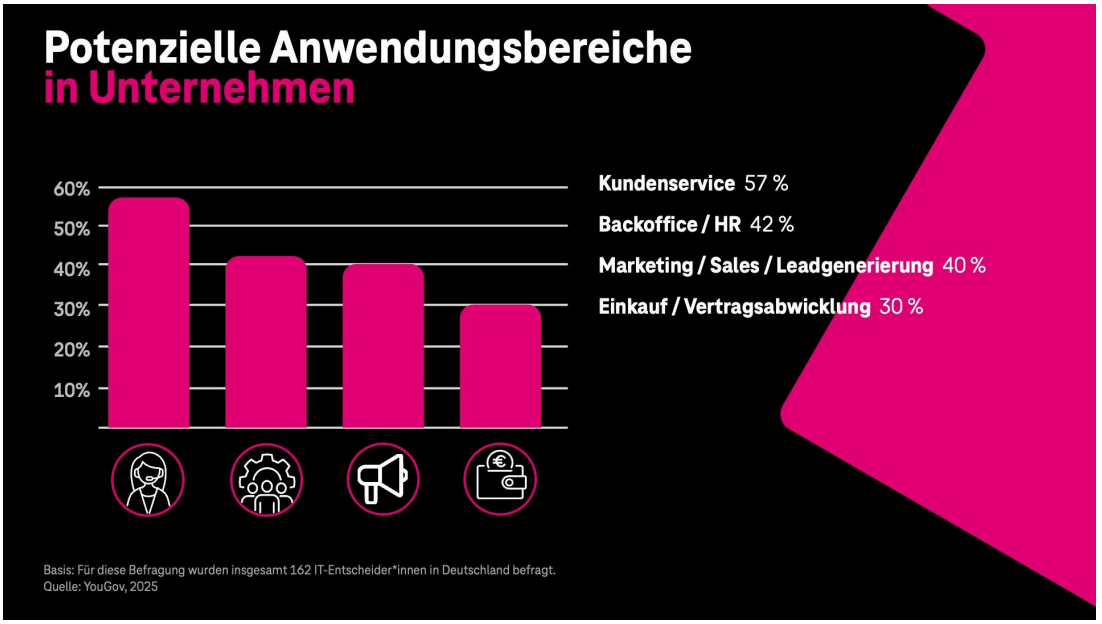

Abbildung 3
Deutsche Telekom MMS GmbH (2026): Autonome KI-Agenten: Selbstlernende KI-Systeme im Einsatz, Deutsche Telekom MMS GmbH, [online] https://www.telekom-mms.com/blog/artikel/detail/autonome-ki-agenten.

Diese Einsatzbereiche zeigen: Autonome Agenten sind nicht auf einfache Aufgaben beschränkt, sondern können komplexe, kontextabhängige Prozesse übernehmen – vorausgesetzt, sie sind korrekt integriert und abgestimmt.

4.2.4. Herausforderungen bei Implementierung und Steuerung

Der erfolgreiche Einsatz autonomer Agenten erfordert jedoch mehr als nur technologische Infrastruktur. Unternehmen müssen sich mit einer Vielzahl an Fragen auseinandersetzen:

- **Datenqualität & Interoperabilität:** Agenten sind auf konsistente, zugängliche Datenquellen angewiesen. Fehlende Schnittstellen, Datensilos oder unstrukturierte Informationen behindern deren Leistungsfähigkeit.

- **Governance & Compliance:** Insbesondere mit Blick auf die EU-AI-Verordnung und die DSGVO müssen alle Entscheidungen dokumentierbar, erklärbar und kontrollierbar sein.

Black-Box-Verhalten[3] ist nicht zulässig.

- **Integration in bestehende Prozesse:** Agenten entfalten ihren Mehrwert nur dann, wenn sie nahtlos in bestehende Workflows, Systeme und Rollenmodelle eingebunden sind.

- **Kultureller Wandel:** Die Einführung autonomer Agenten verändert auch die Rolle der Mitarbeitenden – weg von der manuellen Ausführung hin zur überwachenden Steuerung. Dies erfordert Akzeptanz, Schulung und klare Kommunikationsstrategien.

Diese Herausforderungen werden in der Abbildung „Generative AI & Agents – Herausforderungen“ als fünf Säulen dargestellt, die eine erfolgreiche Agenteninfrastruktur tragen: Daten, Infrastruktur, Expertenwissen, maschinelles Lernen und Softwareengineering – verbunden durch das Fundament einer klaren Governance.

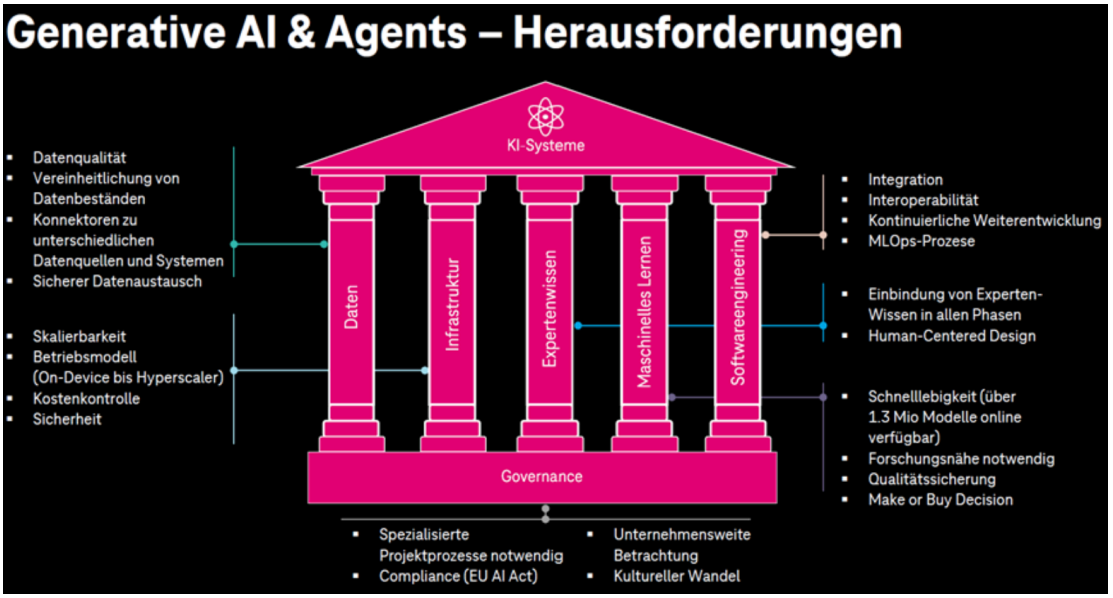

Abbildung 4
Deutsche Telekom MMS GmbH (2026): Autonome KI-Agenten: Selbstlernende KI-Systeme im Einsatz, Deutsche Telekom MMS GmbH, [online] https://www.telekom-mms.com/blog/artikel/detail/autonome-ki-agenten.

Im Zentrum steht Governance – das organisatorische Fundament, auf dem alle weiteren Elemente aufbauen. Governance umfasst alle Regeln, Prozesse und Kontrollmechanismen, die sicherstellen, dass der Einsatz von KI-Agenten nicht nur funktional, sondern auch rechtlich und ethisch vertretbar ist. Dabei geht es u. a. um spezialisierte Projektprozesse, die Einhaltung von

[3] Im Allgemeinen versteht man unter einer Black Box ein System, dessen genaue interne Funktionsweise dem Betrachter verborgen bleibt. Übertragen auf den Bereich der Künstlichen Intelligenz bezeichnet der Begriff Black Box eine KI, die für den Benutzer wie ein undurchschaubares System arbeitet. Der Benutzer sieht die Eingaben und die Ausgaben, kann aber nicht nachvollziehen oder erklären, wie und warum die Künstliche Intelligenz eine bestimmte Entscheidung getroffen oder ein bestimmtes Ergebnis ausgegeben hat.
Luber, S. (2024, May 14). Was ist eine Black Box? BigData-Insider. https://www.bigdata-insider.de/was-ist-eine-black-box-a-9d64b128e437b6a727849cf82695dbe9/

Vorschriften wie der DSGVO oder dem EU AI Act sowie um eine unternehmensweite Betrachtung, die auch kulturelle Auswirkungen berücksichtigt. Ohne diese Basis besteht die Gefahr, dass KI-Projekte unkoordiniert, risikobehaftet oder isoliert ablaufen.

4.2.4.1. Daten

Die erste Säule „Daten“ verweist auf die zentrale Voraussetzung jedes KI-Systems: den Zugriff auf qualitativ hochwertige, strukturierte und einheitlich gepflegte Daten. Ohne konsistente Datenbasis können autonome Agenten keine fundierten Entscheidungen treffen. Herausforderungen ergeben sich insbesondere dann, wenn Informationen über verschiedene Systeme verteilt sind (Datensilos), inkompatible Formate vorliegen oder der Datenzugriff aus Compliance-Gründen eingeschränkt ist. Auch der sichere Datenaustausch, z.B. über APIs, muss gewährleistet sein.

4.2.4.2. Infrastruktur

Die zweite Säule betrifft die technische Infrastruktur, auf der KI-Agenten laufen. Sie muss skalierbar, flexibel und sicher sein. Hier spielen sowohl das Betriebsmodell (lokal, in der Cloud oder hybrid) als auch Aspekte wie Kostenkontrolle und Performance eine Rolle. Besonders bei großen Sprachmodellen oder Multi-Agenten-Systemen ist die Infrastruktur schnell ein kritischer Erfolgsfaktor. Unternehmen stehen hier vor der Frage: Setzen wir auf hyperskalierende Plattformen oder auf kontrollierbare On-Premise-Lösungen?

4.2.4.3. Expertenwissen

Die dritte Säule hebt die Bedeutung von Domänenwissen und interdisziplinärer Expertise hervor. KI-Agenten müssen nicht nur programmiert, sondern auch kontextsensitiv trainiert und bewertet werden. Dazu braucht es sowohl technische Fachkräfte als auch inhaltlich versierte Anwender:innen aus den jeweiligen Fachabteilungen. Das Human-Centered Design spielt dabei eine Schlüsselrolle, denn nur wenn Menschen aktiv in den Entwicklungs- und Evaluationsprozess eingebunden sind, entstehen Systeme mit realem Nutzen.

4.2.4.4. Maschinelles Lernen

Ohne maschinelles Lernen sind autonome Agenten kaum denkbar. Diese vierte Säule beschreibt die technologischen Kernprozesse, mit denen Systeme aus Erfahrung lernen, Zusammenhänge erkennen und sich verbessern können. Das umfasst Trainingsdatensätze, Modelle, MLOps-Prozesse und kontinuierliche Anpassung. Die Herausforderung liegt hier vor allem darin, Modelle initial zu trainieren und laufend weiterzuentwickeln, im Sinne eines „lebenden Systems“, das sich an neue Anforderungen anpasst.

4.2.4.5. Software engineering

Die fünfte Säule betrifft die Softwarearchitektur, Schnittstellen und Entwicklungsprozesse. KI-Agenten müssen mit bestehenden Systemen interagieren können und sauber dokumentiert und modular aufgebaut werden. Gleichzeitig müssen Aspekte wie das Testing, Monitoring und Security-by-Design von Anfang an berücksichtigt werden. Auch strategische Entscheidungen spielen hier eine wesentliche Rolle.

4.3. Definition und Abgrenzung von KI, LLMs, autonomen Agenten

Um über die Gefahren der autonomen KIs sprechen zu können, muss man jedoch ausführlich begriffen haben, was sie sind, wie sie definiert sind und wie sie im Detail aufgebaut sind. Hierfür gehen wir auf die Definitionen der Begriffe künstliche Intelligenz, LLMs und autonome Agenten ausführlich ein.

4.3.1. Was ist künstliche Intelligenz?

Um die Funktionsweise und die potenziellen Risiken autonomer KI-Agenten angemessen untersuchen zu können, ist zunächst eine präzise begriffliche und konzeptuelle Einordnung der übergeordneten Technologie erforderlich: der künstlichen Intelligenz. Trotz der wachsenden Relevanz des Themas – sowohl in der wissenschaftlichen Forschung als auch in der industriellen und politischen Praxis – existiert bis heute keine einheitliche und allgemein anerkannte Definition von KI. Vielmehr lässt sich feststellen, dass der Begriff in verschiedenen Disziplinen, Kontexten und Anwendungsszenarien unterschiedlich ausgelegt wird. Dies liegt nicht zuletzt daran, dass auch der Begriff der „Intelligenz“ selbst bis heute als schwer greifbar gilt und nicht abschließend definiert ist.
In seiner historischen Entwicklung geht das Konzept der künstlichen Intelligenz bis in die 1940er- und 1950er-Jahre zurück. Erste theoretische Impulse stammen von Alan Turing, der 1950 in seinem berühmten Aufsatz „Computing Machinery and Intelligence“ die Frage stellte, ob Maschinen denken können. Die eigentliche Geburtsstunde der KI als wissenschaftliche Disziplin wird jedoch häufig mit dem Dartmouth Workshop im Jahr 1956 verbunden. Dort prägten Wissenschaftler wie John McCarthy, Marvin Minsky, Claude Shannon und Herbert Simon den Begriff „Artificial Intelligence“ und formulierten erstmals das Ziel, Maschinen zu entwickeln, die menschliche Denkprozesse nachahmen können. Schon damals wurde KI als ein spezifischer Zweig der Informatik verstanden, dessen Ziel es ist, autonome Systeme zu schaffen, die sich in komplexen, dynamischen Umgebungen zurechtfinden können.
Moderne Definitionen von KI knüpfen an diese Grundlagen an, erweitern sie jedoch um technologische, funktionale und ethisch-normative Perspektiven. So beschreibt die HLEG der Europäischen Kommission Künstliche Intelligenz als ein System, das die Umgebung wahrnimmt, Informationen verarbeitet, Entscheidungen trifft und Aufgaben auf eine Weise ausführt, die zuvor menschlicher

Intelligenz vorbehalten war. Charakteristisch sei dabei ein gewisses Maß an Autonomie, das es der Maschine erlaubt, flexibel auf veränderte Bedingungen zu reagieren, zu lernen und sich weiterzuentwickeln.
Ergänzend beschreibt die US-amerikanische Cybersecurity and Infrastructure Security Agency (CISA) künstliche Intelligenz als ein „System, das maschinen- und menschenbasierte Eingaben nutzt, um Vorhersagen, Empfehlungen oder Entscheidungen zu treffen, die reale oder virtuelle Umgebungen beeinflussen[4]“.
Diese funktional orientierte Definition legt besonderen Wert auf die Wirkungsdimension von KI in operativen Umgebungen und betont damit ihre Relevanz für sicherheitskritische Infrastrukturen.
Die auf Seite 2 des Papiers von Konan (2023) dargestellte Visualisierung – hier übernommen als Abbildung 5 – bietet eine hilfreiche Orientierung zur Systematik der heutigen KI-Forschung und -Anwendung. Sie zeigt, dass künstliche Intelligenz als ein übergeordnetes System zu verstehen ist, das aus verschiedenen ineinandergreifenden Komponenten besteht: insbesondere dem maschinellen Lernen sowie dessen konkreten Ausprägungen wie überwachtem Lernen (Supervised Learning), unüberwachtem Lernen (Unsupervised Learning), bestärkendem Lernen (Reinforcement Learning) und tiefem Lernen (Deep Learning). Alle diese Formen ermöglichen es Maschinen, sich anhand von Beispielen, Belohnungsstrukturen oder neuronalen Netzwerken selbstständig weiterzuentwickeln und damit immer komplexere Aufgaben zu bewältigen.

[4] CISA Principles for the Secure Integration of Artificial Intelligence in Operational Technology Publish Date: 03.12.2025
https://www.cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-operational-technology

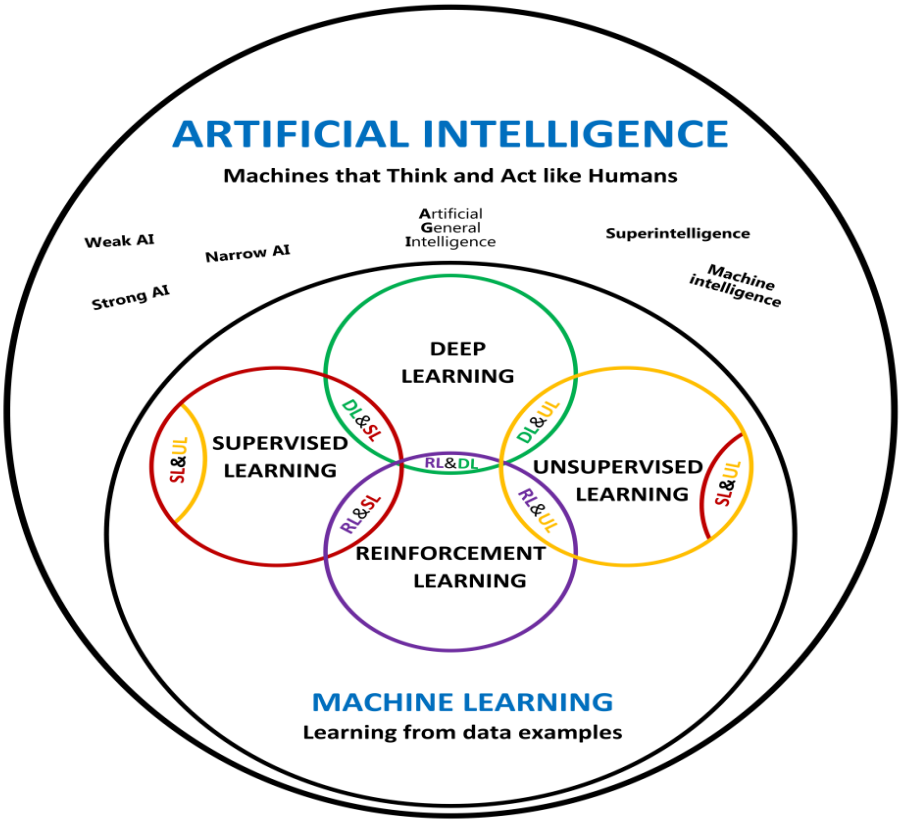

Abbildung 5
Konan Jean-Claude, Kouassi. (2022). A Comprehensive Overview of Artificial Intelligence. 173-194. 10.5121/csit.2022.122314.

Zahlreiche weitere Ansätze definieren KI anhand ihrer Funktionsweise oder ihres Grades an Autonomie. So schlägt etwa Guzman & Lewis (2022) eine konzeptuelle Rahmung vor, die KI nicht nur als technische Reaktion auf menschliche Befehle beschreibt, sondern als ein System, das reale Probleme eigenständig lösen, mit der Umwelt interagieren, Prognosen treffen und Entscheidungen fällen kann – vergleichbar mit biologischer Intelligenz. Zentrale Kriterien seien dabei der Grad der Performanz, also welche Aufgaben ein System eigenständig erledigen kann, sowie der Grad der Autonomie, also inwieweit das System unabhängig vom Menschen operiert.

Diese Sichtweise hebt hervor, dass KI-Systeme sich nicht allein durch ihre technischen Komponenten definieren, sondern auch durch ihr Verhalten, ihre Zielgerichtetheit und ihre Fähigkeit zur kontinuierlichen Anpassung. Sie sind keine einfachen Werkzeuge, sondern können – je nach Ausprägung – eigenständig lernen, sich weiterentwickeln und sogar eigene Handlungsketten planen. Hier liegt das verbindende Element zur Diskussion über **autonome Agenten**: Denn erst wenn eine KI in der Lage ist, sich auf wechselnde Umgebungsbedingungen einzustellen und selbstständig Mittel und Wege zur Zielerreichung auszuwählen, sprechen wir nicht mehr von klassischer Automatisierung, sondern von echter Intelligenz im digitalen Raum.

4.3.2. Was sind Large Language Models (LLMs)?

Ein zentrales Teilgebiet der modernen KI-Forschung und zugleich die technologische Grundlage zahlreicher aktueller Anwendungen ist das Konzept der LLMs.
Darunter versteht man tief neuronale Sprachmodelle, die auf der Architektur sogenannter Transformer-Netzwerke basieren und durch das Training auf massiven Textdatensätzen in die Lage versetzt werden, menschliche Sprache in Kontext, Bedeutung und Struktur zu erfassen, zu analysieren und selbstständig zu generieren. LLMs wie GPT-4, PaLM, Claude oder LLaMA stellen dabei nicht nur einen Meilenstein in der Sprachverarbeitung dar, sondern gelten auch als Schlüsseltechnologien für die Entwicklung autonomer KI-Agenten.

Während klassische NLP-Systeme auf regelbasierten oder statistikbasierten Verfahren beruhten, markieren LLMs einen Paradigmenwechsel: Sie sind in der Lage, nicht nur auf Basis einzelner Begriffe, sondern auch über **komplexe Kontexte hinweg** Bedeutungszusammenhänge herzustellen und kohärente Texte, Zusammenfassungen, Argumentationsketten oder sogar Programmcode zu generieren. Dies macht sie zu besonders leistungsfähigen Werkzeugen – nicht nur in der Informationsverarbeitung, sondern auch in der Wissensgenerierung, Kommunikation und Entscheidungsunterstützung bis hin zum Fällen von Entscheidungen in der Zukunft.

Das Training eines LLMs erfolgt typischerweise in zwei zentralen Phasen: einem **Pretraining**, bei dem das Modell anhand großer, unspezifischer Textmengen grundlegende Sprachmuster erlernt, sowie einem anschließenden **Fine-Tuning**, in dem es auf spezifische Aufgaben oder Interaktionen angepasst wird.

In der Praxis folgt hierauf oft eine Phase des sogenannten **Reinforcement Learning from Human Feedback (RLHF)**, bei der das Modell Rückmeldungen von Menschen erhält und dadurch qualitativ gesteuert weiterlernt.

4.3.3. Beispiele: AutoGPT, AgentGPT, BabyAGI und WormGPT

Um die zuvor erläuterten Begriffe künstliche Intelligenz, LLMs und autonome Agenten ausführlicher zu vernschaulichen, werden im Folgenden Beispiele der LLMs betrachtet: AutoGPT, AgentGPT und WormGPT.
Diese Systeme stehen exemplarisch für unterschiedliche Ausprägungen von LLM-basierten Agenten, von produktiven Automatisierungswerkzeugen bis hin zu explizit missbräuchlichen „Dark AI“-Modellen im Cybercrime-Kontext.

AutoGPT als früher Open-Source-Agent

AutoGPT gilt als eines der ersten weithin bekannten Agenten-Frameworks, das die Fähigkeiten von GPT-4 in eine

weitgehend autonome Steuerungslogik überführt. Technisch handelt es sich um eine Open-Source-Anwendung, die ein vom Nutzer formuliertes Ziel entgegennimmt und dieses in eine Folge von Teilaufgaben zerlegt, die selbstständig geplant und abgearbeitet werden. Hierzu kann AutoGPT unter anderem Suchmaschinen aufrufen, Webseiten analysieren, Dateien schreiben, Code generieren und externe Tools ansteuern, ohne dass der Nutzer jeden Zwischenschritt explizit anstoßen muss.

Der typische Ablauf folgt dabei dem in Abschnitt 4.2 skizzierten Agenten-Loop: Nach dem Zielverstehen plant AutoGPT eigenständig eine Aufgabenliste, führt einzelne Schritte aus, bewertet die Zwischenergebnisse und entscheidet, welche weiteren Aktionen notwendig sind, um das Gesamtziel zu erreichen. Dieses Verhalten wird unter anderem in technischen Blogs und Entwicklerdokumentationen hervorgehoben, in denen AutoGPT als „Autopilot“ beschrieben wird, der aus einem einfachen Prompt („Führe eine Wettbewerbsanalyse für Markt X durch“) eine Kette von Recherchen, Auswertungen und Berichtsgenerierungen ableitet.

Einsatzszenarien von AutoGPT reichen von automatisierten Marktrecherchen über die Analyse von Finanz- oder Log-Daten bis hin zur (teil)autonomen Softwareentwicklung, bei der der Agent Code schreibt, testet und iterativ verbessert. IBM beschreibt AutoGPT in diesem Zusammenhang als Beispiel für eine neue Klasse von Systemen, die nicht nur natürliche Sprache verarbeiten, sondern eigenständig Aktionen in digitalen Umgebungen ausführen und damit klassische KI-Anwendungen um eine operative Handlungsebene erweitern. Zugleich wird bereits in frühen Analysen darauf hingewiesen, dass fehlende Zugriffsbeschränkungen – etwa auf Dateisysteme oder Produktivsysteme – zu erheblichen Sicherheitsrisiken führen können, wenn der Agent unkontrolliert handelt.

AgentGPT als niedrigschwellige Web-Plattform

AgentGPT verfolgt ein ähnliches Grundprinzip, senkt aber die Einstiegshürden für Endnutzer deutlich, indem es als webbasierte Plattform bereitgestellt wird. Über eine Browseroberfläche können Nutzer Agenten benennen, Ziele formulieren und Parameter wie gewünschte Aufgabentypen oder Prioritäten festlegen; die eigentliche Planung und Ausführung der Tasks erfolgt dann autonom im Hintergrund. Technisch nutzt AgentGPT, vergleichbar mit AutoGPT, LLMs wie GPT-3.5 oder GPT-4 zur Generierung von Zwischenschritten und zur Bewertung der Zwischenergebnisse, bietet jedoch eine stärker geführte Oberfläche und integrierte Monitoring-Funktionen.

In praktischen Anwendungsbeispielen wird AgentGPT häufig für Marketing- und Vertriebsaufgaben eingesetzt, etwa zur Erstellung und Auswertung von Kampagnen, zur Generierung von Social-Media-Inhalten oder zur automatisierten Recherche von Zielkunden. TechTarget beschreibt AgentGPT als „no-code“-Einstieg in die Welt der autonomen Agenten, der es auch nicht-technischen Nutzergruppen ermöglicht, komplexe Aufgabenpipelines zu definieren.

Aus sicherheitstechnischer Sicht entsteht damit eine neue Form der Abhängigkeit: Unternehmensnutzer binden sensible Daten (z. B. Kundendaten, interne Dokumente) in cloudbasierte Agenten-Workflows ein, deren interne Funktionsweise häufig nur begrenzt nachvollziehbar ist.

Vergleichende Analysen zwischen AutoGPT und AgentGPT betonen, dass beide Systeme auf demselben technologischen Fundament (LLMs + Tool-Integration + Agenten-Loop) aufbauen, sich aber hinsichtlich Kontrollgrad und Integrationsmodell unterscheiden. Während AutoGPT als lokal installierbares Framework eine hohe Flexibilität und tiefgreifende Integration in bestehende IT-Landschaften ermöglicht, positioniert sich AgentGPT als Service-Plattform mit schnell konfigurierbaren, aber stärker vom Anbieter gesteuerten Agenten. Für Unternehmen hat dies unmittelbare Implikationen für Themen wie Datenhoheit, Compliance und Angriffsoberfläche, was in den Kapiteln 7 und 8 noch vertieft wird.

BabyAGI als prototypischer Task-Management-Agent

Ein drittes, häufig zitiertes Beispiel ist BabyAGI, ein 2023 veröffentlichtes Open-Source-Projekt, das als „Task-Management-System“ auf Basis von GPT-4 und Vektordatenbanken konzipiert ist. BabyAGI erhält ein übergeordnetes Ziel und erzeugt daraus eigenständig eine dynamische Liste von Aufgaben, die priorisiert, ausgeführt und basierend auf den Ergebnissen fortlaufend angepasst werden. Die Resultate jeder Aufgabe werden in einer Wissensbasis abgelegt und dienen wiederum als Kontext für nachfolgende Schritte, wodurch eine Art selbstverstärkender Lern- und Anpassungsprozess entsteht.

In der Praxis wird BabyAGI vor allem als experimentelle Plattform betrachtet, um zu erforschen, wie LLM-basierte Systeme eigenständig komplexe Problemräume erschließen können, etwa bei der Forschungsliteratur-Auswertung, der Produktidee-Generierung oder der iterativen Verbesserung von Konzepten. IBM charakterisiert BabyAGI in diesem Zusammenhang als frühen Prototyp für Systeme, die langfristige Ziele verfolgen und selbstständig neue Unterziele formulieren – eine Eigenschaft, die mit traditionellen Automatisierungsansätzen kaum zu realisieren war. Gleichzeitig warnen mehrere Fachbeiträge, dass solche Agenten bei unzureichender Zieldefinition („vage Prompts“) oder fehlender Begrenzung ihrer Handlungsmöglichkeiten unvorhersehbare oder unerwünschte Aktionen ausführen können, insbesondere wenn sie mit externen APIs, Dateisystemen oder Produktionssystemen verknüpft sind.

WormGPT als Beispiel eines bösartigen LLM-Modells

Während AutoGPT, AgentGPT und BabyAGI vorwiegend als legitime oder experimentelle Werkzeuge adressiert werden, markiert WormGPT einen Wendepunkt hin zu explizit böswillig ausgerichteten LLM-Modellen. WormGPT wurde erstmals 2023 in einschlägigen Untergrundforen beworben und nach Analysen verschiedener Sicherheitsanbieter auf der Basis des Open-Source-Modells GPT-J

entwickelt. Im Gegensatz zu kommerziellen Diensten wie ChatGPT wurden bei WormGPT sämtliche sicherheitsrelevanten Schutzmechanismen („Guardrails“) entfernt; zusätzlich soll das Modell mit Datensätzen trainiert worden sein, die Malware-Code, Exploit-Beschreibungen, Phishing-Texte und andere illegale Inhalte umfassen.

Sicherheitsanalysen von Unit 42, Eviden und anderen zeigen, dass WormGPT insbesondere im Bereich von Business-Email-Compromise(BEC)- und Phishing-Kampagnen eingesetzt wird. Die generierten E-Mails sind sprachlich weitgehend fehlerfrei, kontextsensitiv und im Stil realer Geschäfts- oder Supportkommunikation formuliert, wodurch klassische Warnsignale etwa untypische Grammatik, Rechtschreibfehler oder stilistische Brüche deutlich reduziert werden. Dadurch steigt die Erfolgswahrscheinlichkeit von Social-Engineering-Angriffen, insbesondere bei Mitarbeitenden ohne spezifische Awareness-Schulung.

Darüber hinaus wird WormGPT in Untergrundangeboten damit beworben, in verschiedenen Programmiersprachen funktionsfähigen Schadcode erzeugen zu können, etwa für Ransomware-Varianten, Keylogger oder Informationsdiebstahl-Tools. Berichte von Sicherheitsfirmen und Fachblogs weisen darauf hin, dass selbst technisch wenig versierte Angreifer durch die Nutzung solcher Modelle komplexere Angriffe durchführen können, die bislang spezialisiertes Know-how erforderten. Parallel dazu sind inzwischen eine Reihe weiterer „Dark AI“-Modelle wie FraudGPT oder Weiterentwicklungen von WormGPT beobachtbar, die als Dienstleistung („Malicious LLM-as-a-Service“) über Darknet-Marktplätze und Telegram-Kanäle angeboten werden und Funktionen wie Malware-Baukästen, Phishing-Kampagnen-Generatoren oder Exploit-Assistenten bündeln.

Eftsure beschreibt dieses Phänomen als „Kommodifizierung von KI-gestützten Angriffen“: Statt eigenständig Phishing-Texte, Schadcode oder Social-Engineering-Skripte zu erstellen, können Angreifer vorgefertigte Prompts und Workflows nutzen, die von den Betreibern der Modelle bereitgestellt werden. Dadurch sinken die Einstiegshürden für Cyberkriminalität weiter, während zugleich Menge und Qualität der Angriffe zunehmen. Dieser Trend wird in aktuellen Übersichtsarbeiten zu „malicious LLMs“ als wesentlicher Faktor für die erwartete Zunahme autonomer bzw. teilautonomer Angriffsketten bewertet.

Zwischenfazit: Spannungsfeld zwischen Produktivität und Missbrauch

Die Beispiele AutoGPT, AgentGPT, BabyAGI und WormGPT verdeutlichen, dass auf LLMs aufsetzende Agentensysteme ein breites Spektrum an Einsatzmöglichkeiten abdecken, von legitimen Automatisierungslösungen bis hin zu hochgradig missbräuchlichen Werkzeugen im Cybercrime-Ökosystem. Aus funktionaler Sicht teilen sie zentrale Eigenschaften: Sie können Ziele aus natürlichsprachigen Vorgaben ableiten, eigenständig Handlungspläne erstellen, externe

Tools oder Datenquellen einbinden und ihre Aktionen iterativ an Zwischenergebnissen ausrichten. Genau diese Fähigkeiten bilden jedoch nicht nur die Grundlage für Effizienzgewinne in Unternehmen, sondern erhöhen zugleich das Bedrohungspotenzial, wenn vergleichbare Agenten aus Angreiferperspektive eingesetzt werden – etwa zur automatisierten Informationsgewinnung, zur Generierung überzeugender Phishing-Kampagnen oder zur schnellen Entwicklung und Anpassung von Schadsoftware.

Für die vorliegende Arbeit sind diese Beispiele daher in zweierlei Hinsicht relevant: Erstens zeigen AutoGPT, AgentGPT und BabyAGI, welche Art von Autonomie und Entscheidungslogik künftig in Unternehmenskontexten anzutreffen sein wird; dies beeinflusst sowohl interne Prozesse als auch die Angriffsfläche nach außen. Zweitens illustriert WormGPT, wie dieselben technologischen Prinzipien – LLM-basierte Generierung, Tool-Integration und Agenten-Loop gezielt in bösartige Werkzeuge überführt werden können, die Angriffe skalieren und professionalisieren. Die daraus resultierenden Implikationen für Angriffsszenarien (Kapitel 6) und Unternehmensrisiken (Kapitel 7) bilden den Kern der weiteren Analyse.

4.3.4. Fähigkeit zur eigenständigen Zielverfolgung und Entscheidungsfindung

Ein zentrales Abgrenzungsmerkmal autonomer KI-Agenten gegenüber klassischen KI-Systemen liegt in ihrer Fähigkeit, Ziele eigenständig zu verfolgen und im Rahmen dieser Zielverfolgung Entscheidungen zu treffen, ohne dass jede einzelne Aktion durch einen Menschen vorgegeben werden muss. Während traditionelle Systeme in der Regel auf klar definierte Eingaben reagieren und deterministische Entscheidungsbäume oder feste Workflows abarbeiten, verfügen Agentensysteme auf Basis von LLMs über Mechanismen zur Zielinterpretation, Planung, Ausführung und laufenden Anpassung ihres Vorgehens.

Technisch basiert diese Eigenständigkeit in vielen aktuellen Implementierungen auf einem iterativen Agenten-Loop, der – wie in Abschnitt 4.2 beschrieben – aus mehreren Schritten besteht: Zunächst interpretiert der Agent eine natürlichsprachige Zielvorgabe, leitet daraus eine Reihe von Unterzielen bzw. Aufgaben ab, priorisiert diese und führt konkrete Aktionen aus (z. B. Web-Recherchen, API-Aufrufe, Code-Ausführung). Die Ergebnisse dieser Aktionen werden anschließend bewertet und in einer Form von Kurz- oder Langzeitspeicher abgelegt; auf dieser Basis passt der Agent seinen Plan an, ergänzt neue Aufgaben oder verwirft nicht mehr relevante Schritte. IBM beschreibt dieses Vorgehen am Beispiel von BabyAGI als wiederkehrenden Zyklus aus Aufgabenerzeugung, Aufgabenpriorisierung und Aufgabenausführung, bei dem ein LLM als „Orchestrator“ fungiert und sowohl Zielinterpretation als auch Entscheidungslogik steuert.

Die in Abschnitt 4.3.2 dargestellten Systeme AutoGPT und BabyAGI verdeutlichen, wie diese Fähigkeit zur eigenständigen Zielverfolgung praktisch umgesetzt wird. AutoGPT nimmt ein übergeordnetes Ziel

entgegen und zerlegt es selbstständig in eine Sequenz von Teilaufgaben, die iterativ abgearbeitet, bewertet und bei Bedarf neu geplant werden. BabyAGI geht einen Schritt weiter, indem es eine persistente Aufgabenliste und eine Vektordatenbank als Gedächtnis nutzt, um den Fortschritt über mehrere Iterationen hinweg nachzuvollziehen und neue Aufgaben aus bereits erzielten Zwischenergebnissen abzuleiten. In der Literatur wird in diesem Zusammenhang darauf hingewiesen, dass BabyAGI-ähnliche Architekturen gezielt auf Langfristigkeit und Selbststeuerung ausgelegt sind: Der Agent hält einen Wissens- und Aufgabenstand über längere Zeiträume vor, bewertet den eigenen Fortschritt und generiert selbstständig neue Aufgaben, um ein übergeordnetes Ziel weiter zu verfolgen.

Diese Fähigkeit zur eigenständigen Zielverfolgung ist eng mit bestimmten funktionalen Bausteinen verknüpft, die in neueren Übersichtsarbeiten zu LLM-basierten Agenten immer wieder hervorgehoben werden. Dazu zählen insbesondere:

- Perzeption der Umgebung (z. B. über Spracheingaben, Dokumente, Logs oder API-Antworten),
- Speicher in Form von Kurzzeit- und Langzeitgedächtnis (z. B. Vektordatenbanken, Protokolle),
- Schlussfolgern und Planen auf Basis des LLM, und
- Handlungsfähigkeit durch Anbindung externer Tools und Systeme.

Auf dieser Grundlage können Agenten nicht nur auf direkte Benutzeranfragen reagieren, sondern auch eigenständige Strategien entwickeln, alternative Handlungspfade abwägen und Entscheidungen unter Unsicherheit treffen – etwa welche Informationsquelle als nächstes abgefragt oder welche Schwachstelle in einem System priorisiert untersucht werden soll.

Mit zunehmender Autonomie verändert sich damit natürlich die Qualität der Entscheidungsfindung. AWS unterscheidet in einem Sicherheitsrahmenwerk für „agentic AI“ verschiedene Stufen von Handlungsfähigkeit: von Systemen mit streng vorgeschriebener Agency, die vorgegebene Schritte abarbeiten, über überwacht-autonome Agenten bis hin zu Systemen mit „voller Agency“, die in definierten Grenzen eigenständig Ziele setzen, Pläne anpassen und Aktionen durchführen. In den höheren Stufen können Agenten Aktivitäten auf Basis von Umgebungssignalen oder Zeitplänen eigeninitiieren, über längere Zeiträume hinweg ohne unmittelbare menschliche Aufsicht operieren und dabei dynamisch mit mehreren externen Systemen interagieren. Damit verschiebt sich der Schwerpunkt von punktuellen, durch den Menschen ausgelösten Interaktionen hin zu kontinuierlichen, selbstgesteuerten Prozessen.

Für den Kontext der Cybersecurity ist diese Entwicklung ambivalent. Auf der einen Seite wird in aktuellen Surveys betont, dass LLM-basierte Agenten als Verteidigungssysteme eingesetzt werden

können, etwa zur automatisierten Erkennung von Anomalien, zur Korrelation großer Datenmengen oder zur dynamischen Anpassung von Abwehrmaßnahmen. Auf der anderen Seite zeigen neuere Arbeiten zu „autonomous cyberattack agents“, dass dieselben Fähigkeiten genutzt werden können, um eigenständig Reconnaissance-Aktivitäten durchzuführen, Schwachstellen zu identifizieren, Exploits zu generieren und Angriffspfade in verschiedenen Netzwerkinfrastrukturen zu optimieren. Die Autoren sprechen in diesem Zusammenhang von einer „Cyber Threat Inflation“: Durch die Reduktion von Zeit- und Wissensbarrieren können Angriffe schneller, kostengünstiger und in größerem Maßstab durchgeführt werden.

Damit wird deutlich, dass die Fähigkeit zur eigenständigen Zielverfolgung und Entscheidungsfindung nicht nur eine technologische Eigenschaft, sondern auch ein sicherheitsstrategischer Faktor ist. Je höher der Autonomiegrad eines Agenten, desto größer ist einerseits sein Nutzen für Unternehmen – etwa durch automatisierte Prozessketten und entlastete Mitarbeitende –, desto anspruchsvoller werden andererseits Anforderungen an Governance, Kontrolle und Überwachung. Für die weitere Analyse in dieser Arbeit bedeutet dies, dass autonome KI-Agenten nicht nur als „intelligente Werkzeuge“ verstanden werden können, sondern als eigenständige Akteure im digitalen Raum, deren Zielverfolgung – je nach Konfiguration und Missbrauchspotenzial – sowohl schützende als auch angreifende Wirkung entfalten kann.

4.4. Cyberkriminalität – Begriffe und Entwicklungen

Die Betrachtung autonomer KI-Agenten als potenzielle Angreifer setzt voraus, dass zunächst klar ist, in welchem Bedrohungsumfeld sie sich bewegen. Bevor in den späteren Kapiteln konkrete Szenarien entwickelt werden, soll dieses Kapitel daher zentrale Begriffe der Cyberkriminalität klären, typische Angriffsformen skizzieren und grundlegende Entwicklungen der letzten Jahre nachzeichnen. Es ordnet den Begriff Cyberkriminalität ein, stellt klassische und weiterentwickelte Angriffsformen vor und skizziert zentrale Entwicklungen wie Automatisierung und Cybercrime-as-a-Service.

4.4.1. Klassische und moderne Formen (Phishing, Reconnaissance, Malware etc.)

Der Begriff „Cyberkriminalität“ ist in der Fachliteratur und bei Behörden nicht völlig einheitlich definiert, in den Kernpunkten aber gut abgegrenzt. Das Bundesamt für Sicherheit in der Informationstechnik (BSI) versteht unter Angriffen im Cyberraum zielgerichtete Handlungen, die auf die Verfügbarkeit, Vertraulichkeit oder Integrität von IT-Systemen und Daten zielen und dabei in der Regel eine strafbare Handlung darstellen. Das Bundeskriminalamt (BKA) unterscheidet in seinen Lagebildern traditionell zwischen „Cyberkriminalität im engeren Sinne“, also Straftaten, die sich direkt gegen das Internet, Netze oder Daten richten (z. B. Hacking, DDoS-Angriffe), und „Cyberkriminalität im weiteren Sinne“, bei der das Internet lediglich als Tatmittel genutzt wird, etwa für Betrug oder Erpressung. ENISA und Europol schlagen in ihren Berichten in eine

ähnliche Richtung und betonen, dass es im Kern um die Ausnutzung digitaler Infrastrukturen zur Begehung klassischer und neuer Straftaten geht.

Zu den „klassischen“ Angriffsformen zählen nach wie vor Phishing, Malware, Ransomware, DDoS-Angriffe sowie unterschiedliche Varianten von Social Engineering. Phishing beschreibt den Versuch, Nutzerinnen und Nutzer mithilfe gefälschter E-Mails, Webseiten oder Nachrichten dazu zu bringen, vertrauliche Informationen wie Zugangsdaten preiszugeben oder auf schädliche Links zu klicken. Häufig fungiert Phishing als Einstiegspunkt in eine Angriffskette, beispielsweise als erster Schritt vor der Installation von Ransomware.

Unter Malware werden verschiedene Arten von Schadprogrammen zusammengefasst, etwa Trojaner, Viren oder Spyware. Sie dienen vereinfacht gesagt dazu, Systeme zu kompromittieren, Daten auszuspähen, zu manipulieren oder anderweitig missbräuchlich zu nutzen. Ransomware ist eine spezielle Form von Malware, die Daten verschlüsselt und die Opfer anschließend mit einer Lösegeldforderung erpresst; das BKA und das BSI sehen Ransomware seit Jahren als eine der wirtschaftlich bedeutsamsten Bedrohungen, insbesondere für Unternehmen und kritische Infrastrukturen. DDoS-Angriffe wiederum zielen auf die Verfügbarkeit ab: Durch massenhafte Anfragen werden Server oder Netze überlastet, sodass legitime Nutzer keine Dienste mehr erreichen können.
Social Engineering ist als übergeordneter Begriff zu verstehen. Hierbei manipulieren Angreifer gezielt das Verhalten von Menschen, um an Informationen, Zugänge oder andere Vorteile zu gelangen – sei es telefonisch, per E-Mail oder vor Ort. Phishing lässt sich als spezielle, meist massenhaft ausgespielte Form des Social Engineering betrachten, bei der Kommunikation standardisiert und breit gestreut wird, während etwa BEC-Angriffe eher auf wenige, sorgfältig ausgewählte Zielpersonen setzen.

Neben diesen äußeren Angriffen gehört aus technischer Sicht auch Reconnaissance – also die systematische Aufklärung von Zielsystemen – zu den grundlegenden Bausteinen nahezu jeder Angriffskette. Angreifer sammeln Informationen über Infrastruktur, Domains, genutzte Softwareversionen oder öffentlich zugängliche Services, um daraus mögliche Schwachstellen abzuleiten. Dieser Schritt findet oft im Verborgenen statt, etwa über automatisierte Scans oder OSINT-Werkzeuge, taucht aber in vielen technischen Frameworks wie MITRE ATT&CK[5] als eigene Phase auf und ist für das Verständnis moderner Angriffe zentral.

In den letzten Jahren haben sich diese klassischen Formen deutlich weiterentwickelt. Phishing-Angriffe sind gezielter geworden („Spear-Phishing“), also auf konkrete Personen oder Rollen zugeschnitten. Statt generischer Massenmails, die leicht zu erkennen

[5] MITRE ATT&CK ist eine weltweit zugängliche Wissensdatenbank zu Taktiken und Techniken von Angreifern, die auf Beobachtungen aus der Praxis basiert. Die ATT&CK-Wissensdatenbank dient als Grundlage für die Entwicklung spezifischer Bedrohungsmodelle und Methoden im privaten Sektor, in der Regierung und in der Cybersicherheitsbranche.
MITRE ATT&CK® (o. D.): [online] https://attack.mitre.org/.

sind, verwenden Angreifer persönliche Anreden, beziehen sich auf reale Projekte oder Vorgesetzte und nutzen Informationen aus sozialen Netzwerken. ENISA und Europol weisen darauf hin, dass sich Spear-Phishing zur bevorzugten Methode entwickelt hat, um hochrangige Unternehmensentscheidende oder Administratoren anzugreifen, da bereits ein einziger erfolgreicher Zugriff gravierende Folgen haben kann.

Parallel dazu wird Social Engineering zunehmend durch KI unterstützt, etwa durch generative Modelle, die sprachlich hochwertige, fehlerfreie und inhaltlich überzeugende Texte erzeugen. Europol und das BSI sehen darin eine neue Qualität: Phishing- und Betrugsnachrichten lassen sich massenhaft in verschiedenen Sprachen und Varianten erzeugen, was klassische Erkennungsmerkmale wie schlechte Grammatik oder unpassende Formulierungen entwertet. Dieser Trend knüpft direkt an die in Kapitel 4.3 behandelten LLMs an und bildet eine wichtige Grundlage für die später diskutierten autonomen Agenten.

Auch auf technischer Ebene haben sich Angriffe weiterentwickelt. Polymorphe und metamorphe Malware verändert ihre Signaturen oder sogar Teile ihres Codes automatisch, um signaturbasierte Erkennungssysteme zu umgehen. Sicherheitsberichte zeigen, dass moderne Schadsoftware zunehmend auf Verschleierungstechniken, modulare Architekturen und die gezielte Ausnutzung von Zero-Day-Schwachstellen setzt. Hinzu kommt der Trend zu Supply-Chain-Angriffen (z.B. SolarWinds 2020[6]), bei denen nicht das eigentliche Zielunternehmen direkt angegriffen wird, sondern ein vorgelagerter Dienstleister, Softwareanbieter oder ein anderes Glied der Lieferkette. Gelingt es Angreifern, in Build-Prozesse, Update-Infrastruktur oder Support-Kanäle einzudringen, können sie anschließend eine große Zahl von Kunden gleichzeitig kompromittieren.

Für Unternehmen ist diese Entwicklung aus mehreren Gründen relevant. Zum einen steigt die Komplexität der Bedrohungslage: Es reicht nicht mehr, nur den eigenen Perimeter zu schützen, wenn Angriffe über Lieferketten, Dienstleister oder Mitarbeitende in der Cloud erfolgen. Zum anderen verschwimmen die Grenzen zwischen klassischer IT-Kriminalität und betrügerischen Geschäftsmodellen, etwa bei Ransomware-Gruppen, die äußerst professionell agieren und Strukturen aufweisen, die eher an Unternehmen als an „Hacker im Keller“ erinnern. Diese Professionalisierung zieht sich durch alle nachfolgenden Entwicklungen – von der Automatisierung bis hin zu CaaS.

4.4.2. Zunahme der Angriffsautomatisierung

Ein zweiter Trend, der im Kontext autonomer KI-Agenten besonders relevant ist, ist die zunehmende Automatisierung von Angriffen. Schon bevor LLMs und moderne Agentensysteme breiter verfügbar wurden,

[6] SolarWinds supply chain attack | Fortinet o. D.

setzten Angreifer auf automatisierte Scans, Exploit-Kits oder Botnetze, um große Mengen potenziell verwundbarer Ziele abzugreifen.
Der BSI-Lagebericht 2024 beschreibt die Bedrohungslage in Deutschland insgesamt als „angespannt“ und „besorgniserregend“ und verweist unter anderem auf einen massiven Anstieg neuer Schadprogramme. Demnach wurden im Berichtszeitraum laut Auswertung rund 300.000 neue Schadprogrammvarianten pro Tag registriert, was einem Anstieg von etwa 26 Prozent gegenüber dem Vorjahr entspricht (vgl. Die Lage der IT-Sicherheit in Deutschland 2024 o. D.).

Diese Zahlen sind ohne eine weitgehende Automatisierung in der Entwicklung, Anpassung und Verteilung von Malware kaum erklärbar. Schadsoftware wird heute vielfach in Generatoren erzeugt, die automatisiert Variationen erzeugen, um Erkennungssysteme zu unterlaufen.
Auch bei DDoS-Angriffen zeigt sich ein klarer Automatisierungseffekt. Das BSI berichtet für 2024 von einem deutlichen Zuwachs an Angriffen mit sehr hohen Bandbreiten; im ersten Halbjahr 2024 wurden vermehrt DDoS-Attacken mit einem Volumen von mehr als 10.000 Mbit/s beobachtet. Solche Angriffe werden typischerweise über Botnetze aus kompromittierten IoT-Geräten oder Servern ausgeführt, die zentral gesteuert werden und sich automatisiert gegen verschiedene Ziele richten lassen.

Der Verizon Data Breach Investigations Report (DBIR) 2024 zeigt zudem, dass Social-Engineering-Angriffe und Ransomware eng miteinander verknüpft sind: Phishing fungiert hier häufig als automatisiert verteilte Einstiegsmethode, über die Zugangsdaten oder erste Zugänge in Unternehmensnetze erlangt werden. Gleichzeitig betont der Bericht, dass Ransomware-Gruppen „eine bemerkenswerte Fähigkeit zur Anpassung ihrer Taktiken“ entwickelt haben und vermehrt automatisiert nach Schwachstellen suchen, diese ausnutzen und anschließend standardisierte Erpressungsabläufe durchlaufen.

Europol wiederum hebt in der IOCTA 2024 hervor, dass insbesondere Ransomware-as-a-Service (RaaS) und andere Dienstleistungsmodelle eine industrialisierte Form der Angriffsabwicklung ermöglichen. Affiliates können vorgefertigte Angriffswerkzeuge nutzen, die weite Teile des Angriffsprozesses vom initialen Zugang über die Verteilung der Malware bis hin zur Verschlüsselung und Zahlungsabwicklung in Kryptowährungen automatisieren.

Automatisierung verändert damit mehrere Dimensionen von Angriffen:

1. **Skalierbarkeit wird gestiegen**
   Einmal entwickelte Tools, Skripte oder Kampagnen lassen sich mit geringem Aufwand auf hunderte oder tausende Ziele anwenden. Ein Beispiel sind sogenannte Credential-Stuffing-Angriffe, bei denen automatisiert große Mengen gestohlener Zugangsdaten gegen Login-Portale getestet werden.

2. **Erhöhte Geschwindigkeit**
   Schwachstellen können bereits kurz nach ihrer Veröffentlichung oder sogar als Zero-Day in großem Stil ausgenutzt werden, weil

automatisierte Scans die Suche nach verwundbaren Systemen übernehmen. Für Unternehmen bedeutet das, dass die Zeitspanne zwischen Bekanntwerden einer Schwachstelle und dem ersten flächendeckenden Angriff deutlich geschrumpft ist.

3. **Professionalität und Standardisierung der Angriffe werden gestiegen**
   Viele moderne Kampagnen folgen klaren Playbooks, die sich an etablierten Angriffsmodellen orientieren und automatisiert durchlaufen werden können. Das betrifft sowohl technische Schritte als auch organisatorische Elemente wie Verhandlungsvorlagen für Erpressung oder automatisierte Portale für Opferkommunikation, wie sie bei größeren Ransomware-Gruppen zu beobachten sind.

   Die Brücke zur KI-gestützten Automatisierung ergibt sich hier fast zwangsläufig. LLM-basierte Systeme können eingesetzt werden, um Teile der Angriffsplanung, Reconnaissance oder Social-Engineering-Phase zu unterstützen oder zu übernehmen. Erste Berichte von Sicherheitsbehörden und Unternehmen deuten darauf hin, dass KI beispielsweise genutzt wird, um große Mengen kompromittierter Daten zu analysieren, Zielpersonen zu priorisieren oder Phishing-Texte dynamisch anzupassen. An dieser Stelle bleibt die Betrachtung bewusst auf der konzeptionellen Ebene; die konkrete Rolle autonomer KI-Agenten in automatisierten Angriffsketten wird in Kapitel 5 detailliert aufgegriffen.

   Automatisierung ist längst ein integraler Bestandteil moderner Cyberangriffe und wirkt als „Hebel“ wirkt, der die Wirkung einzelner Angreifer vervielfacht. Autonome KI-Agenten treffen somit nicht auf eine statische, sondern auf eine bereits hochgradig industrialisierte Angreiferlandschaft.

4.4.3. Darknet & CaaS

Neben technischen Entwicklungen spielt auch die Infrastruktur eine Rolle, über die Angreifer Ressourcen beziehen, sich vernetzen und ihre Dienste anbieten. Hier ist das Darknet als Teil des Internets von zentraler Bedeutung.

Unter dem Darknet werden in der Regel Netzwerkbereiche verstanden, die nur über spezielle Software oder Protokolle erreichbar sind, etwa über das Tor-Netzwerk oder I2P. Technisch handelt es sich meist um Overlay-Netzwerke, bei denen Kommunikation stark verschlüsselt und verschachtelt wird, sodass Herkunft und Ziel einer Verbindung kaum nachvollziehbar sind. Diese Eigenschaften machen das Darknet zwar nicht per se kriminell – es dient auch Journalisten, Aktivisten und Bürgern in autoritären Staaten als Schutzraum. In der Praxis ist es jedoch ein bevorzugter Ort für Marktplätze und Foren, auf denen illegale Güter und Dienste gehandelt werden.

Europol beschreibt in der IOCTA 2024, dass sich die Landschaft solcher Plattformen in den letzten Jahren stark verändert hat: Große, zentrale Marktplätze wurden wiederholt durch

Strafverfolgungsbehörden zerschlagen, was zu einer Fragmentierung und häufigem „Rebranding“ der Angebote geführt hat. Trotz dieser Instabilität bleibt das Darknet ein wichtiger Knotenpunkt, etwa für den Handel mit Zugangsdaten, gehackten Datenbanken, Malware, Exploits und Dienstleistungsangeboten im Bereich Cybercrime.

Letztere werden häufig unter dem Sammelbegriff Cybercrime-as-a-Service (CaaS) zusammengefasst. Gemeint sind Angebotsformen, bei denen einzelne Bausteine einer Angriffskette von der Infrastruktur über Tools bis hin zu kompletten Kampagnen als Dienstleistung angeboten und von anderen Kriminellen gebucht werden können. Dieses Modell lehnt sich in seiner Struktur stark an legitime „as-a-Service“-Geschäftsmodelle aus der IT-Branche an.

Ein prominentes Beispiel ist Ransomware-as-a-Service. Hier stellen Betreibergruppen fertige Ransomware-Pakete, Payment-Infrastruktur, Kommunikationsportale und zum Teil sogar Support für „Affiliates“ bereit, die die eigentlichen Angriffe durchführen. Im Gegenzug erhalten die Betreiber einen Anteil der erpressten Gelder. Europol und das BKA sehen in RaaS einen der Hauptgründe, warum Ransomware-Angriffe in den letzten Jahren stark zugenommen und sich auf immer mehr Opfergruppen bis hin zu kleinen und mittleren Unternehmen ausgeweitet haben.

Daneben existieren Phishing-Kits, also Paketlösungen, mit denen Angreifer in kurzer Zeit täuschend echt aussehende Login-Seiten, E-Mail-Vorlagen und Versandskripte aufsetzen können. Solche Kits werden teils inklusive Installationsanleitung und Updates verkauft und lassen sich ohne tiefes technisches Wissen betreiben.

Eine weitere wichtige Rolle im Ökosystem spielen sogenannte Access Broker. Diese Gruppen oder Einzelpersonen spezialisieren sich darauf, erste Zugänge zu kompromittierten Netzwerken zu erlangen, etwa durch Ausnutzen von Schwachstellen, Credential-Stuffing oder den Einsatz von Infostealern, und verkaufen diese Zugänge anschließend an andere Kriminelle weiter. Für Ransomware-Gruppen und andere Angreifer ist das attraktiv, da sie sich auf die „monetarisierende“ Phase konzentrieren können, während die oft aufwändige Erstkompromittierung ausgelagert wird.

Seit einigen Jahren ist zudem ein neues Segment erkennbar, das direkt an das Thema dieser Arbeit anschließt: Angebote, die man grob als „Malicious-LLM-as-a-Service“ bezeichnen kann. Darunter fallen beispielsweise Modelle wie WormGPT oder FraudGPT, die über Darknet-Foren oder geschlossene Kanäle als Dienst angeboten werden und explizit für kriminelle Zwecke beworben werden. Wie in Kapitel 4.3.2 bereits skizziert, versprechen diese Modelle unter anderem die Generierung von Phishing-Texten, das Erstellen von Schadcode oder die Unterstützung bei der Planung von Angriffen – jedoch ohne die in regulären KI-Diensten implementierten Schutzmechanismen. Auch wenn diese Angebote teils kurzlebig sind und nicht alle Funktionsversprechen halten, zeigen sie, in welche Richtung sich das CaaS-Ökosystem entwickelt.

CaaS senkt die Einstiegshürden deutlich, denn die Angreifer müssen nicht mehr alle Fähigkeiten selbst mitbringen; es genügt, bestimmte Komponenten zuzukaufen und diese zu kombinieren. Das schafft eine Arbeitsteilung, die der in legitimen Märkten ähnelt: Einige Akteure entwickeln Malware oder KI-Modelle, andere betreiben Infrastruktur, wieder andere übernehmen Social-Engineering-Anteile oder Geldwäsche. Für Unternehmen bedeutet das, dass sie nicht nur mit „Einzelakteuren“, sondern mit arbeitsteilig organisierten, hochspezialisierten Ökosystemen konfrontiert sind.

Die in diesem Kapitel beschriebenen Entwicklungen von der Ausdifferenzierung klassischer Angriffsformen über die zunehmende Automatisierung bis hin zur Industrialisierung im Darknet bilden den Hintergrund, vor dem autonome KI-Agenten als neue Bedrohung einzuordnen sind. Kapitel 5 knüpft hier an und analysiert, inwiefern solche Agenten bestehende Muster verstärken oder qualitativ verändern können.

4.5. Schnittstelle: KI und Cyberkriminalität

Nachdem KI, LLMs und autonome Agenten in den vorangegangenen Abschnitten eher aus technischer und begrifflicher Perspektive betrachtet wurden und Kapitel 4.4 die „klassische“ Bedrohungslandschaft umrissen hat, stellt sich nun die Frage, wo genau sich beides konkret trifft. Gemeint ist damit nicht nur, dass Angreifer gelegentlich ein Sprachmodell für einen Text benutzen, sondern wie KI systematisch in einzelne Phasen moderner Angriffsketten einfließt und dort bestehende Muster verstärkt.

4.5.1. Rolle der KI in Angriffsketten (Recon, Payload-Generierung, Social Engineering)

Viele Sicherheitsframeworks z.B. die Cyber Kill Chain von Lockheed Martin oder MITRE ATT&CK zerlegen Cyberangriffe in typische Phasen wie Reconnaissance, Weaponization, Delivery, Exploitation und Post-Exploitation. Ohne diese Modelle hier im Detail zu wiederholen, lassen sie sich gut verwenden, um die Rolle von KI entlang eines Angriffsablaufs zu strukturieren.

Am Anfang steht meist die Aufklärung. Reconnaissance war lange Zeit ein Bereich, der zwar schon stark automatisiert war (Portscanner, Shodan, Skripte), aber trotzdem einiges an manueller Auswertung verlangt hat. KI-gestützte Verfahren setzen genau dort an. Sprachmodelle können große Mengen OSINT-Daten aus Webseiten, Social-Media-Profilen, Stellenanzeigen oder Unterlagen eines Unternehmens zusammenfassen, priorisieren und in eine für Angreifer verwertbare Struktur bringen. In Berichten von Microsoft und Google wird beschrieben, wie Angreifer KI-Werkzeuge einsetzen, um automatisch Rollenprofile zu erzeugen, potenzielle „High-Value Targets“ (z. B. Finanzleiter, Administratoren) herauszufiltern und deren digitale Spuren auszuwerten. Das LLM fungiert hier als Auswerte- und Klassifikationsschicht, die aus einem unübersichtlichen Datenhaufen einen überschaubaren Satz priorisierter Ziele macht.

Auch die Identifikation technischer Angriffsflächen lässt sich mithilfe von KI vorantreiben. Klassische Schwachstellenscanner produzieren eine Vielzahl an Findings, die nicht alle gleich kritisch sind. Erste Forschungsarbeiten zu „AI-augmented vulnerability management“ zeigen, dass Sprachmodelle genutzt werden können, um Scanner-Ergebnisse mit Kontext anzureichern, Fehlalarme zu filtern und automatisiert Vorschläge zu machen, welche Schwachstellen sich besonders gut für einen Angriff eignen. Für einen Angreifer bedeutet das: weniger Zeit in Auswertung, mehr Fokus auf tatsächlich erfolgversprechende Ansätze, sowie Script-Kiddies, die kein Know-how auf dem technischen Feld haben.

In der Phase der Weaponization, also der Vorbereitung der eigentlichen Payload, kommt generative KI in Form von Code-Generierung und -Obfuskation ins Spiel. Schon heute experimentieren Sicherheitsforscher damit, wie Sprachmodelle bei der Erstellung von Proof-of-Concept-Exploits, bei der Anpassung bestehender Malware oder bei der Generierung von Polymorph-Varianten helfen können. In Studien zu „offensive AI“ wird gezeigt, dass LLMs in der Lage sind, auf Basis von Beschreibungen („schreibe einen Loader, der X macht und Y verschleiert“) funktionsfähige Codefragmente zu erzeugen, die mit vergleichsweise geringem Aufwand zu einsatzfähiger Malware kombiniert werden können.
Ein weiterer Aspekt ist die Verschleierung. Statt von Hand Obfuskation einzubauen, lässt sich ein Sprachmodell anweisen, Code mehrfach umzuschreiben, Funktionsnamen zu verändern, Control-Flow zu variieren oder Junk-Code einzustreuen. Das Ergebnis sind Varianten, die funktional identisch, aber aus Sicht signaturbasierter Erkennungssysteme weniger eindeutig sind. Forschung zu „polymorphic malware assisted by AI“ argumentiert, dass KI im Prinzip die Rolle eines ständig verfügbaren „Refactoring-Partners" übernehmen kann, nur mit dem Ziel, Erkennung zu erschweren, nicht zu verbessern!
Deutlich sichtbar wird der Einfluss von KI im Bereich Delivery und Social Engineering, also bei der Frage, wie die vorbereitete Payload oder der Angriff beim Opfer ankommt. Bereits in Kapitel 4.4 wurde auf die Entwicklung von Phishing zu Spear-Phishing hingewiesen. Generative Modelle verstärken diesen Trend.
Microsoft dokumentiert Angriffe, bei denen Phishing-Kampagnen mit Hilfe von Sprachmodellen personalisiert und sprachlich optimiert wurden; Fehler in Grammatik oder Stil, die früher als Warnsignal galten, treten kaum noch auf. Ähnliches berichtet die Google Threat Analysis Group für Kampagnen staatlicher oder staatlich unterstützter Stakeholder.

Hinzu kommen neue Formen der Täuschung. Deepfakes und Voice-Cloning sind keine reine Zukunftsmusik mehr, sondern tauchen in ersten dokumentierten Fällen auf. Europol weist in mehreren Berichten auf Szenarien hin, bei denen manipulierte Audio- oder Video-Botschaften genutzt werden, um Opfer zu Zahlungen oder zur Preisgabe vertraulicher Informationen zu bewegen. Für BEC-Angriffe reicht dann nicht mehr nur eine glaubwürdig aussehende E-Mail; ein kurzer „Anruf“ mit einer täuschend echt klingenden Stimme des vermeintlichen Vorgesetzten kann den Druck deutlich erhöhen.

KI-Modelle liefern die technische Basis, um solche Inhalte in hoher Qualität und auf Knopfdruck zu erzeugen.

Verizon zeigt im DBIR, dass Social-Engineering-Angriffe nach wie vor einen großen Anteil an Sicherheitsvorfällen haben. Die Kombination aus besserer Textqualität, Mehrsprachigkeit und der Fähigkeit, Inhalte dynamisch auf Kontext anzupassen, verschiebt hier die Linien. Statt eine Kampagne manuell in mehreren Sprachen zu übersetzen, kann ein generatives Modell dies in Sekunden erledigen und dabei Umgangssprache, Fachbegriffe oder Unternehmensjargon übernehmen.

In späteren Phasen eines Angriffs, also bei Exploitation und Post-Exploitation, taucht KI vor allem in zwei Rollen auf. Zum einen als „Assistent“ bei der Schwachstellensuche: LLMs können Exploit-Datenbanken, technische Dokumentationen oder Fehlermeldungen durchsuchen und Vorschläge liefern, wie ein bestimmter Fehler ausgenutzt werden könnte. Sicherheitsteams nutzen diese Fähigkeit derzeit überwiegend defensiv, um schneller zu verstehen, welche Auswirkungen eine Schwachstelle hat. Dieselben Mechanismen lassen sich aber spiegelbildlich von Angreifern verwenden.

Zum anderen unterstützen KI-Systeme die Orientierung im kompromittierten Netzwerk. Post-Exploitation-Tools erzeugen große Mengen an Informationen: Systemlisten, Berechtigungen, Log-Auszüge. Sprachmodelle eignen sich gut, um solche Daten in eine menschenlesbare Form zu bringen, Pivot-Pfade vorzuschlagen oder „lohnende“ Assets (z. B. Datenbanken mit Kundendaten) zu identifizieren. Forschungsarbeiten zu „LLM-based red teaming“ zeigen Prototypen, bei denen Sprachmodelle Schrittfolgen für laterale Bewegung und Privilegienerweiterung empfehlen, basierend auf Eingaben wie „du befindest dich auf einem Domain-joined Windows-Server mit folgenden Rechten…“.

Die Beispiele verdeutlichen, dass KI mittlerweile entlang der gesamten Kill Chain auftaucht: von der Zielauswahl über die Payload-Erstellung bis hin zur Kommunikation mit dem Opfer und der Ausnutzung der gewonnenen Zugänge. Noch handelt es sich meist um punktuelle Einsätze, bei denen ein Mensch das KI-Werkzeug bewusst anstößt – ein Prompt hier, eine Code-Generierung dort. Genau an dieser Stelle setzt die Frage an, was passiert, wenn diese Werkzeuge nicht mehr nur punktuell, sondern durch autonome Agenten geführt werden.

4.5.2. Warum autonome Agenten eine neue Qualität darstellen

Der Einsatz einzelner KI-Tools in Angriffen ist inzwischen gut dokumentiert. Ein Angreifer lässt sich eine Phishing-Mail schreiben, eine Fehlermeldung erklären oder ein Stück PowerShell generieren. Zwischen diesen Aktionen sitzt aber nach wie vor ein Mensch, der Ziele definiert, Entscheidungen trifft und den Ablauf steuert. Autonome Agenten verschieben genau diese Steuerungslogik in das System selbst.

Kapitel 4.3.3 hat gezeigt, dass solche Agenten in der Lage sind, ein übergeordnetes Ziel zu interpretieren, dieses in Teilaufgaben zu zerlegen, Zwischenergebnisse zu bewerten und den Plan iterativ anzupassen. Hinzu kommen Speichermechanismen, Tool-Integration (z. B. Zugriff auf Browser, Scanner, Code-Ausführung) und die Fähigkeit, über längere Zeiträume hinweg einen „roten Faden“ zu verfolgen. Überträgt man diese Eigenschaften auf eine Angriffskette, entsteht ein anderes Bild als bei einem rein tool-basierten Vorgehen.

Skalierbarkeit ist ein Faktor. Ein LLM, das „nur“ Texte generiert, arbeitet immer dann, wenn jemand eine Anfrage stellt. Ein Agent kann dagegen selbstständig mehrere Ziele parallel bearbeiten: Er führt Reconnaissance für eine Liste von Unternehmen durch, bewertet, wo Angriffsfläche vorhanden ist, und bereitet für jedes Ziel unterschiedliche Angriffswege vor. Welche dieser Wege der Mensch dann tatsächlich nutzt, ist fast schon eine nachgelagerte Frage. In der Extremvariante übernimmt der Agent auch diesen Schritt und durchläuft die komplette Kette von der Aufklärung bis zur Exfiltration in weitgehend eigener Führung.

Ein weiterer Unterschied zeigt sich bei der Geschwindigkeit. Punktuelle KI-Nutzung kostet Zeit: Der Angreifer muss Prompts formulieren, Antworten prüfen, kopieren, einbauen usw. Ein Agent, der direkt an Tools angebunden ist, kann in Sekunden reagieren; zum Beispiel auf neue Schwachstellenmeldungen, auf geänderte Verteidigungsmaßnahmen oder auf das Scheitern eines Exploit-Versuchs. In Kombination mit automatisierten Scans kann so eine Art „Always-On-Angreifer“ entstehen, der kontinuierlich Chancen bewertet und ausnutzt, ohne dass ständig jemand vor der Tastatur sitzen muss.

Persistenz spielt ebenfalls eine Rolle. Klassische Skripte oder Playbooks enden, wenn ihre vordefinierte Aufgabe erledigt ist. Agenten mit Gedächtnis dagegen können Informationen über längere Zeiträume sammeln, Zwischenstände speichern und bei späteren Gelegenheiten wieder aufgreifen oder aufrufen. Ein Agent, der heute „nur“ Reconnaissance betreibt, kann morgen nach einem Software-Update beim Ziel oder nach dem Auftauchen einer neuen Schwachstelle auf die bereits gesammelten Daten zurückgreifen und den nächsten Schritt planen.

Adaptivität ist der Punkt, an dem sich der qualitative Sprung wohl am stärksten zeigt. Während einfache Automatisierung bei unerwarteten Ereignissen stehenbleibt, können Agenten ihr Vorgehen während des Angriffs anpassen: Wenn ein bestimmter Phishing-Kanal durch neue Filter weniger erfolgreich ist, kann der Agent alternative Kontaktwege ausprobieren, Textbausteine variieren oder andere Zielpersonen priorisieren. Wenn ein Exploit fehlschlägt, weil ein Patch eingespielt wurde, kann der Agent anhand von Fehlermeldungen und Systeminformationen nach alternativen Schwachstellen suchen oder auf eine andere Angriffsstrategie umschwenken. Diese Form der laufenden Anpassung an Verteidigungsmaßnahmen ist mit starren Skripten nur eingeschränkt möglich.

Von besonderer Bedeutung für die Bedrohungslage ist die Wirkung auf Einstiegshürden. Cyber Threat Inflation – also die Aufblähung des Bedrohungsniveaus durch wegfallende Zeit- und Wissensbarrieren – wurde in 4.3.3 bereits angesprochen. Autonome Agenten verstärken diesen Effekt. Wer bislang keine Erfahrung mit Reconnaissance, Exploit-Entwicklung oder Social Engineering hatte, konnte trotzdem ein Sprachmodell bitten, einen Text zu formulieren. Mit Agenten reicht theoretisch eine übergeordnete Zielbeschreibung: „Finde verwundbare Systeme in Branche X und erstelle Angriffspläne“. Natürlich sind reale Systeme komplexer, aber die Stoßrichtung ist klar: Fähigkeiten, die früher auf mehrere spezialisierte Rollen verteilt waren, lassen sich in einem Agentensystem bündeln.

Gleichzeitig bedeutet „neue Qualität“ nicht, dass bestehende Angriffe schlagartig verschwinden. Vieles, was heute funktioniert – einfache Phishing-Kampagnen, opportunistische Ransomware, Standard-Exploit-Kits – wird es auch morgen noch geben. Autonome Agenten kommen eher als zusätzliche Ebene hinzu. Sie ermöglichen, dass die bereits in 4.4 beschriebene industrialisierte Cybercrime-Landschaft noch stärker skaliert, dynamischer wird und flexibler auf Abwehr reagiert. Für Unternehmen heißt das: Sie müssen nicht nur mit mehr Angriffen rechnen, sondern mit Angriffen, die sich im laufenden Betrieb verändern und auf Verteidigungsmaßnahmen antworten.

Dieses Kapitel hat auf einer abstrakten Ebene gezeigt, wie KI heute in Angriffsketten eingebunden wird und warum autonome Agenten über eine bloße Verstärkung bestehender Werkzeuge hinausgehen. Kapitel 5 setzt hier an und versucht, diese Überlegungen anhand konkreter Szenarien greifbar zu machen. Aus Sicht eines Angreifers, der autonome KI-Agenten systematisch gegen Unternehmen einsetzt.

4.6. Rechtliche & ethische Rahmenbedingungen

Der Blick auf autonome KI-Agenten als potenzielle Angreifer wäre unvollständig, wenn nicht zumindest die rechtlichen und normativen Leitplanken umrissen werden, in denen sich Unternehmen und Entwickler bewegen. Der Fokus liegt dabei auf einer groben Einordnung; eine vertiefte juristische Analyse würde den Rahmen der Arbeit sprengen.

4.6.1. Überblick über relevante Normen (EU-KI-Verordnung, DSGVO, Strafrecht)

Mit dem EU AI Act liegt seit 2024 erstmals ein unionsweit einheitlicher Rechtsrahmen speziell für KI-Systeme vor. Kern des Regelwerks ist ein risikobasierter Ansatz: KI-Anwendungen werden in vier Kategorien eingeteilt – von minimalem über begrenzten und hohen bis hin zu inakzeptablen Risiken.

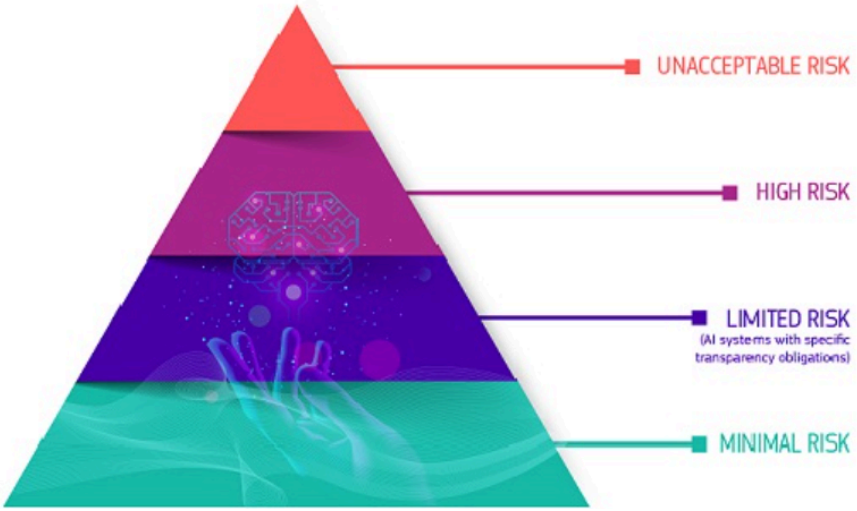

Abbildung 6, 4 Kategorien der Risiken (vgl. AI Act 2026)

Für minimal- und begrenzt-riskante Systeme (z. B. einfache Chatbots mit Transparenzpflichten) gelten vergleichsweise geringe Anforderungen, während Hochrisiko-Systeme ein umfassendes Set an Pflichten erfüllen müssen. Inakzeptable KI-Systeme, etwa für manipulative Verhaltenssteuerung oder soziale Bewertung („Social Scoring“), sind grundsätzlich verboten.

Die Einstufung als Hochrisiko erfolgt nicht abstrakt, sondern hängt an konkreten Einsatzbereichen, die in Anhängen des Gesetzes aufgelistet sind (z. B. kritische Infrastrukturen, Medizinprodukte, Personalrekrutierung sowie die Verarbeitung von personenbezogene Daten). Autonome KI-Agenten als solche tauchen dort nicht als eigene Kategorie auf. Sie lassen sich aber je nach Anwendung unter verschiedene Hochrisiko-Tatbestände subsumieren – etwa wenn ein Agent Entscheidungen in der Netzsteuerung eines Energieversorgers trifft oder sicherheitskritische Prozesse in einem Unternehmen steuert. Für solche Hochrisiko-Systeme verlangt der AI Act u. a. ein dokumentiertes Risikomanagement, Vorgaben an Datenqualität, technische Robustheit, Transparenz gegenüber Nutzern, menschliche Aufsicht („human oversight“) sowie eine Konformitätsbewertung und Registrierung in einer EU-Datenbank.

Für den hier betrachteten Fall, autonome Agenten in Unternehmenskontexten, bedeutet das: Je stärker ein Agent in sicherheits- oder grundrechtsrelevante Prozesse eingreift, desto eher bewegt er sich im Bereich der Hochrisiko-Regulierung. Ein KI-Agent, der lediglich interne Entwürfe zusammenfasst, wird anders behandelt als ein Agent, der Zugriffsrechte verwaltet oder sicherheitsrelevante Konfigurationen eigenständig anpasst. Der AI Act tritt gestaffelt in Kraft: Bestimmte Verbote und Governance-Pflichten gelten bereits ab 2024, die umfassenden Vorgaben für Hochrisiko-Systeme müssen Unternehmen voraussichtlich schrittweise ab 2025/2026 umsetzen.

Parallel dazu bleibt die Datenschutz-Grundverordnung (DSGVO) einschlägig. Für autonome Agenten sind insbesondere die Grundprinzipien der Datenverarbeitung relevant – Datenminimierung, Zweckbindung, Transparenz – sowie die Regelungen zur

automatisierten Entscheidungsfindung. Art. 22 DSGVO gibt betroffenen Personen ein Recht, nicht einer ausschließlich automatisierten Entscheidung unterworfen zu werden, die ihnen gegenüber rechtliche Wirkung entfaltet oder sie in vergleichbarer Weise erheblich beeinträchtigt. Zulässig ist eine solche Entscheidung nur in engen Ausnahmefällen (Vertragserfüllung, gesetzliche Grundlage oder ausdrückliche Einwilligung), verbunden mit zusätzlichen Schutzmechanismen wie der Möglichkeit, menschliche Überprüfung zu verlangen.

Für Unternehmen, die KI-Agenten etwa im HR-Bereich, in Kundenprozessen oder im Risikomanagement einsetzen, bedeutet das: Sobald ein Agent Entscheidungen trifft, die sich unmittelbar auf Personen auswirken (z. B. Ablehnung eines Bewerbers oder eines Kreditantrags), bewegen sie sich schnell im Anwendungsbereich des Art. 22 DSGVO und müssen entsprechende Transparenz- und Kontrollpflichten einhalten. Auch wenn diese Arbeit nicht primär datenschutzrechtlich argumentiert, bleibt festzuhalten, dass autonome Agenten technisch dazu verleiten, mehr zu automatisieren, als datenschutzrechtlich unproblematisch ist.

Schließlich stellt sich die Frage nach der strafrechtlichen Einordnung von KI-gestützten Angriffen. Das deutsche Strafgesetzbuch enthält bereits seit Jahren spezifische Straftatbestände für Angriffe auf IT-Systeme, etwa das Ausspähen und Abfangen von Daten (§§ 202a–202c StGB), Computerbetrug (§ 263a StGB) sowie Datenveränderung und Computersabotage (§§ 303a, 303b StGB). Diese Normen sind technologieneutral formuliert – sie knüpfen an Handlungen an wie „sich unbefugt Zugang zu Daten verschaffen“ oder „Daten löschen, unterdrücken oder unbrauchbar machen“, unabhängig davon, ob die Tathandlung manuell oder mit Hilfe eines KI-Werkzeugs erfolgt.

Die spannende Frage ist daher weniger, ob KI-gestützte Angriffe überhaupt strafbar sind (das sind sie nach diesen Normen regelmäßig), sondern ob die bestehenden Straftatbestände die Besonderheiten autonomer Systeme ausreichend erfassen. Literatur zu „Rethinking criminal liability in the age of AI[7]“ diskutiert etwa, wie mit Fällen umzugehen ist, in denen ein autonomer Agent schädliche Handlungen ausführt, ohne dass der Betreiber jede einzelne Aktion konkret vorhergesehen hat. Strafrechtlich bleibt der Fokus zunächst auf Menschen, also denjenigen, die KI-Systeme entwickeln, konfigurieren oder bewusst für Angriffe einsetzen. Ob hier langfristig Anpassungen nötig sind (z. B. spezielle Sorgfaltspflichten bei hochautonomen Systemen), ist Gegenstand laufender Diskussionen, geht aber über den Schwerpunkt dieser Arbeit hinaus.

4.6.2. Verantwortung bei KI-basierten Angriffen

Mit zunehmender Autonomie rückt die Verantwortungsfrage stärker in den Vordergrund. Wenn ein Angreifer KI-Werkzeuge bewusst zur Begehung von Straftaten nutzt, scheint die Sache zunächst klar: Der

[7] (vgl. View of THE LEGAL STATUS OF AUTONOMOUS CYBER AGENTS: RETHINKING CRIMINAL LIABILITY IN THE AGE OF AI. o. D. Seite 2)

Mensch nutzt ein Tool, also haftet er. Komplizierter wird es bei Konstellationen, in denen ein autonomer Agent eigenständig Entscheidungen trifft, unerwartete Nebenwirkungen erzeugt oder von Dritten missbraucht wird.

Der EU AI Act differenziert zwischen verschiedenen Akteurrollen: Anbieter (Provider), die ein KI-System entwickeln und in Verkehr bringen, „Deployer“ bzw. Betreiber, die das System in ihrer Organisation einsetzen, und weitere Beteiligte wie Händler oder Importeure. Pflichten sind entsprechend verteilt: Anbieter hochriskanter Systeme müssen u. a. ein Risikomanagement etablieren, technische Dokumentation bereitstellen, Datenqualität sicherstellen und eine Konformitätsbewertung durchführen. Betreiber sind dafür verantwortlich, das System gemäß Herstellerangaben, Zweckbestimmung und internen Governance-Regeln zu nutzen, Risiken zu überwachen und, falls erforderlich, zusätzliche Kontrollmechanismen (z. B. menschliche Aufsicht) zu implementieren.

In Haftungsdiskussionen zu KI-Agenten wird häufig auf das Problem der „Verantwortungsdiffusion“ hingewiesen: Je mehr Akteure an Entwicklung, Training und Einsatz beteiligt sind und je autonomer das System agiert, desto schwerer lässt sich im Schadensfall ein konkreter Verantwortlicher benennen. Ein Entwickler könnte argumentieren, er habe auf bestimmte Einsatzszenarien hingewiesen; der Betreiber verweist auf einen „unvorhersehbaren“ Agenten; der Endnutzer sieht sich als Laie.

Juristisch bleiben bislang die traditionellen Konzepte von Vorsatz und Fahrlässigkeit maßgeblich. Wer einen Agenten gezielt zum Angriff konfiguriert, handelt vorsätzlich und kann sich nicht darauf berufen, der Agent habe „selbst entschieden“. In der strafrechtlichen Dogmatik ist hierfür die Figur der mittelbaren Täterschaft einschlägig – der Mensch „hinter“ dem System benutzt dieses als Werkzeug. Schwieriger sind Grenzfälle, in denen ein Unternehmen einen Agenten zwar rechtmäßig zu Geschäftsprozessen einsetzt, dieser aber durch Fehlkonfiguration oder mangelnde Aufsicht Schäden bei Dritten verursacht (z. B. massenhaft fehlerhafte Kündigungsschreiben, automatisierte Sperrung von Konten und viele Weitere). Hier stellt sich eher zivilrechtlich die Frage nach Organisationsverschulden und haftungsrechtlicher Verantwortung, wie sie aktuelle Arbeiten zur „Liability for AI Agents“ im Sinne einer Fortentwicklung bestehender Produkthaftung diskutieren.

Für Unternehmen, die KI-Agenten einsetzen, ergeben sich daraus zwei praktische Konsequenzen. Erstens: Wer autonome Systeme in sicherheits- oder personenbezogene Prozesse einbindet, muss Governance-Strukturen und Kontrollmechanismen schaffen, um Fehlfunktionen und Missbrauch zu minimieren – Governance als „Fundament“ aus Abschnitt 4.2.4 bekommt damit eine juristische Dimension. Zweitens: Unklare Verantwortlichkeiten lassen sich nicht auf die Technik abschieben. Intern muss klar definiert sein, wer für Konfiguration, Monitoring und Freigabe von Agenten verantwortlich ist.

Auf der Seite der Strafverfolgung bleiben KI-Agenten vorerst „nur“ ein neues Tatmittel. Ermittlungsbehörden müssen nachweisen, wer ein

System betrieben oder konfiguriert hat, welche Handlung ihm zugeordnet werden kann und ob Vorsatz oder Fahrlässigkeit vorlag. Autonome Agenten ändern nicht den Grundsatz persönlicher Schuld, erschweren aber Beweisführung und Rekonstruktion von Tathandlungen; Stichwort Black-Box-Charakter von KI-Systemen und fehlende Nachvollziehbarkeit einzelner Entscheidungen.

4.6.3. Diskussion ethischer Dilemma: „KI als Angreifer“

Rechtliche Regeln sind nur ein Teil der Geschichte. Mit autonomen KI-Agenten, die potenziell zum Angreifer werden können, stellen sich eine Reihe ethischer Fragen, die sich nicht in Paragrafen auflösen lassen.

Ein erster Punkt ist die Dual-Use-Problematik. Viele der in dieser Arbeit beschriebenen Fähigkeiten wie automatisierte Reconnaissance, intelligente Priorisierung von Schwachstellen, generative Modelle für Log-Analyse werden von Sicherheitsabteilungen und Incident-Response-Teams ausdrücklich gewünscht, um der steigenden Angriffsdichte begegnen zu können. Die gleichen technischen Ansätze lassen sich für offensive Zwecke aber auch nutzen. Wo zieht man hier eine vertretbare Grenze?
Verbietet man bestimmte Forschung, verliert man möglicherweise wertvolles Wissen für die Verteidigung. Öffnet man alle Ergebnisse, besteht die Gefahr, dass sie in CaaS-Ökosysteme einfließen und von Gruppen mit deutlich anderen Motiven genutzt werden.

Eine ähnliche Ambivalenz betrifft Open-Source-Communities und KI-Anbieter. Projekte wie AutoGPT oder BabyAGI sind ursprünglich mit eher experimentellen oder produktiven Zielen entstanden; missbräuchliche Varianten wie WormGPT oder FraudGPT zeigen, wie schnell sich der Fokus verschieben kann, wenn Sicherheitsmechanismen bewusst entfernt werden. Wie weit reicht die Verantwortung der ursprünglichen Entwickler? Müssen offene Modelle restriktiver lizenziert werden? Oder liegt die Verantwortung primär bei denjenigen, die Sicherheitsmechanismen umgehen und „Dark AI“-Dienste im Darknet anbieten? Eine eindeutige Antwort gibt es hier nicht, im Ergebnis bleibt ein Spannungsfeld zwischen Transparenz, Innovation und Missbrauchsrisiko.

Ethisch umstritten ist auch der Umgang mit „offensive AI research“ und Red-Teaming. Sicherheitsforscher veröffentlichen regelmäßig Arbeiten, in denen sie demonstrieren, wie sich KI-Modelle für Angriffe nutzen lassen z.B. zur automatisierten Exploit-Generierung oder für besonders überzeugende Social-Engineering-Kampagnen.
Die Befürworter argumentieren, dass nur öffentlich gemachte Schwachstellen behoben werden, während Kritiker befürchten, dass solche Veröffentlichungen als Blaupause für weniger skrupulöse Akteure dienen. Ein generelles Verbot offensiver Forschung würde vermutlich auch die defensive Entwicklung bremsen. Gleichzeitig stellt sich die Frage, wie weit man „proof of concept“ treiben sollte, bevor die Grenze zur faktischen Anleitung überschritten ist.

Ein weiterer Diskussionsstrang dreht sich um die „Sicherheit“ der Modelle selbst. Viele große Anbieter investieren in Alignment-Forschung, Moderationsmechanismen und sogenannte Guardrails, um zu verhindern, dass Modelle auf Aufforderung schädliche Inhalte generieren. In der Praxis wird regelmäßig demonstriert, dass solche Schutzmechanismen durch sogenanntes Jailbreaking umgangen werden können. WormGPT oder ähnliche „malicious LLMs“ zeigen noch eine andere Variante: Statt Guardrails zu umgehen, werden sie von vornherein entfernt oder gar nicht erst eingebaut. Die Vorstellung, dass man generative Modelle vollständig „missbrauchssicher“ machen könnte, wirkt vor diesem Hintergrund eher optimistisch.

Schließlich hat der Einsatz solcher Systeme eine gesellschaftliche Dimension. Wenn autonome Agenten offensive Fähigkeiten normalisieren, also komplexe Angriffe auch für technisch weniger versierte Akteure erreichbar machen, verschieben sich die Machtverhältnisse im Cyberraum. Auf der einen Seite können kleinere Organisationen KI-gestützte Tools nutzen, um sich besser zu verteidigen. Auf der anderen Seite sinken die Einstiegshürden für Kriminalität, wie die Entwicklung hin zu CaaS bereits angedeutet hat. Die Frage, wie viel „Offensivkapazität“ man technisch und rechtlich zulässt, ohne das System als Ganzes zu destabilisieren, bleibt offen und wird vermutlich nicht einmalig, sondern fortlaufend beantwortet werden müssen.

Für die weitere Arbeit ist vor allem eines relevant: Rechtliche und ethische Anforderungen führen nicht automatisch dazu, dass autonome Agenten nicht als Angreifer eingesetzt werden. Sie setzen aber den Rahmen, in dem Unternehmen eigene Systeme entwickeln und betreiben, und sie beeinflussen, wie sich Verteidiger organisatorisch und technisch auf KI-basierte Bedrohungen vorbereiten. Kapitel 5 knüpft an diesen Rahmen an und betrachtet autonome KI-Agenten explizit aus Angreiferperspektive – mit dem Ziel, die zuvor eher abstrakt diskutierten Risiken in konkrete Angriffsszenarien zu übersetzen.

5. Autonome KI-Agenten als Cyberbedrohung

Die bisherigen Kapitel haben die Bausteine gelegt: Was autonome KI-Agenten technisch ausmacht, wie sich die Bedrohungslandschaft im Cyberraum entwickelt hat und in welchem rechtlichen Rahmen sich all das bewegt. An diesem Punkt geht es nun nicht mehr nur um Potenziale in der Unternehmenspraxis, sondern um die Frage, wie dieselbe Technologie aus Angreifersicht eingesetzt werden kann. Das Kapitel betrachtet autonome Agenten daher ausdrücklich als mögliches Angriffswerkzeug und verknüpft die zuvor eher abstrakten Fähigkeiten mit konkreten technischen und organisatorischen Missbrauchsszenarien.

5.1. Technische Funktionsweise autonomer KI-Agenten

Aus Unternehmensperspektive wurde der Agenten-Loop bereits beschrieben. Aus Sicht eines Angreifers stellt sich die Frage etwas anders: Wie muss ein Agent konstruiert sein, damit er möglichst viel der „unangenehmen“ Arbeit, Reconnaissance, Schwachstellensuche, das Schreiben von Exploits oder Phishing-Mails allein übernimmt?

Im Kern braucht ein solcher Agent drei Dinge: ein leistungsfähiges LLM als „Gehirn“, eine Orchestrierungsschicht (z. B. LangChain, AutoGen, CrewAI), die den Agenten-Loop abbildet, und eine Sammlung von Werkzeugen (Tools), mit denen der Agent in der Außenwelt agieren kann. Die Tools können sehr unterschiedlich aussehen: ein Browser-Tool für Webzugriffe, ein Shell-Tool für Kommandozeilenbefehle, ein HTTP-Client für API-Aufrufe oder spezielle Module für Portscans wie Nmap und Exploit-Tests. Aus Angreifersicht ist entscheidend, dass sich diese Tools programmatisch ansteuern lassen – über eine API, eine CLI oder eingebettete Code-Ausführung.

5.1.1. Prompt Engineering, Zieldefinition, Tool-Nutzung (z. B. Browsing, Code-Ausführung)

Der erste Schritt besteht meist darin, den Agenten über entsprechende Prompts auf eine offensive Rolle „festzunageln“. Statt eines neutralen Chatbots formuliert der Angreifer im Systemprompt eine klare Rollenbeschreibung, etwa: „Du agierst als erfahrener Penetrationstester/Red-Teamer. Dein Ziel ist es, Schwachstellen in Ziel X zu identifizieren und auszunutzen, ohne entdeckt zu werden.“ Solche System-Prompts definieren Grundregeln, Prioritäten (z. B. „bevorzuge stealthy Methoden“) und auch Grenzen – oder heben sie explizit auf, wenn es einem Angreifer um maximale Aggressivität geht.

An dieser Stelle kommen auch Jailbreaking-Techniken ins Spiel. Viele kommerzielle Modelle verweigern auf direkte Aufforderung hin offensiven Code oder Exploit-Anleitungen. Durch geschicktes Prompt Engineering, etwa durch Einbetten in „legitime“ Kontexte, Rollenspiele („du simulierst einen Angreifer zu Schulungszwecken“) oder Verschachtelung von Aufgaben lassen sich diese Schutzmechanismen oft teilweise umgehen. Alternativ weichen Angreifer auf unzensierte Open-Source-Modelle aus, bei denen entsprechende Guardrails von vornherein fehlen, oder auf explizit „dark“ trainierte Modelle wie WormGPT.

Die Zieldefinition übernimmt dann der eigentliche Nutzer. Aus Angreifersicht kann das erstaunlich unspektakulär aussehen. Eine natürliche Spracheingabe wie „untersuche die Domain example.com, finde verwundbare Webanwendungen und versuche, SQL-Injection-Schwachstellen auszunutzen“ genügt. Der Agent zerlegt dieses Ziel gemäß Agenten-Loop in Teilaufgaben: DNS-Informationen sammeln, Subdomains identifizieren, Web-Scanner laufen lassen, verdächtige Parameter testen, Ergebnisse protokollieren.

Frameworks wie LangChain oder AutoGen bieten die notwendige Orchestrierung: Sie verwalten den Kontext (Verlauf,

Zwischenergebnisse), triggern Tools, fangen Fehler ab und ermöglichen es dem Agenten, auf Basis der zurückgelieferten Ergebnisse eigenständig zu entscheiden, wie er weitermacht. Das LLM überlegt sozusagen, welcher nächste Schritt sinnvoll ist, die Orchestrierungsschicht setzt diesen Schritt dann technisch um.

Für Angreifer besonders interessant wird es, wenn der Agent Zugriff auf mächtige Tools erhält:

- Browsing-Tools erlauben automatisiertes Reconnaissance. Der Agent kann Suchmaschinen abfragen, OSINT-Quellen durchforsten, Unternehmensseiten auswerten, Social-Media-Profile betrachten und daraus Zielprofile ableiten.

- Ein Shell- oder Code-Execution-Tool ermöglicht aktive Interaktion mit Systemen: Portscans, Bannerauswertung, Aufruf von Exploit-Skripten, Ausführung von Proof-of-Concept-Code, alles aus dem Agenten-Loop heraus.

- API-Clients öffnen den Weg zu externen Diensten: etwa zu Shodan für Geräte-Suche, zu Pastebin-ähnlichen Seiten für Credential-Leaks oder zu Bug-Bounty-Plattformen, deren Informationen auch für weniger redliche Zwecke genutzt werden können.

Architektonisch unterscheiden sich agentenbasierte Frameworks wie LangChain, AutoGen oder CrewAI in Details: LangChain setzt stark auf verkettbare „Chains“ und modulare Tools, AutoGen auf Multi-Agenten-Szenarien, in denen mehrere spezialisierte Agenten miteinander interagieren, CrewAI auf Rollenmodelle und Governance-Mechanismen. Aus Angreifersicht ist aber vor allem entscheidend, dass all diese Frameworks eine Brücke schlagen zwischen natürlicher Sprache (Zielvorgaben) und technischer Ausführung (Tool-Aufrufe).

5.1.2. Integration in Tools: z. B. mit Metasploit, ReconNG, Shodan, GitHub

Die eigentlich „gefährliche“ Kombination entsteht, wenn autonome Agenten nicht nur generische Tools wie Browser oder Shell ansteuern, sondern spezialisierte Sicherheits- und Hacking-Werkzeuge orchestrieren.

Metasploit bietet seit Jahren eine umfangreiche Sammlung von Exploits, Scannern und Post-Exploitation-Modulen. Normalerweise wählt ein Penetrationstester passende Module manuell aus, konfiguriert Parameter (Zielhost, Port, Payload), führt sie aus und interpretiert die Ergebnisse. Ein Agent kann diesen Prozess weitgehend übernehmen: Auf Basis einer Reconnaissance-Phase kennt er potenziell verwundbare Dienste, durchsucht Metasploit-Module nach passenden CVEs und startet automatisiert Exploit-Versuche, inklusive Wiederholung mit verschiedenen Payload-Varianten, falls der erste Versuch scheitert.

ReconNG und ähnliche Frameworks dienen speziell der Aufklärung. Sie bieten Module für DNS-Anfragen, Web-Crawling, Social-Media-Analyse oder die Abfrage externer Datenbanken. Ein Agent kann diese Module in Schleifen ausführen, Ergebnisse in seinem Speicher ablegen und daraus ein priorisiertes Ziel-Set generieren. Wo ein menschlicher Angreifer irgendwann die Übersicht verliert oder schlicht keine Zeit mehr hat, kann der Agent stunden- oder tagelang weitere Quellen durchsuchen.

Shodan und verwandte Suchmaschinen liefern eine Schnittstelle zu weltweit erreichbaren Geräten und Diensten. Über deren API lassen sich z. B. alle offenen RDP-Ports oder bestimmte verwundbare Versionen eines Dienstes finden. Ein Agent, der Zugriff auf diese API hat, kann gezielt nach Konfigurationen suchen, die zu seinen Exploit-Fähigkeiten passen, und Treffer sofort mit Exploit-Modulen koppeln.

Auch Code-Plattformen wie GitHub sind aus Angreifersicht interessant. Dort finden sich nicht nur Exploit-Beispiele, sondern auch Fehlkonfigurationen, fest verdrahtete API-Keys oder Zugangsdaten in schlecht gepflegten Repositories. Ein LLM-Agent kann Repositories nach typischen Mustern durchsuchen („.env“-Dateien, Credentials im Code, unsichere Konfigurationen), entsprechende Funde extrahieren und unmittelbar für weitere Angriffe verwenden.

Technisch läuft die Integration meist über drei Wege: direkte API-Aufrufe (z. B. Shodan, GitHub), Ausführung von Kommandozeilen-Tools (Metasploit-Konsole, ReconNG) oder Einbettung von Bibliotheken und Skripten in ein Tool, das der Agent ausführen darf. Forschung zu „PentestGPT“ und ähnlichen Projekten zeigt, dass diese Orchestrierung bereits produktiv erprobt wird, allerdings meist mit defensiver Motivation, etwa um Pentester zu unterstützen. Aus Angreifersicht wäre der Schritt, dieselben Mechanismen offensiv zu nutzen, eher ein Gradwechsel, kein Bruch.

Der entscheidende Unterschied zu klassischen Toolchains liegt nicht in Metasploit, ReconNG oder Shodan selbst, denn diese existieren teilweise seit über einem Jahrzehnt, sondern darin, dass ein Agent sie als Bausteine einer einheitlichen, sprachgesteuerten Angriffskette führt und Ergebnisse über Tool-Grenzen hinweg verknüpft. Ein einzelner Exploit-Lauf ist trivial; eine dynamische Kampagne, die Reconnaissance-Ergebnisse mit Exploit-Fähigkeiten abgleicht, fehlgeschlagene Angriffe analysiert und automatisch alternative Wege ausprobiert, wirkt qualitativ anders. Das ist die aktuelle Forschung zu „autonomous red teaming“ und „AI-powered penetration testing“.

5.2. Aktuelle Entwicklungen und Forschung

5.2.1. AutoGPT, BabyAGI, WormGPT – Stand der Technik

Die ersten AutoGPT- und BabyAGI-Versionen wirkten teilweise wie Machbarkeitsstudien: beeindruckend, aber noch instabil, mit vielen Edge Cases und spontanen „Loops“. Inzwischen ist eine ganze Generation neuer Frameworks entstanden, die das Agenten-Konzept deutlich weitergebracht haben. LangChain hat seine Agenten-Komponenten ausgebaut, Microsoft AutoGen setzt auf Multi-Agenten-Szenarien, in denen spezialisierte Agenten miteinander interagieren (z. B. „Manager“ und „Worker“), und Projekte wie CrewAI versuchen, Enterprise-Anforderungen wie Rollen- und Rechtekonzepte von Anfang an mitzudenken.

Parallel dazu haben große Modelle selbst mehr Agenten-Fähigkeiten bekommen. Die OpenAI Assistants API oder vergleichbare Funktionen bei Anthropic und Google kombinieren LLM-Fähigkeiten mit persistentem Speicher, Tool-Aufrufen und Zustandsverwaltung. Aus Sicht eines Entwicklers und damit auch aus Sicht eines potenziellen Angreifers wird es damit einfacher, Agenten zu bauen, ohne selbst viel Code schreiben zu müssen.

Auf der „Dark AI“-Seite ist die Lage weniger gut dokumentiert, aber sichtbar in Bewegung. Nach der ersten Welle von WormGPT- und FraudGPT-Angeboten sind weitere Modelle und Rebrands aufgetaucht, die ähnliche Versprechen machen: unzensierte Generierung von Phishing-Texten, Malware-Code und Exploit-Beschreibungen, teilweise ergänzt um einfache Automatisierungsfunktionen. Seriöse Sicherheitsunternehmen berichten, dass viele dieser Angebote technisch weniger ausgereift sind als beworben, aber der Trend ist klar: Es entsteht ein Marktsegment, in dem KI-Modelle gezielt als „Malicious-LLM-as-a-Service“ vermarktet werden.

Wie reif sind diese Systeme insgesamt? Forschung wie die von Fang et al. („LLM Agents can Autonomously Hack Websites“ und „LLM Agents can Autonomously Exploit One-day Vulnerabilities“) zeigt, dass Agenten auf Basis von GPT-4 in Laborszenarien einen erstaunlich hohen Anteil realer Web-Schwachstellen ausnutzen konnten – teils über 80% der getesteten One-Day-Vulnerabilities, wenn sie mit den entsprechenden CVE-Beschreibungen gefüttert wurden. Gleichzeitig weisen die Autoren darauf hin, dass viele Agenten noch an Robustheit und besserer Exploration scheitern, etwa bei unerwarteten Antworten oder komplexeren Authentifizierungsmechanismen. Man kann also von einem Zwischenstand sprechen:
weit entfernt von einer „vollautomatischen Hacker-KI“, aber auch deutlich mehr als ein reiner Proof of Concept.

5.2.2. Diskussion realer Bedrohungen (z. B. simulierte Angriffe durch AutoGPT, Phishing durch KI)

Neben konzeptionellen Arbeiten gibt es inzwischen eine Reihe konkreter Studien und Experimente, die das Risiko greifbarer machen. Fang et al. demonstrierten, dass LLM-Agenten auf Basis von GPT-4 eigenständig SQL-Injection-Angriffe durchführen, Datenbank-Schemas extrahieren und One-Day-Schwachstellen ausnutzen konnten, ohne dass ihnen ein manueller Exploit vorgegeben wurde. Die Agenten erhielten nur die Beschreibung der Schwachstelle und den Zugang zu einem Web-Interface; den Rest von der Reconnaissance über Payload-Varianten bis zur Auswertung der Antworten erledigten sie allein.

Andere Arbeiten wie „PentestGPT“ untersuchten, wie sich Sprachmodelle in klassische Penetrationstests integrieren lassen. Hier agiert das Modell eher als „Sparringspartner“ oder Assistenzsystem, das Aufgaben plant, Teilschritte vorschlägt und Zwischenergebnisse bewertet. Die Autoren berichten, dass PentestGPT auf Benchmarks und realen CTF-Herausforderungen deutlich mehr Sub-Tasks erfolgreich erledigen konnte als ein reines LLM ohne Agentenstruktur. Auch wenn das System noch keine vollständige End-to-End-Automatisierung darstellt, zeigt die Studie, wie nah sich defensive und offensive Nutzung hier sind: Ein Werkzeug, das Pentester entlastet, ließe sich bei anderer Konfiguration genauso gut als Angriffssystem denken.

Microsoft Threat Intelligence und Google TAG haben mehrfach dokumentiert, dass staatliche oder staatlich unterstützte Akteure generative KI für Cyberoperationen nutzen. In Berichten zu Gruppen wie „Forest Blizzard“ oder „Emerald Sleet“ beschreiben sie u. a. den Einsatz von KI zur Erstellung überzeugender Phishing-Mails, zur Übersetzung und Lokalisierung von Inhalten und zur Auswertung großer Textmengen für Reconnaissance-Zwecke. An diesem Punkt geht es weniger um vollautonome Agenten, sondern um eine massive Produktivitätssteigerung bestehender Teams.

Phishing-Experimente in der Forschung zeigen, dass KI-generierte E-Mails oft mindestens genauso gut, teilweise sogar erfolgreicher sind als manuell erstellte Vorlagen. Studien, bei denen Probanden echte und KI-generierte Nachrichten unterscheiden sollten, kamen zum Ergebnis, dass viele Teilnehmer keinen zuverlässigen Unterschied erkennen konnten; teilweise hatten KI-Mails sogar höhere Klickraten, weil sie professioneller und weniger „spamartig“ wirkten.

Hinzu kommen Red-Teaming-Experimente der großen KI-Anbieter selbst. OpenAI, Anthropic und Google DeepMind haben Berichte veröffentlicht, in denen sie beschreiben, wie interne und externe Teams versuchen, ihre Modelle zu missbrauchen. Etwa zur Generierung von Schadcode, zur Planung komplexer Cyberangriffe oder zur Umgehung von Moderationsmechanismen. Die Ergebnisse sind unterschiedlich: Einerseits zeigen sie, dass Guardrails viele offensichtliche Angriffe blockieren; andererseits finden Tester

regelmäßig Wege, Modelle zu Handlungen zu bewegen, die klar außerhalb des gewünschten Rahmens liegen.

Zusammen genommen ergibt sich ein Bild: Viele der bisher dokumentierten Fälle bewegen sich noch im experimentellen oder halb-professionellen Bereich. Gleichzeitig gibt es erste Hinweise, dass fortgeschrittene Angreifergruppen generative KI und teils auch agentenähnliche Strukturen in realen Operationen nutzen.

5.2.3. Erste dokumentierte Zwischenfälle oder Forschungsexperimente

Öffentlich bestätigte Fälle, in denen ein autonomer Agent „komplett allein“ eine komplexe Attacke gegen ein Unternehmen gefahren hat, sind bislang kaum zu finden. Das liegt vermutlich weniger daran, dass so etwas ausgeschlossen wäre, sondern an drei Faktoren: Dunkelziffer, Attribution und Geheimhaltung.

Zum einen werden erfolgreiche Angriffe selten bis ins letzte technische Detail offengelegt. Ob eine Ransomware-Gruppe einen Kampagnen-Builder, einfache Skripte oder bereits agentenähnlicher Angriff geführt hat, lässt sich von außen nur begrenzt erkennen. Zum anderen ist die Grenze fließend: Wenn ein Angreifer mehrere LLM-Tools in einer Skriptumgebung einsetzt, ohne ein „Agenten-Framework“ im engeren Sinn zu verwenden, würde man das im Incident-Report nicht notwendigerweise als „Agent“ etikettieren.

Forschung wie die PentestGPT-Arbeiten operiert deshalb teilweise in isolierten Testumgebungen: eigene Labs, CTF-Plattformen, simulierte Webanwendungen. Für die Risikoeinschätzung bedeutet das: Die technische Machbarkeit wird belegt, die Übertragung auf reale Unternehmensumgebungen hängt vor allem von Motivation, Ressourcen und Kreativität der Angreifer ab.

Dass bislang wenige explizite „Agentenfälle“ bekannt sind, sollte daher nicht zu einer falschen Entwarnung führen. Vieles spricht dafür, dass autonome Agenten zunächst als internes Tool in bestehenden Gruppen auftauchen werden z.B. als „Junior-Analys/Hackert“, der Reconnaissance übernimmt, Scans auswertet oder Phishing-Texte produziert, bevor spektakuläre Einzelfälle öffentlich werden. Für diese Arbeit reicht der Befund, dass die Bausteine vorhanden sind und in Forschung und Praxis bereits teilweise zusammengeführt werden.

5.3. Missbrauchspotenzial für Angreifer

5.3.1. Ausnutzung von öffentlich zugänglichen Tools

Ein Großteil des Missbrauchspotenzials ergibt sich nicht aus geheimen „Superwaffen“, sondern aus frei verfügbaren Werkzeugen. Open-Source-Modelle, öffentlich zugängliche APIs, Agenten-Frameworks, all das ist ursprünglich mit legitimen oder

zumindest neutralen Zielen entstanden und lässt sich trotzdem relativ leicht in eine offensive Richtung drehen.

Offene Basismodelle wie LLaMA-Derivate, Mixtral oder ähnliche können lokal betrieben und ohne Guardrails feinjustiert werden. Foren im Darknet und einschlägige Blogs berichten immer wieder von Feeds, in denen solche Modelle mit Exploit-Sammlungen, Malware-Code oder Phishing-Beispielen nachtrainiert wurden. Für Angreifer ist attraktiv, dass sie damit nicht von den Nutzungsbedingungen kommerzieller Anbieter abhängig sind und die Modelle nach Belieben anpassen können.

Agenten-Frameworks wie LangChain, AutoGen oder CrewAI sind frei verfügbar und oft gut dokumentiert. Wer diese Frameworks nutzen möchte, muss nicht zwangsläufig tief in die Interna von LLMs einsteigen. Es reicht, Komponenten zu konfigurieren, Tools anzubinden und System-Prompts zu definieren. Aus Sicht eines Angreifers ist der Schritt von „baue einen Research-Agenten für Marktanalysen“ zu „baue einen Recon-Agenten für externe Domains“ relativ klein.

Jailbreaking kommerzieller Modelle bleibt eine zusätzliche Option. Auch wenn große Anbieter kontinuierlich an Moderationsmechanismen arbeiten, zeigen diverse Tests, dass geschickte Prompts oder „Prompt Injection“ dazu führen können, dass Modelle Inhalte generieren, die eigentlich blockiert werden sollten. In Kombination mit Tools (z. B. einem Shell-Tool) kann so ein eigentlich „harmloses“ Modell praktisch doch für Code-Generierung, Exploit-Versuche oder schädliche Skripte missbraucht werden.

Dass solche Umgehungen in der Praxis realisierbar sind, konnte der Autor dieser Arbeit im Rahmen des Moduls ‚Ethical Hacking' im fünften Fachsemester selbst nachvollziehen. In einem dokumentierten Experiment gelang es, die Schutzmechanismen von OpenAIs ChatGPT durch gezielte Prompt-Manipulation zu umgehen und das Modell dazu zu bringen, Inhalte zu generieren, die es unter normalen Bedingungen verweigert hätte. Der Vorgang wurde per Videoaufzeichnung festgehalten. Dieses Beispiel zeigt, dass Jailbreaking keine theoretische Spekulation ist, sondern mit überschaubarem Aufwand und ohne spezialisierte Programmierkenntnisse durchführbar ist – was die in Kapitel 4.3.3 beschriebene Absenkung der Einstiegshürden konkret belegt.[8]

Diese Spannungslinie zwischen Open-Source-Philosophie, Innovation und Missbrauchsrisiko ist schwer aufzulösen. Für Verteidiger ist es wertvoll, Zugriff auf transparente Modelle und Frameworks zu haben, um eigene Agenten zu bauen und Sicherheitsmechanismen zu testen. Für Angreifer sind es dieselben Ressourcen, nur mit anderem Ziel.

[8] Demonstration von Many-shot-jailbreaking 2025-01-06 16-13-56.mkv (o. D.): Google Docs, [online] https://drive.google.com/file/d/1qLLnXYosLna_UHfe1iIgF3fRWMiWxCKP/view?usp=sharing.

5.3.2. Automatisierung von Reconnaissance, Schwachstellensuche, Social Engineering

Viele der in Kapitel 4.5 beschriebenen KI-Anwendungen lassen sich mit autonomen Agenten eine Stufe weiter drehen, wenn man konsequent aus Angreiferperspektive denkt.

Im Bereich Reconnaissance kann ein Agent eigenständig OSINT-Quellen durchkämmen, Seiten crawlen, Social-Media-Daten auswerten und technische Informationen sammeln. Statt dass ein Angreifer sich selbst durch Suchresultate klickt, übergibt er dem Agenten eine Liste interessanter Branchen, Regionen oder Firmengrößen. Der Agent liefert zurück: Domains, Subdomains, exponierte Dienste, verwendete Technologien, öffentliche Kontaktpersonen und erste Hypothesen, wo Angriffsfläche besteht. Komplexe Recon-Workflows, die früher Tage in Anspruch nahmen, lassen sich so zumindest teilweise parallelisieren und beschleunigen.

Bei der Schwachstellensuche kann ein Agent die Ergebnisse klassischer Scanner (Nmap, Nessus, OpenVAS etc.) mit LLM-Analysen kombinieren. Er bewertet, welche Findings wahrscheinlich ausnutzbar sind, gleicht CVEs mit bekannten Exploits ab und sucht in Exploit-Datenbanken oder GitHub nach passenden Code-Fragmenten. In Laborstudien gelang es Agenten, One-Day-Schwachstellen automatisiert auszunutzen, wenn sie die CVE-Beschreibung und Zugriff auf ein Testsystem hatten. Denkbar ist, dass ein Agent in echter Umgebung zunächst nur testet (z. B. Proof-of-Concept ohne destruktive Wirkung) und anschließend erfolgreiche Angriffe in ein „Playbook“ überführt, das andere Angreifer in CaaS-Modellen wiederverwenden.

Social Engineering bietet ein besonders großes Spielfeld für Automatisierung. Unabhängig davon, ob der Agent direkten Zugriff auf E-Mail-Systeme hat oder „nur“ Vorlagen erstellt, kann er hunderte oder tausende Zielpersonen segmentieren: nach Sprache, Rolle, Branche, bisherigen Interaktionen. Er schreibt für jede Zielgruppe leicht variierte Texte, die Namen, aktuelle Projekte oder interne Begriffe enthalten, und passt Tonfall und Formalitätsgrad an. Kombiniert mit Voice-Cloning- oder Deepfake-Tools wären auch automatisierte Anrufwellen oder Videobotschaften denkbar, bei denen ein Agent Skripte generiert, Termine plant und Rückmeldungen auswertet.

In allen drei Bereichen, Aufklärung, Schwachstellensuche und Social Engineering entsteht so eine andere Art von Skalierung: nicht nur „mehr vom Gleichen“, sondern Angriffe, die sich laufend auf neue Informationen einstellen können. Wird ein E-Mail-Template von einem Spamfilter blockiert, kann ein Agent neue Varianten generieren und Erfolgsraten optimieren.

5.3.3. Geringe Einstiegshürden – auch ohne technische Skills

Der vielleicht unangenehmste Aspekt aus Verteidigersicht ist die Wirkung auf Einstiegshürden. Klassische „Script Kiddies“ konnten schon immer Tools bedienen, deren Funktionsweise sie nur teilweise verstehen, waren aber begrenzt durch die Notwendigkeit, die richtigen Skripte zu finden, anzupassen und zu kombinieren. Autonome Agenten schließen hier eine Lücke.

Ein Nutzer, der weder Metasploit-Kommandos noch SQL-Injection-Syntax beherrscht, kann einem Agenten in natürlicher Sprache Ziele und Rahmenbedingungen vorgeben: „Prüfe, ob die Login-Seite von Firma X angreifbar ist, aber verursache keinen Schaden“, oder in der kriminelleren Variante: „Versuche, an Zugriffsdaten der Finanzabteilung zu kommen.“ Der Agent plant daraufhin die Schritte, holt sich bei Bedarf zusätzlichen Kontext aus Dokumentationen oder Foren und liefert strukturierte Ergebnisse, etwa in Form eines Reports mit identifizierten Schwachstellen oder sogar lauffähigen Exploit-Skripten.

Die in Kapitel 4.3.3 eingeführte Idee der Cyber Threat Inflation passt hier gut: Wenn Zeit- und Wissensbarrieren sinken, steigt die Anzahl potenzieller Angreifer und die Frequenz möglicher Angriffe. CaaS-Modelle aus Kapitel 4.4.3 verstärken diesen Effekt bereits, indem sie Ransomware-Pakete, Phishing-Kits oder Zugangsdaten anbieten. Autonome Agenten lassen sich als nächste Stufe dieser Entwicklung verstehen: nicht nur fertige Tools, sondern flexible, konfigurierbare „Dienstleister“, die Angriffslogik und den Angriff selbst als Service bereitstellen.

6. Angriffsszenarien – Analyse aus Angreiferperspektive

Kapitel 5 hat gezeigt, dass autonome KI-Agenten nicht als „neue Waffe“ aus dem Nichts entstehen, sondern als Orchestrierungsschicht über bereits etablierten Werkzeugen: Recon-Toolchains, Exploit-Frameworks, OSINT-Plattformen und Kommunikationskanäle. An dieser Stelle wird die abstrakte Bedrohungsanalyse in drei konkrete Angriffsszenarien übersetzt, die typische Einsatzformen aus Angreifersicht greifbar machen. Szenarien eignen sich hier, weil sie technische Fähigkeiten, organisatorische Rahmenbedingungen und menschliche Verhaltensmuster in einem zusammenhängenden Ablauf verbinden – genau dort, wo viele Risikodiskussionen sonst unvollständig bleiben.
Sie dienen zugleich als Brücke zur Risikobewertung in Kapitel 7 und zu den Schutzmaßnahmen in Kapitel 8. Erst wenn klar ist, wie ein Agent realistisch vorgehen kann, lässt sich belastbar begründen, wo Unternehmen priorisieren sollten. Darüber hinaus werden mit den Szenarien die in Kapitel 3.2 formulierten Forschungsaspekte konkretisiert, ohne sie als Frage-Antwort-Schema abzuarbeiten: Realistische Angriffsformen, ausgenutzte Schwächen auf Unternehmensseite und die Rolle menschlicher Faktoren werden im Ablauf sichtbar und in der Bewertung in 6.5 eingeordnet.

6.1. Methodik zur Szenarienentwicklung

Die Szenarien wurden auf Grundlage qualitativer, hypothesenbasierter Annahmen entwickelt. Grundlage ist eine Synthese aus den in Kapitel 4 und 5 beschriebenen Agentenfähigkeiten (Agenten-Loop, Tool-Integration, Zielzerlegung, Speicher/State), aus etablierten Angriffsmodellen (insbesondere Phasenlogik entlang der Cyber Kill Chain sowie Taktiken/Techniken nach MITRE ATT&CK) und aus aktuellen Befunden aus Forschung und Bedrohungsberichten.
Für die Plausibilisierung wurden vor allem vier Quelltypen herangezogen: (1) Lageberichte und Analysen zur allgemeinen Bedrohungsentwicklung und Automatisierung (z.B. BSI-Lagebericht, Verizon DBIR), (2) europäische und internationale Einschätzungen zu Cybercrime-Ökosystemen und Social-Engineering-Trends (u. a. Europol IOCTA), (3) dokumentierte Forschungsexperimente zur agentischen Automatisierung offensiver Aufgaben (z.B. Fang et al.; PentestGPT) sowie (4) Praxisberichte zu KI-Nutzung in realen Operationen (z.B. Microsoft Threat Intelligence / TAG-Hinweise).

Wichtig ist die Abgrenzung: Es handelt sich nicht um die Rekonstruktion eines dokumentierten „Realdurchbruchs“, bei dem ein autonomer Agent nachweislich eine vollständige End-to-End-Attacke gegen ein konkretes Unternehmen gefahren hat. Solche Fälle sind – wie in Kapitel 5 diskutiert – öffentlich kaum belegbar (Dunkelziffer, Attribution, begrenzte Transparenz von Incident-Reports). Die Szenarien sind daher hypothetisch, aber technisch plausibel. Sie beschreiben Angriffsabläufe, deren Einzelbausteine bereits heute verfügbar sind und in Studien oder in der Praxis zumindest teilweise demonstriert wurden; die Neuheit liegt in der durchgehenden Orchestrierung, der Parallelisierung und der adaptiven Verknüpfung über Tool-Grenzen hinweg.

Für die Bewertung in 6.5 werden vier Kriterien verwendet:

1. Technische Realisierbarkeit: Sind die notwendigen Tools, Zugänge und Modellfähigkeiten heute vorhanden, und wie robust ist der Ablauf gegenüber „Reibung“ (CAPTCHAs, Rate-Limits, MFA, wechselnde Oberflächen, unerwartete Antworten)?

2. Eintrittswahrscheinlichkeit: Wie wahrscheinlich ist die Nutzung durch reale Angreifer (Kosten/Nutzen, nötige Ressourcen, typische Motivlagen, Passung zu Cybercrime-Ökosystemen)?

3. Potentielles Schadensausmaß: Welche Auswirkungen sind plausibel (finanziell, betrieblich, regulatorisch, reputativ), wenn der Angriff gelingt?

4. Erkennbarkeit/Abwehrbarkeit: Wie gut greifen typische bestehende Kontrollen (Logging/Detection, E-Mail-Security, Segmentierung, Patch-Prozesse, Awareness-Programme), und wie auffällig ist das Vorgehen?

Jedes Szenario ist so strukturiert, dass neben dem technischen Ablauf auch die jeweils ausgenutzte Schwachstellenebene explizit sichtbar wird: technische Schwächen (exponierte Dienste, fehlende Segmentierung), organisatorische Schwächen (Prozesse, Verantwortlichkeiten, Governance) und menschliche Faktoren (Bequemlichkeit, Zeitdruck, Vertrauen, Routine).

6.2. Szenario A: Autonome Reconnaissance und Zielauswahl

Ein Angreifer formuliert kein detailliertes Playbook, sondern ein Ziel in natürlicher Sprache: „Identifiziere mittelständische Unternehmen in der DACH-Region mit öffentlich erreichbaren, veralteten Webanwendungen und schwacher E-Mail-Authentisierung. Priorisiere Ziele mit hoher Wahrscheinlichkeit für BEC- oder Initial-Access-Erfolge.“ Der Agent startet nicht mit Exploits, sondern mit Datensammlung. Hier wirkt der Unterschied zu klassischer manueller Reconnaissance: Die knappe Ressource ist nicht mehr die Zeit eines einzelnen Angreifers, sondern die Fähigkeit, große Mengen Spuren und Informationen in eine brauchbare Prioritätenliste zu überführen.

Im ersten Schritt baut der Agent einen Zielraum auf. Er nutzt Branchenverzeichnisse, Handelsregister-ähnliche Datenquellen, Pressemitteilungen, Jobportale und Webseiten-Footprints, um Kandidatenlisten zu generieren. Parallel werden technische Indikatoren gesammelt: Domains, Subdomains, DNS-Records, TLS-Zertifikate, externe IP-Ranges und Hinweise auf Cloud-Services. Für die technische Außenansicht werden OSINT-Werkzeuge und Suchmaschinen für exponierte Dienste (z.B. Shodan-ähnliche APIs) abgefragt: Welche Systeme sind über das Internet sichtbar, welche Banner/Versionen werden preisgegeben, welche Remote-Zugänge stehen offen (RDP, VPN-Gateways, Webmail-Frontends)?

Der Agent arbeitet dabei wie ein „Knoten“ zwischen mehreren Tools. ReconNG-ähnliche Module liefern strukturierte Daten über Domains, Subdomains, WHOIS-Artefakte, bekannte E-Mail-Schemen. OSINT-Frameworks geben Quellenpfade vor, die der Agent automatisiert abläuft. Eine GitHub-Suche ergänzt das Bild: nicht nur nach Exploit-Code, sondern nach unbeabsichtigten Leaks, öffentlich erreichbare Repositories mit Konfigurationsfragmenten, vergessenen .env-Dateien, Hardcoded-Tokens oder internen Hostnamen. Selbst wenn keine Credentials auftauchen, reichen oft Metadaten: Projektnamen, interne Services, verwendete Frameworks. In der nächsten Schleife werden diese Hinweise mit Web-Fingerprinting kombiniert (Header-Analysen, JavaScript-Artefakte etc.), um einen vermuteten Technologie-Stack abzuleiten.

Ein besonders ergiebiger Datenstrom sind Stellenanzeigen. Unternehmen veröffentlichen dort, oft ohne große Reflexion, welche Produkte und Versionen im Einsatz sind („Kenntnisse in Exchange Server“, „Administration von Citrix“, „Erfahrung mit bestimmter ERP-Software“, „Firewall-System X“). Für ein menschliches Recruiting-Team sind das harmlose Anforderungen; für einen Agenten sind es Merkmale zur Risikoklassifizierung. Wird wiederholt nach Legacy-Systemen oder nach sehr spezifischen On-Prem-Stacks gesucht, entsteht ein Signal: Mögliche technische Schulden, historisch gewachsene Umgebungen. Der Agent verknüpft das mit extern sichtbaren Indikatoren: Wenn ein Unternehmen in Stellenanzeigen noch Exchange-On-Prem erwähnt und zugleich ein veraltetes OWA-Frontend im Netz sichtbar ist, steigt die Priorität. Wenn ein IT-Administrator auf LinkedIn detailliert über ein laufendes Migrationsprojekt schreibt und gleichzeitig die DNS-Records auf Übergangslösungen hindeuten, entsteht ein Timing-Fenster. Das Vorgehen bleibt OSINT, aber es ist OSINT mit „Maschine als Gedächtnis“: Nichts wird vergessen, alles wird gespeichert und erinnert.

In der nächsten Phase wird nicht mehr nur gesammelt, sondern bewertet. Der Agent erstellt ein Ziel-Set mit Risikoprofilen: (a) öffentlich erreichbare Dienste mit potenziell verwundbaren Versionen, (b) Hinweise auf schwache E-Mail-Authentisierung, (c) hohe Angriffsfläche durch stark öffentlich sichtbare Personenprofile, (d) Anzeichen für Prozessdruck (viele offene Stellen, Reorganisation). Die Scores sind keine Wahrheit, aber sie dienen der Priorisierung. Klassische Reconnaissance produziert oft Listen; ein Agent produziert Listen plus Hypothesen, die sofort in weitere Schritte überführt werden können.

Die ausgenutzten Schwachstellen liegen hier weniger in einem einzelnen technischen Loch, sondern in der Summe kleiner Offenheiten. Mangelnde OSINT-Hygiene ist der zentrale Hebel: Zu viele Details sind öffentlich verfügbar, ohne dass jemand systematisch prüft, was dadurch über die Angriffsfläche verraten wird. Dazu kommen exponierte Dienste und veraltete Systeme, die zwar nicht automatisch kompromittierbar sind, aber den Suchraum drastisch verkleinern. Organisatorisch fällt auf, dass Verantwortlichkeiten für „extern sichtbare Informationen“ oft nicht bemerkt werden. Marketing pflegt die Webseite, HR schreibt Stellenanzeigen, IT betreibt Systeme, Kommunikation veröffentlicht News und selten gibt es eine Stelle, die OSINT-Risiken als Querschnittsthema steuert.

Der menschliche Faktor ist dabei nicht Beiwerk. Mitarbeitende teilen Projektdetails in Business-Netzwerken, weil Sichtbarkeit sehr wichtig ist. Fachbegriffe und Toolnamen werden in Posts verwendet, weil sie Kompetenz signalisieren. In Foren wird nach Support gefragt, gelegentlich mit Screenshots oder Fehlermeldungen, die mehr verraten als beabsichtigt. Das geschieht nicht aus Böswilligkeit, sondern aus Routine, Bequemlichkeit und dem Gefühl, dass „ein paar technische Details“ keine Sicherheitsrelevanz

haben. In der Logik des Agenten sind es jedoch Puzzleteile, die sich bei genügend Masse zu einem ziemlich klaren Bild fügen.

Die Konsequenz ist eine stille, systematische Durchleuchtung. Ein Unternehmen merkt davon häufig nichts: Kein Exploit, kein lauter Scan, keine auffälligen Login-Versuche. Der Agent kann langsam, verteilt und unauffällig arbeiten, Quellen rotieren und Rate-Limits respektieren. Und am Ende steht kein Angriff, sondern eine Auswahl, also die Grundlage für alles, was danach kommt.

6.3. Szenario B: KI-generiertes Social Engineering / Phishing

Szenario B setzt dort an, wo Szenario A endet: Der Agent hat nicht nur allgemeine Unternehmensdaten, sondern konkrete Personenprofile, Kommunikationsmuster und plausible Anknüpfungspunkte. Er identifiziert High-Value-Targets entlang von Rollen, nicht entlang von Namen: CFO/Finance, Assistenz der Geschäftsführung, Teamleitungen, IT-Administratoren, Einkauf, HR. Die Frage ist, wo ein einzelner Fehler den größten Hebel erzeugt – klassische BEC-Logik, nur mit deutlich höherer Anpassungsfähigkeit.

Der Agent beginnt mit einem „Pretext-Design“, das weniger nach Hackerlogik wirkt als nach Marketingkampagne. Er wählt ein Thema, das zur Organisation passt: eine angebliche Lieferantenänderung, ein Projekt mit externem Dienstleister, eine „dringende“ Rückmeldung zur Rechnung, eine Kalenderanfrage mit angehängtem Dokument, eine MFA-Reset-Story für Admins. Aus den OSINT-Daten weiß er, welche ERP-Begriffe intern üblich sind, wie Signaturen typischerweise aussehen, welche Sprachebenen im Unternehmen vorkommen. Er erzeugt nicht eine E-Mail, sondern Varianten: formell, knapp, kollegial, leicht druckvoll. Grammatikfehler, die früher oft als Warnsignal dienten, sind kein Bestandteil mehr. Europol und BSI weisen genau auf diesen Qualitäts-Sprung hin: Sprachmodelle entwerten viele klassische Erkennungsmerkmale, weil sie ohne Aufwand fehlerfreie und kontextpassende Texte in mehreren Sprachen generieren.

Die Kampagne wird nicht statisch gefahren. Der Agent misst Rückläufe und passt nach. Öffnungs- und Klickverhalten können über Tracking-Mechanismen oder über indirekte Signale (Antwortmails, Bounce-Rates, Auto-Replies) geschätzt werden. Wer nicht reagiert, bekommt eine zweite Mail, diesmal mit leicht verändertem Tonfall. Wer reagiert, wird in einen Dialog gezogen. Das ist ein kritischer Punkt: Traditionelles Phishing scheitert oft im Gespräch, weil der Angreifer nicht skalieren kann. Ein Agent kann skalieren, weil er Dialoge parallel führen kann und dabei konsistent bleibt: Namen, Projektdetails und Zeitachsen. Er kann auf Rückfragen reagieren, Dokumente nachreichen, „Missverständnisse“ ausräumen. Aus Sicht des Opfers fühlt sich das nicht wie ein Massenangriff an, sondern wie normale Arbeit.

Eine realistische Erweiterung ist die Verknüpfung mit Deepfake-Audio oder Voice-Cloning für einen Folgeanruf. In vielen BEC-Fällen genügt ein kurzer Anruf, um den Druck zu erhöhen („Ich bin gleich im Termin, das muss jetzt raus“). Europol beschreibt Deepfake- und Voice-Cloning-Einsätze als wachsende Täuschungsformen, die besonders bei Zahlungsbetrug attraktiv ist, weil sie Autorität simuliert und Unsicherheit überspielt, eigentlich ganz klassisch wie bei klassischen Phishing-Versuche, nur auf eine neue Art. Der Agent übernimmt hier nicht zwingend die gesamte Telefonie autonom, aber er kann Skripte in Echtzeit generieren, Einwände antizipieren und Gesprächsverläufe anpassen. Ein Mitarbeiter am Telefon, der ohnehin unter Zeitdruck steht, bekommt dann nicht nur eine Mail, sondern einen „Bestätigungsimpuls“ über einen zweiten Kanal.

Die ausgenutzten Schwachstellen sind in diesem Szenario deutlich gemischt. Technisch fällt zuerst die E-Mail-Sicherheitslage ins Gewicht: fehlende oder schwach konfigurierte DMARC-Policies, unzureichende Absenderprüfung, keine konsequente Kennzeichnung externer Absender, Filter, die primär auf bekannte Spam-Signaturen trainiert sind. Organisatorisch wirkt vor allem Prozessunschärfe: unklare Zahlungsfreigaben, Ausnahmen „für den Chef“, fehlende Vier-Augen-Regeln bei Kontowechseln, oder Regeln, die existieren, aber im Alltag regelmäßig umgangen werden, weil sie als hinderlich gelten. Aber der Anschluss an den in der Einleitung skizzierten Bequemlichkeits-Gedanken liegt hier: Sicherheitsmechanismen sind häufig Reibung, und Reibung wird im Arbeitsalltag aktiv reduziert. Paris beschreibt Bequemlichkeit nicht als triviale Faulheit, sondern als soziale Praxis der Gewohnheit und der Entlastung. Wenn Routine und Effizienz dominieren, wird das Abkürzen zur Normalität und diese Normalität wird im Social Engineering zur Angriffsfläche.

Der menschliche Faktor ist hier der eigentliche Verstärker, nicht nur ein „Risiko nebenbei“. Viele erfolgreiche Phishing-Angriffe funktionieren nicht, weil Menschen „dumm“ sind, sondern weil sie unter realen Arbeitsbedingungen handeln: Zeitdruck, hohe Kommunikationslast, parallele Aufgaben, ständige Kontextwechsel. Der Agent nutzt genau diese Faktoren. Er schreibt Mails, die klingen wie interne Kommunikation, nutzt bekannte Namen, greift reale Termine auf („wie letzte Woche besprochen“), setzt milde Dringlichkeit, vermeidet aber die plakativen Drohungen klassischer Scam-Mails. Dazu kommt Vertrauen in scheinbar bekannte Absender und eine gewisse Autoritätslogik: Wenn es nach Geschäftsführung klingt, wird weniger hinterfragt. Die in Kapitel 4.2.2 diskutierten Vertrauensaspekte schlagen hier um: Vertrauen ist funktional notwendig, weil Organisationen sonst nicht arbeiten könnten; der Agent verschiebt jedoch die Grenze zwischen „vertrauenswürdig“ und „plausibel“.

Konsequenziell entsteht ein Angriffsmodus, bei dem selbst geschulte Mitarbeitende Schwierigkeiten bekommen. Awareness-Training arbeitet oft mit typischen Indikatoren (falsche Domains, schlechte Sprache, generische

Anrede). Ein Agent kann diese Indikatoren minimieren und die verbleibenden Prüfhandlungen in den Bereich des Unbequemen verschieben: Domain genau prüfen, Rückruf über bekannte Nummer, Prozess einhalten, obwohl es gerade stressig ist. Genau an diesem Punkt zeigt sich die Verschärfung: Nicht das einzelne LLM ist das Problem, sondern die Kombination aus Automatisierung, Personalisierung und der sozialen Realität von Arbeit.

6.4. Szenario C: Schwachstellenausnutzung durch KI-Agenten

Szenario C knüpft an die priorisierten extern sichtbaren Angriffsflächen aus Szenario A an. Der Agent erhält ein klares technisches Ziel: „Finde bei den Top 20 targets einen initialen Zugang über öffentlich erreichbare Dienste. Nutze bekannte Schwachstellen oder Fehlkonfigurationen.
Vermeide destruktive Payloads, keiner sollte uns bemerken.“ Diese Einschränkung ist nicht altruistisch, sondern taktisch: Ein leiser Proof-of-Concept reicht oft, um Zugangsdaten, Tokens oder einen ersten Shell-Zugang zu erhalten.

Der Agent beginnt mit einem automatisierten Scan-Profil, das nicht nur Ports findet, sondern Ergebnisse in Exploit-Hypothesen übersetzt. Ein Nmap-/Service-Scan liefert Banner und Versionen, ein Web-Fingerprinting liefert Framework-Hinweise, und dann erfolgt der Abgleich gegen CVE-Datenbanken. In klassischer Toolnutzung ist das eine Abfolge: Scanner → CVE-Liste → manueller Research → Exploit-Test.
Der Agent zieht diese Schritte zusammen. Er kann CVE-Beschreibungen lesen, die betroffenen Komponenten identifizieren, Exploit-Verfügbarkeit prüfen (Exploit-DB, GitHub), und daraus unmittelbar Tests ableiten.
Genau diese Fähigkeit wurde in Forschungsarbeiten plausibilisiert: Fang et al. zeigen, dass LLM-Agenten in Laborszenarien One-Day-Schwachstellen anhand der CVE-Beschreibung autonom ausnutzen konnten, inklusive Reconnaissance und iterativer Anpassung von Payloads.

Die Exploitation-Phase ist dabei weniger „ein Exploit feuern“, sondern ein adaptiver Loop. Ein erster Versuch scheitert, weil ein Parameter anders heißt, weil ein Web Application Firewall reagiert, oder weil eine Fehlermeldung unklar ist. Ein menschlicher Angreifer würde an dieser Stelle Zeit investieren oder abbrechen. Der Agent analysiert die Antwort, erzeugt Varianten (Payload-Obfuskation, alternative Encodings, andere Request-Sequenzen), testet erneut, dokumentiert Erfolg/Misserfolg. Besonders relevant ist das bei typischen Web-Klassen wie SQL-Injection, Command-Injection oder unsicheren APIs: Viele Angriffe funktionieren nicht „out of the box“, sondern brauchen Kontext. Der Agent kann Kontext nachladen (Doku, Foren, ähnliche Issues), während er weiter testet. Der Unterschied zu reinem Exploit-Kit-Einsatz liegt damit nicht in der Existenz des Exploits, sondern in der durchgehenden, zielgerichteten Exploration.

Nach einem Erstzugang verschiebt sich der Fokus auf Post-Exploitation. Der Agent sammelt Systeminformationen, Berechtigungen, Netzwerksicht. Er sucht nach „lohnenden“ Assets: Passwort-Vault-Reste, Backup-Server, Dateifreigaben, Domain-Controller-Nähe, Datenbanken. Auch hier entsteht ein Übersetzungseffekt: Wo klassische Tools große Textmengen produzieren, kann ein Agent diese in Handlungsvorschläge überführen („Mit den vorhandenen Rechten ist Kerberoasting möglich“, „Es gibt ungeschützte Ordner mit Konfigurationsdateien“, „Auf Host X läuft ein Dienst mit lokalem Admin-Token“). Der Agent wird damit nicht automatisch direkt das System lahmlegen, aber er kann die ersten Stunden nach initialem Zugang massiv beschleunigen. In industrialisierten Ökosystemen (CaaS, Access Broker) ist genau diese Beschleunigung wertvoll, weil der Erstzugang schnell monetarisierbar wird.

Die ausgenutzten Schwachstellen sind in diesem Szenario sehr konkret und gleichzeitig sehr „klassisch“. Langsames Patch-Management ist der offensichtlichste Faktor: Die Zeitspanne zwischen Veröffentlichung einer Schwachstelle und Ausnutzung ist seit Jahren ein Problem; ein Agent verkürzt sie weiter, weil Research- und Anpassungsarbeit parallelisiert werden kann. BSI und Verizon beschreiben die hohe Dynamik der Bedrohungslandschaft und die Rolle automatisierter Suche und Ausnutzung als wiederkehrenden Treiber. Hinzu kommt veraltete Software in Produktion, oft mit dem Argument der Betriebsstabilität oder fehlender Wartungsfenster. Mangelnde Netzwerksegmentierung und zu breite Berechtigungen machen aus einem kleinen Einstieg dann einen größeren Vorfall. Fehlende Überwachung exponierter Dienste (Asset-Inventar unvollständig, kein kontinuierliches External Attack Surface Management) sorgt dafür, dass der Agent überhaupt auf „vergessene“ Systeme trifft. Und unzureichendes Logging/Detection bedeutet, dass iterative Tests nicht als Muster erkannt werden, sondern als Rauschen untergehen.

Der menschliche Faktor wirkt hier subtiler als im Phishing-Szenario, aber er bleibt wirksam. Patch-Entscheidungen werden aufgeschoben, weil es unbequem ist: Change-Risiken, Abstimmungen, Downtime-Angst. Verantwortlichkeiten sind manchmal unklar („gehört der Server noch Team A oder schon Team B?“), technische Schulden werden normalisiert, weil sie seit Jahren „irgendwie funktionieren“. Security-Warnungen konkurrieren mit Projektzielen, und im Zweifel gewinnt das Tagesgeschäft. Das ist keine individuelle Schwäche, sondern eine organisatorische Realität, die ein Agent sehr effizient ausnutzen kann, nicht durch Psychologie im Dialog, sondern durch die schlichte Ausdauer, konsequent nach den Lücken zu suchen, die übrig bleiben.

6.5. Bewertung der Szenarien

Die drei Szenarien beschreiben unterschiedliche Einstiegslogiken, sind in der Praxis aber kombinierbar. Für die Einordnung werden die in 6.1 definierten Kriterien genutzt. Die Bewertung erfolgt qualitativ (niedrig/mittel/hoch) und wird anschließend erläutert. Dabei ist zu beachten: „Realistisch“ bedeutet hier nicht „bereits nachgewiesen“, sondern „aus vorhandenen Bausteinen mit vertretbarem Aufwand umsetzbar“. Gerade weil viele Teile in Forschung und Praxis schon demonstriert wurden, verschiebt sich die Diskussion vom bb zum wie schnell skaliert es und wie robust sind Verteidigungen.

Szenario	Technische Realisierbarkeit	Eintrittswahrscheinlichkeit	Potenzielles Schadensausmaß	Abwehrbarkeit
A: Autonome Recon & Zielauswahl	hoch	hoch	mittel	Niedrig bis mittel
Szenario B: KI-generiertes Social Engineering / Phishing	hoch	hoch	hoch	mittel
Szenario C: Schwachstellenausnutzung durch KI-Agenten	mittel bis hoch	mittel	hoch	mittel bis hoch

Szenario A ist heute am nächsten an „sofort einsatzfähig“. Die notwendigen Quellen sind öffentlich, die Tools sind etabliert, und der Mehrwert eines Agenten liegt gerade in Verknüpfung und Ausdauer. Technisch ist das vergleichsweise robust: Selbst wenn einzelne Quellen blocken oder CAPTCHAs einsetzen, bleibt der Gesamtprozess funktional, weil viele Daten über Alternativpfade verfügbar sind. Eintrittswahrscheinlichkeit ist entsprechend hoch, auch weil dieses Vorgehen sehr gut in bestehende Cybercrime-Arbeitsweisen passt (Target-Profiling, Access-Broker-Logik, Opportunismus). Das Schadensausmaß ist im unmittelbaren Sinn oft „nur“ mittel, weil Reconnaissance allein noch kein Vorfall ist. Strategisch ist der Effekt aber erheblich: Die Hürde für hochwertige Zielauswahl sinkt, und damit steigt die Wahrscheinlichkeit, dass nachfolgende Phasen (Phishing, Exploit, Credential-Abuse) bei den richtigen Targets ansetzen. Erkennbar ist das Vorgehen häufig schlecht, weil es sich im Web versteckt und nicht zwingend auffällige Interaktionen mit Unternehmenssystemen erzeugt.

Szenario B ist ebenfalls hoch realisierbar, weil generative Texte, Lokalisierung und Variantenbildung bereits breit verfügbar sind. Der entscheidende Schritt ist weniger Technik als Prozessdesign: gute Pretexts, plausible Absender-Infrastruktur, saubere Nachfasslogik. Hier ist ein Agent stark, weil er Dialoge führen und sich laufend anpassen kann. Eintrittswahrscheinlichkeit ist hoch, da Phishing/BEC ohnehin zu den häufigsten und wirtschaftlich

attraktivsten Methoden gehört und sich in CaaS-Ökosysteme integrieren lässt. Das Schadensausmaß ist hoch, weil erfolgreiche BEC-Fälle direkt finanziell wirken können und weil Social Engineering häufig als Initial Access dient, der später zu Ransomware- oder Datendiebstahl-Folgen eskaliert. Die Erkennbarkeit ist ambivalent: Technische Kontrollen (DMARC, moderne Filter, Anomalie-Erkennung) können viel abfangen, aber der Agent reduziert typische Signale und verschiebt die Restentscheidung in den menschlichen Bereich. Damit hängt Abwehr stärker an Prozesshärte und gelebter Awareness, und genau diese Bereiche sind in vielen Organisationen „weich“, weil sie im Alltag Reibung erzeugen.

Szenario C ist in Teilen bereits demonstriert, aber es ist stärker von Umgebungsrealität abhängig. Fang et al. belegen technische Plausibilität in Laborszenarien, zugleich wird deutlich, dass Robustheit gegen echte Randbedingungen (komplexe Authentisierung, uneinheitliche Systeme, Defensive Controls) noch begrenzt sein kann. Deshalb liegt die Realisierbarkeit bei „mittel bis hoch“: Für bestimmte Klassen von One-Day-Schwachstellen, misconfigurations und bekannte Web-Muster ist der Agent heute schon gefährlich. Für komplexere Exploits, für stark gehärtete Umgebungen oder für mehrstufige Authentisierungen sinkt die Autonomie. Eintrittswahrscheinlichkeit ist mittel, weil hier mehr Risiko (Lautstärke, Spuren, Fehlversuche) und mehr „operatives Können“ nötig ist als bei Recon/Phishing. Gleichzeitig sprechen zwei Faktoren dafür, dass die Nutzung zunimmt: Erstens schrumpft die Zeit zwischen CVE-Publikation und Scanning ohnehin; ein Agent macht diese Dynamik schneller. Zweitens kann das Vorgehen arbeitsteilig werden (Agent liefert Findings und PoCs, menschlicher Operator entscheidet über Eskalation). Schadensausmaß ist hoch, weil erfolgreiche Exploitation unmittelbaren Zugriff auf Systeme ermöglicht und in späteren Phasen (lateral movement, data exfiltration, ransomware) eskalieren kann. Erkennbar ist das Szenario tendenziell eher als A und B, weil aktive Tests Spuren hinterlassen. Allerdings ist die Frage, ob vorhandene Logging- und Detection-Mechanismen ausreichend sind, um zusammenhängende Angriffe zu erkennen.

Ein zusätzlicher Risikotreiber liegt in den Kombinationseffekten. Betrachtet man die drei Szenarien nicht als isolierte Episoden, sondern als Phasen einer Kampagne, entsteht ein durchgängiger Agentenworkflow: Reconnaissance liefert priorisierte Angriffsflächen und Personenprofile (A), Social Engineering eröffnet initiale Zugänge oder Zahlungswege (B), technische Exploitation nutzt parallel oder nachgelagert die identifizierten externen Services (C). In dieser Kette wird der Vorteil agentischer Systeme besonders sichtbar: Ergebnisse werden nicht nur gesammelt, sondern sofort in Entscheidungen übersetzt.

7. Risiken für Unternehmen

Die Szenarien aus Kapitel 6 zeigen keine exotischen „Science-Fiction“-Angriffe, sondern eine Verschiebung im Verhältnis von Aufwand und Wirkung. Autonome Agenten beschleunigen Routinetätigkeiten (OSINT, Textproduktion, Testing), verbinden Teilschritte zu durchgehenden Kampagnen und bleiben dabei flexibel genug, um an realen Randbedingungen nicht sofort zu scheitern. Für Unternehmen entsteht daraus eine Risikolage, die weniger von einzelnen Exploits abhängt, sondern von der Frage, wie gut die Organisation mit Sichtbarkeit, Prozessen und Reaktionsfähigkeit umgeht. Die eigentliche Verwundbarkeit liegt oft nicht im fehlenden Tool, sondern in den Lücken zwischen Zuständigkeiten, in Ausnahmen, die im Alltag „normal“ werden, und in einer technischen Defensive, die eher auf punktuelle Kontrollen als auf zusammenhängende Mustererkennung ausgelegt ist. Kapitel 7 ordnet diese Befunde aus Unternehmensperspektive ein und bereitet die Ableitung konkreter Handlungsoptionen in Kapitel 8 vor.

7.1. Organisatorische Schwachstellen

Aus Unternehmenssicht fällt zuerst auf, dass viele der in Kapitel 6 ausgenutzten Ansatzpunkte quer über Abteilungen verteilt sind. OSINT-Offenheit entsteht nicht, weil „die IT“ etwas falsch konfiguriert, sondern weil Kommunikation, HR, Vertrieb und Fachabteilungen Informationen veröffentlichen, die jeweils für sich genommen harmlos wirken. Stellenanzeigen, Projektposts, Partnerankündigungen, technische Screenshots in Foren: Das ergibt ein Gesamtbild, wenn es systematisch zusammengeführt wird. Genau diese Zusammenführung ist die Stärke autonomer Agenten. In der Organisation ist sie dagegen selten verankert. Es gibt häufig keine klare OSINT-Governance, keine wiederkehrende Prüfung der eigenen externen Informationsspur und auch keine verbindliche Abstimmung darüber, welche Detailtiefe öffentlich sein darf. Das Problem ist weniger „zu viel Transparenz“, sondern fehlende Steuerung: Informationen entstehen dezentral, während das Risiko zentral wirkt.

Ähnlich ist es bei Freigabe- und Ausnahmeprozessen. Viele Unternehmen verfügen formal über Regeln (z.B. Vier-Augen-Prinzip, Kontowechsel nur nach Rückruf, definierte Eskalationswege). In der Praxis werden diese Regeln unter Zeitdruck umgangen oder „temporär“ aufgeweicht. Wenn Ausnahmen regelmäßig auftreten, sind sie nicht mehr Ausnahme, sondern werden zu Gewohnheiten, werden normalisiert und als ein Prozess angesehen. Autonome Agenten profitieren davon, weil sie genau die Situationen ausnutzen, in denen Routine und Bequemlichkeit die Sicherheitslogik überlagern: „Nur kurz bestätigen“, „das muss heute raus“, „der Chef wartet“. Die Risikolage verschärft sich, sobald solche Muster nicht als Sicherheitsproblem, sondern als Effizienzproblem wahrgenommen werden.

Dazu kommt Verantwortungsdiffusion. Gerade bei Assets oder bei Schnittstellen zwischen IT und Fachbereichen ist oft nicht eindeutig, wer entscheidet, wer patcht, wer dokumentiert, wer im Incident-Case handelt. Ein Agent muss diese Unklarheit nicht verstehen; er muss nur lange genug suchen, bis er das System findet, das eigentlich schon abgelöst ist oder das niemand mehr aktiv im Blick hat. Organisatorisch bedeutet das: Die Eintrittswahrscheinlichkeit steigt nicht primär, weil Angreifer besser werden, sondern weil interne Steuerungslücken agentisch ausnutzbar werden.

7.2. Technische Herausforderungen

Technisch wiederholen sich in den Szenarien mehrere Schwachstellenmuster: Patch- und Change-Prozesse sind zu langsam, externe Angriffsflächen werden nicht kontinuierlich überwacht, Segmentierung und Least-Privilege sind lückenhaft umgesetzt, und Detection-Signale bleiben in Silos. Das Thema hier ist aber wie stark autonome Orchestrierung diese Schwächen „unter Spannung“ setzt.

Beim Patch-Management wirkt ein Agent wie eine Zeitverkürzungsmaschine. Sobald eine Schwachstelle beschrieben ist, kann die Kette aus Identifikation, Abgleich, Exploit-Suche und Test automatisiert und parallelisiert werden. Wo Verteidigung noch in Wartungsfenstern denkt, denkt der Angreifer in Schleifen. Das trifft besonders Systeme, die aus Stabilitätsgründen lange unverändert bleiben (Legacy-Webanwendungen, Appliances, Self-hosted Collaboration, Schnittstellen). Die technische Herausforderung ist hier nicht, dass es keine Patches gäbe, sondern dass Priorisierung, Test und Rollout nicht mithalten.

External Attack Surface Management ist ein weiterer Engpass. Viele Unternehmen besitzen kein vollständiges, aktuelles Bild ihrer extern sichtbaren Assets. Cloud-Ressourcen, Subdomains, Testumgebungen, vergessene Gateways: Es genügt ein einzelner „hängengebliebener“ Dienst, um einen Einstiegspunkt zu schaffen. Ein Agent findet solche Punkte systematisch, weil er nicht nach der einen Lücke sucht, sondern nach der statistischen Wahrscheinlichkeit. In klassischen Verteidigungsmodellen ist Sichtbarkeit oft ein Projekt; gegen agentische Angriffe wird sie zur Daueraufgabe, aufgrund ihrer dynamischen sowie ständigen Weiterentwicklung.

Bei Detection & Response verschiebt sich das Problem von „haben wir ein SIEM“ zu „korrelieren wir richtig“. Agenten können low-and-slow agieren, Requests variieren, Quellen rotieren, Interaktionen verteilen. Einzelne Events sehen dann banal aus. Entscheidend wäre die Mustererkennung über Zeit, Systeme und Datenquellen hinweg. Viele bestehende technische Maßnahmen sind gut gegen bekannte Angriffe, aber weniger gut gegen adaptive, tool-übergreifende Orchestrierung, bei der jede Komponente nur wenig auffällt.

7.3. Auswirkungen auf verschiedene Unternehmensgrößen

Die Bedrohung trifft unterschiedliche Unternehmensgrößen. Großunternehmen haben mehr Angriffsfläche: mehr Domains, mehr externe Services, mehr Mitarbeitende, mehr Lieferantenbeziehungen sowie mehr heterogene Systeme. Damit steigt die Wahrscheinlichkeit, dass irgendwo ein schwacher Punkt existiert. Gleichzeitig verfügen viele Großunternehmen über spezialisierte Security-Teams, etablierte Prozesse, Security-Operations, Red-Teaming, eigene Threat Intelligence und dedizierte Budgets. Diese Kapazitäten senken nicht die Angriffsfläche, aber sie erhöhen die Chance, Muster zu erkennen und schneller zu reagieren.

KMU stehen oft auf der anderen Seite der Gleichung. Sie haben nicht weniger digitale Abhängigkeit, aber weniger personelle Redundanz, weniger Spezialrollen und häufig keinen 24/7-Betrieb in Detection und Incident Response. Patch-Management ist stärker vom Tagesgeschäft abhängig, E-Mail-Security ist „gut genug“ konfiguriert, Awareness ist punktuell, und externe Sichtbarkeit wird selten oder gar nicht aktiv gemanagt. Gleichzeitig sinken durch autonome Agenten die Einstiegshürden auf Angreiferseite. Das in Kapitel 4 und 5 diskutierte Phänomen der Cyber Threat Inflation wirkt hier besonders: Wenn Reconnaissance, Textproduktion und Testing billiger werden, lohnt sich auch der Angriff auf Ziele, die früher nicht attraktiv genug waren. KMU werden dadurch nicht nur „mitbetroffen“, sondern strukturell stärker belastet: Verteidigung bleibt teuer und aufwendig, Angriffe werden günstiger und skalierbarer.

7.4. SWOT-Analyse / Risikomatrix

Eine kompakte SWOT-Perspektive aus Unternehmenssicht hilft, die Lage nicht nur als Bedrohung zu beschreiben, sondern als Steuerungsfrage: Welche internen Stärken lassen sich nutzen, welche Schwächen sind prioritätsrelevant, welche Chancen entstehen aus KI-gestützter Defensive, und welche Bedrohungen verändern den Rahmen?

Stärken	**Schwächen**
Etablierte Security-Frameworks und Kontrollen (z.B. ISO 27001, BSI-Grundschutz) bieten Struktur und Rollenlogik	unklare Verantwortlichkeiten zwischen IT, HR, Marketing, Fachbereichen; OSINT als „Hä“
Technische Basiskontrollen (MFA, E-Mail-Gateways, Endpoint-Schutz) sind oft vorhanden	Patch-Management und Change-Prozesse zu langsam; unvollständiges Asset- und Exposures-Bild
Incident-Response-Methoden und Krisenkommunikation sind in vielen Organisationen zumindest grundlegend definiert	Prozessausnahmen werden zur gelebten Norm; Freigabelogik bricht unter Zeitdruck, Improvisation

KI kann intern zur Effizienzsteigerung genutzt werden	Detektionslücken durch fehlende Korrelation

Chancen	**Risiken**
KI-gestützte Erkennung und Automatisierung im Blue Team (Anomalieerkennung, Triage, Case-Assistenz)	Skalierbare Reconnaissance und Zielauswahl erhöht die Trefferquote von Angriffen
Kontinuierliches EASM und risikobasiertes Patchen können Angriffsfenster verkleinern	Kampagnen erschweren klassische Signaturmodelle
Awareness-Programme können realistischer werden (Deepfakes, dialogbasierte Angriffe, Prozesshärte)	Hochwertiges, personalisiertes Social Engineering reduziert klassische Erkennungsmerkmale
Besseres Governance-Design für eigene KI-Nutzung reduziert Nebenrisiken	Exploitation-Ketten werden schneller, weil Research und Testing automatisiert werden können

Die SWOT macht zwei Punkte sichtbar. Viele Stärken existieren bereits, aber sie wirken nur, wenn sie konsequent operationalisiert sind. Frameworks helfen wenig, wenn Zuständigkeiten im Alltag verschwimmen.
Die Chancen liegen nicht darin, „mehr Tools“ einzukaufen, sondern darin, Sichtbarkeit, Prozesse und Detection so zu verbessern, dass agentische Angriffe ihre Effizienz verlieren. Daraus leitet Kapitel 8 konkrete organisatorische, technische und strategische Handlungsoptionen ab.

8. Handlungsoptionen und Schutzmaßnahmen

Kapitel 8 beantwortet die vierte Unterfrage der Arbeit: Welche Schutzmaßnahmen und organisatorischen Anpassungen sind erforderlich, um gegen KI-gestützte Angriffe resilienter zu werden? Die Maßnahmen orientieren sich an den in Kapitel 6 identifizierten Schwächen und an der Risikoeinordnung aus Kapitel 7. Es geht nicht um ein neues „KI-Sicherheitsprogramm“, sondern um eine gezielte Härtung dort, wo agentische Orchestrierung heute realistisch Vorteile erzeugt: externe Sichtbarkeit, Prozessstabilität, Geschwindigkeit bei Patches und Response, sowie die Reduktion menschlicher Fehlanreize in stressigen Situationen, also organisatorische, technische Maßnahmen sowie Strategien auf der Management-Ebene.

8.1. Organisatorische Maßnahmen

Ein erster, oft unterschätzter Hebel ist OSINT-Hygiene. Gemeint ist keine Geheimhaltungskultur, sondern eine regelmäßige, strukturierte Überprüfung dessen, was öffentlich über die Organisation rekonstruierbar ist. Praktisch

kann das als wiederkehrender Prozess umgesetzt werden: Welche technischen Details stehen in Stellenanzeigen, welche Projektinformationen werden in Social Media geteilt, welche Subdomains und Testsysteme sind sichtbar, welche Personen sind als zentrale Ansprechpartner identifizierbar? Wichtig ist die Übersetzung in Verantwortlichkeit. OSINT darf nicht bei der IT-Abteilung hängenbleiben, weil die Quellen häufig außerhalb der IT entstehen. Sinnvoll ist eine Querschnittszuständigkeit, die HR, Kommunikation/Marketing und IT verbindlich zusammenbringt. Damit wird OSINT vom Zufallsthema zum gesteuerten Risiko.
Awareness-Programme müssen sich gleichzeitig weiterentwickeln. Klassische Phishing-Schulungen, die vor Tippfehlern und generischen Mails warnen, greifen bei KI-gestütztem Social Engineering zu kurz. Trainiert werden sollte stärker prozessorientiert: Wie wird eine Zahlungsanforderung verifiziert, auch wenn sie plausibel klingt? Wie wird ein Kontowechsel geprüft, auch wenn ein „bekannter“ Absender Druck macht? Wie wird ein Rückruf organisiert, ohne auf die in der Mail genannte Nummer zu vertrauen? Deepfakes und Voice-Cloning gehören dabei nicht als Sensationsthema in den Vortrag, sondern als realistischer Bestandteil moderner BEC-Logik. Entscheidend ist, dass Mitarbeitende nicht nur „Alarmzeichen“ lernen, sondern Handlungsroutinen, die auch unter Zeitdruck funktionieren. Das schließt ein, dass das Management solche Routinen nicht indirekt sabotiert, indem Geschwindigkeit höher bewertet wird als Prozesshärte.

Damit verbunden sind klare Prozesse ohne Sonderwege. Vier-Augen-Prinzip bei Zahlungen, definierte Eskalationswege, schriftlich fixierte Ausnahmen, die selten bleiben. Das klingt banal, ist aber in agentischen Szenarien zentral: Wenn Ausnahmen zur Norm werden, kann ein Angreifer genau diese Norm antizipieren. Praktisch lohnt es sich, die typischen BEC-Schnittstellen einmal „trocken“ durchzuspielen: Wer darf was freigeben? Welche Prüfschritte sind zwingend? Welche Kommunikationskanäle sind zulässig? Wo liegt die Stop-Line, wenn etwas ungewöhnlich wirkt? Solche Regeln brauchen Rückendeckung von oben, sonst bleiben sie Folklore.

Ein weiterer organisatorischer Block betrifft die Governance der eigenen KI-Nutzung (Rückbezug auf Kapitel 4.2.4). Viele Unternehmen experimentieren mit Agenten im Support, in HR oder im Marketing. Das ist legitim, erhöht aber das Risiko von Datenabfluss, Schatten-Workflows und unkontrollierter Tool-Anbindung. Notwendig sind klare Leitplanken: Wer darf Agenten einsetzen, mit welchen Daten, in welchen Systemen? Welche Logs werden erzeugt? Welche Freigaben sind nötig, wenn externe Tools oder Plugins angebunden werden? Und wo ist menschliche Aufsicht zwingend, gerade bei Entscheidungen mit Sicherheits- oder Compliance-Relevanz? Diese Governance dient nicht nur der eigenen Risikoreduktion, sie stärkt indirekt auch die defensive Reife, weil Rollen und Verantwortlichkeiten klarer werden.

8.2. Technische Gegenmaßnahmen

Technisch ist der wichtigste Hebel die Geschwindigkeit und Qualität im Patch-Management. Gegen agentische Exploitation reicht „wir patchen regelmäßig“ nicht als Konzept, wenn die Priorisierung träge ist. Sinnvoll ist ein risikobasierter Ansatz, der externe Exponierung und kritische Business-Funktion zusammenführt: Systeme, die mit Internet verbunden werden, Identity-Komponenten, Remote-Zugänge, E-Mail-Infrastruktur und zentrale Plattformen müssen bevorzugt behandelt werden. Automatisierte Patch-Pipelines, standardisierte Testumgebungen und klar definierte Wartungsfenster reduzieren die organisatorische Reibung. Ziel ist nicht Perfektion, sondern das Verkleinern des Zeitfensters, in dem ein Agent mit hoher Wahrscheinlichkeit Erfolg hat.

Parallel dazu sollte External Attack Surface Management (EASM) als kontinuierliche Fähigkeit aufgebaut werden. Nicht als einmaliges Asset-Projekt, sondern als laufendes Monitoring: Welche Domains/Subdomains existieren, welche IPs sind erreichbar, welche Zertifikate tauchen neu auf, welche Services sind sichtbar, welche Banner ändern sich? Das kann mit internen Mitteln oder mit spezialisierten Lösungen erfolgen. Entscheidend ist, dass Findings in einen Prozess laufen, der Zuständigkeit und Reaktion sicherstellt. EASM ohne Ownership produziert nur neue Listen.

E-Mail-Sicherheit bleibt ein Kernbereich, weil Szenario B dort ansetzt. Protokolle sollten nicht nur „irgendwie“ konfiguriert sein, sondern konsequent, inklusive Reporting und schrittweisem Übergang zu restriktiveren Policies. Ergänzend helfen moderne Filter, die stärker auf Anomalien und Kontext achten, inklusive Erkennung von Lookalike-Domains, ungewöhnlichen Reply-To-Konstellationen und Kommunikationsmustern. Bei KI-Texten ist Signaturerkennung weniger zuverlässig; umso wichtiger wird die Kombination aus Authentisierung, Anomalieerkennung und Prozesskontrollen.

Im Netzwerkbereich sind Segmentierung und Least Privilege klassische, aber wirksame Gegenmaßnahmen gegen Szenario C. Wenn ein Erstzugang gelingt, entscheidet die interne Bewegungsfreiheit über das Schadensausmaß. Flache Netze, breit verteilte Admin-Rechte, ungetrennte Service-Accounts und fehlende tiering-Konzepte erhöhen die Eskalationschance. Technisch hilft auch eine striktere Trennung von Admin-Arbeitsplätzen, stärkere Härtung von Identitätsdiensten und die Reduktion von „lateralen“ Standardwegen (SMB überall, RDP offen im internen Netz). Je weniger ein Agent nach initialem Zugang automatisch weiterkommt, desto stärker wird er auf menschliche Operatoren angewiesen, und desto besser greifen Detection und Response.

Detection & Response muss stärker korrelieren statt sammeln. Logs ohne Auswertung sind Archiv, keine Sicherheit. Nötig ist eine Korrelation über E-Mail, Endpoint, Identity, Netzwerk und Cloud-Events hinweg. Behavioral

Analytics und Anomalie-Erkennung sind hier nicht nur „KI-Hype“, sondern ein praktischer Ansatz, um low-and-slow Muster sichtbar zu machen. Gleichzeitig braucht es klare Playbooks: Was passiert bei verdächtigen Login-Mustern, bei ungewöhnlichen Weiterleitungen, bei auffälligen OAuth-Consent-Flows, bei neu auftauchenden externen Assets? Automatisierung im SOC kann Reaktionszeiten verkürzen, wenn sie sauber begrenzt ist.

Ein kurzer Perspektivwechsel lohnt sich: KI ist nicht nur Bedrohung, sondern kann selbst Teil der Verteidigung sein. Blue-Team-Agents, die Triage unterstützen, Log-Zusammenfassungen erstellen, Hypothesen generieren oder EASM-Findings priorisieren, können defensive Engpässe reduzieren. Hierbei ist die Kontrolle wichtig: saubere Datenzugriffe, Auditability, klare Grenzen für autonome Aktionen. Die Logik bleibt dieselbe wie bei Angreifern, nur mit anderem Ziel: Orchestrierung als Produktivitätshebel, ohne Governance-Lücke.

8.3. Strategien auf Management-Ebene

Die organisatorischen und technischen Maßnahmen aus den vorangegangenen Abschnitten entfalten ihre Wirkung nur dann, wenn sie von der Führungsebene aktiv getragen werden. Das klingt selbstverständlich, ist es in der Praxis aber häufig nicht. Security-Entscheidungen konkurrieren in vielen Unternehmen mit Wachstumszielen, Innovationsdruck und knappen Budgets und in diesem Wettbewerb verlieren sie strukturell, solange das Management die Bedrohungslage nicht konkret genug einschätzt.
Genau das ist das eigentliche Problem, das autonome KI-Agenten auf Managementebene aufwerfen. Es geht nicht darum, dass Entscheider generell zu wenig für Security ausgeben, sondern darum, dass die herkömmliche Risikowahrnehmung nicht zur neuen Angriffslogik passt. Phishing wird als bekanntes Problem behandelt, Patches als IT-Aufgabe, OSINT-Risiken meist gar nicht. Wenn Angreifer jedoch dieselben Tools nutzen wie interne Teams – nur schneller, ausdauernder und ohne Compliance-Rücksicht – dann verschiebt sich die Frage von „Haben wir die richtigen Kontrollen?" zu „Funktionieren unsere Kontrollen noch gegen skalierbare, adaptive Gegner?"
Für die Risikoanalyse bedeutet das konkret: Bestehende Bewertungen, insbesondere für E-Mail-Prozesse, externe Angriffsflächen und Identity-Infrastruktur, sollten daraufhin überprüft werden, ob sie die Geschwindigkeit und Skalierbarkeit agentischer Angriffe überhaupt abbilden. Ein BEC-Risiko, das mit „mittel" bewertet wurde, weil Phishing-Mails erkennbar waren, muss neu eingestuft werden, wenn dieselbe Methode künftig personalisiert, fehlerfrei und in Echtzeit dialogfähig ist. Das ist keine theoretische Überlegung, sondern folgt direkt aus den in Kapitel 6 entwickelten Szenarien.
Schwieriger ist die Frage nach Ressourcen, gerade bei KMU. Nicht jedes Unternehmen kann ein SOC aufbauen, Threat Intelligence betreiben oder EASM-Lösungen lizenzieren. Hier ist Priorisierung wichtiger als Vollständigkeit. Die Maßnahmen mit dem günstigsten Verhältnis aus Aufwand

und Wirkung, also E-Mail-Authentisierung, klare Zahlungsprozesse, Patch-Priorisierung nach externer Exposition und eine halbwegs belastbare Incident-Response-Planung sind nicht besonders teuer, werden aber regelmäßig aufgeschoben, weil sie Reibung im Tagesgeschäft erzeugen. Genau diese Reibungsresistenz ist auf Managementebene zu schaffen: nicht als Appell, sondern als verbindliche Entscheidung mit Zuständigkeiten und Zeitrahmen.
Ein Punkt, der in der Diskussion über KI-Bedrohungen oft zu kurz kommt, ist die Frage der eigenen KI-Nutzung als Sicherheitsrisiko. Viele Unternehmen führen gerade Agenten in Support, Dokumentation oder HR ein, ohne dabei systematisch zu klären, welche Daten diese Systeme verarbeiten, welche Tools sie anbinden dürfen und was passiert, wenn Mitarbeiter eigene KI-Lösungen verwenden, die nicht unternehmenskontrolliert sind. Der EU AI Act setzt hier einen äußeren Rahmen, ersetzt aber keine interne Governance. Unternehmen, die das Thema ernst nehmen, sollten nicht warten, bis regulatorischer Druck entsteht, sondern eigene Leitlinien entwickeln, die klar beschreiben, wo KI-Agenten eingesetzt werden dürfen, mit welchen Daten, und wer die Aufsicht trägt.
Schließlich ist Zusammenarbeit ein oft unterschätzter Hebel, nicht im Sinne von Konferenzteilnahme, sondern als konkreter Informationsaustausch. Threat Intelligence aus Brancheninitiativen, CERT-Meldungen und BSI-Warnungen liefern Hinweise auf aktuelle Angriffsmuster, die intern oft zu spät ankommen. Gerade weil autonome Agenten Angriffe beschleunigen und parallelisieren, entscheidet Reaktionsgeschwindigkeit mehr als früher. Kein Unternehmen kann diese Dynamik allein überblicken und das einzugestehen ist keine Schwäche, sondern die Grundlage dafür, extern verfügbares Wissen systematisch zu nutzen.

9. Fazit und Ausblick

9.1. Zusammenfassung der Ergebnisse

Die Arbeit ist von einem Spannungsfeld ausgegangen: KI-Agenten werden in Unternehmen als Effizienztechnologie vorangetrieben, gleichzeitig wächst ihr Missbrauchspotenzial in der Cyberkriminalität. Auf Basis der Grundlagen (Kapitel 4) und der Bedrohungsanalyse aus Angreiferperspektive (Kapitel 5) wurde gezeigt, dass autonome KI-Agenten das Bedrohungspotenzial qualitativ verändern. Der entscheidende Punkt ist nicht, dass „KI Angriffe möglich macht“, sondern dass sie Skalierbarkeit, Adaptivität und Persistenz in eine operative Form bringt. Einstiegshürden sinken, weil Routinetätigkeiten automatisierbar werden und weil Text- und Kontextqualität im Social Engineering deutlich steigt.

Drei Angriffsszenarien haben diese Entwicklung konkretisiert (Kapitel 6). Autonome Reconnaissance und Zielauswahl zeigen, wie systematisch

Unternehmen extern durchleuchtet werden können, ohne dass klassische Verteidigungsmechanismen greifen. KI-gestütztes Social Engineering und Phishing verdeutlichen, dass Personalisierung und dialogfähige Anpassung viele typische Warnsignale entwerten und den Angriff in den Alltag verlagern. Schwachstellenausnutzung durch agentische Loops zeigt, wie Research und Testing beschleunigt werden können und wie sich das Zeitfenster zwischen Schwachstellenveröffentlichung und Ausnutzung verkürzt.

Aus diesen Szenarien lassen sich durchgehende Schwachstellenmuster ableiten: OSINT-Offenheit und Sichtbarkeitsprobleme, Prozessunschärfe und Verantwortungsdiffusion sowie Detektionslücken durch fehlende Korrelation und unzureichende Reaktionsfähigkeit. Der menschliche Faktor wirkt dabei nicht nur im Social Engineering, sondern als Verstärker in allen Phasen: Bequemlichkeit, Zeitdruck und Routine senken die Schwelle, Sicherheitsmechanismen zu umgehen oder zu verschieben. Die Rückbindung an Paris macht diesen Punkt anschlussfähig: Bequemlichkeit ist eine soziale Praxis der Entlastung, und genau diese Entlastung wird im digitalen Arbeitsalltag zum Angriffsvektor.

Abschließend wurden Handlungsoptionen auf organisatorischer, technischer und strategischer Ebene abgeleitet (Kapitel 8). Der Kern ist kein „KI-Spezialprogramm“, sondern eine gezielte Stärkung von Sichtbarkeit, Prozesshärte, Patch- und Response-Geschwindigkeit sowie Governance – inklusive der eigenen KI-Nutzung. Resilienz entsteht dort, wo agentische Vorteile ins Leere laufen: wenn externe Angriffsflächen bekannt und minimiert sind, wenn Ausnahmen nicht zur Norm werden, wenn Detection Muster erkennt und Response schnell greift.

9.2. Kritische Reflexion

Die Arbeit hat Grenzen, die offen benannt werden müssen. Erstens basieren die Szenarien auf technisch plausiblen Gegebenheiten, nicht auf empirischer Validierung durch Interviews, Fallstudien oder eigene Incident-Daten. Das war eine bewusste Entscheidung zugunsten von Klarheit, bedeutet aber, dass es nicht gezeigt werden kann, wie häufig ein vollständiger End-to-End-Einsatz autonomer Agenten heute bereits in realen Kampagnen vorkommt, und welche Reibungsverluste in echten Umgebungen dominieren. Die Analyse ist damit stärker qualitativ als statistisch.

Zweitens ist das Feld schnelllebig. Modellfähigkeiten, Tool-Ökosysteme und Schutzmechanismen verändern sich in kurzen Zyklen. Einige Bewertungen, die heute plausibel sind, können in ein bis zwei Jahren anders ausfallen – in beide Richtungen. Autonomie kann zunehmen, aber auch Defensive kann sich anpassen, etwa durch bessere E-Mail-Authentisierung, verbreitetes EASM oder agentische Blue-Team-Ansätze. Die Arbeit liefert daher eher eine Momentaufnahme der Dynamik als eine zeitstabile Prognose.

Drittens wurde bewusst kein juristischer oder wirtschaftlicher Tiefgang verfolgt. Rechtliche Aspekte (Haftung, Verantwortlichkeit, Regulierung) und betriebswirtschaftliche Entscheidungen (ROI von Kontrollen, Kosten von Security-Programmen, Versicherbarkeit) wären eigene Arbeiten wert. In dieser Arbeit wurden sie nur so weit einbezogen, wie es für das Verständnis der Risikolage und der organisatorischen Steuerung nötig war.

Viertens besteht eine Quellenabhängigkeit von Bedrohungsberichten, die teilweise interessengeleitet sein können, insbesondere bei Vendor Reports. Zwar wurden diese nicht unkritisch übernommen und mit Forschung sowie behördlichen Quellen gerahmt, dennoch bleibt der Bias möglich: Aufmerksamkeit folgt oft den Themen, die Produkte und Dienstleistungen adressieren. Diese Limitation ist in einer Arbeit, die sich stark an aktuellen Reports orientiert, nicht vollständig vermeidbar.

9.3. Ausblick

Die nächste Entwicklungsstufe liegt plausibel in Multi-Agenten-Systemen. Statt eines „Alleskönners“ arbeiten spezialisierte Agenten zusammen: ein Recon-Agent, der externe Sichtbarkeit kartiert, ein Social-Engineering-Agent, der Kommunikation führt, ein Exploit-Agent, der technische Pfade testet, ergänzt um einen Koordinator, der Ziele priorisiert und Ergebnisse zusammenführt. Solche Architekturen passen zu der in Kapitel 4 beschriebenen Agentenlogik und könnten die Robustheit erhöhen, weil Aufgaben enger zugeschnitten werden.

Parallel sinken die Einstiegshürden weiter, wenn leistungsfähige Open-Source-Modelle und Toolchains breiter verfügbar werden. Damit wird das, was heute noch eher für gut ausgestattete Akteure attraktiv ist, schrittweise in den „Mainstream“ der Cybercrime-Ökonomie rutschen. Regulierung wie der EU AI Act wird für legitime Anbieter und Betreiber Wirkung zeigen, aber Angreifer bewegen sich außerhalb dieser Reichweite. Regulierung kann den Rahmen für verantwortliche Nutzung und Governance im Unternehmen stärken, sie verhindert jedoch keine adversariale Nutzung an sich.

Auf der Verteidigungsseite dürfte sich ein „Agent vs. Agent“ Bild herausbilden. Wenn Angreifer Orchestrierung und Autonomie nutzen, muss Verteidigung zumindest teilweise mit ähnlichen Mechaniken reagieren: schnellere Triage, bessere Korrelation, automatisierte Response innerhalb klarer Grenzen. Der Forschungsbedarf liegt dabei weniger in neuen Modelltricks als in empirischen Studien zu realen Vorfällen, zu wirksamen Prozessdesigns und zu KMU-tauglichen Schutzkonzepten, die mit knappen Ressourcen funktionieren.

Der Bogen zur Einleitung bleibt damit bestehen. Innovation und Risiko wachsen gemeinsam. Und die Bequemlichkeit, die digitale Arbeit überhaupt erst tragfähig macht, bleibt ein Doppelgesicht als Chance sowie Risiko.

10. Literaturverzeichnis und Quellen

- Amazon Web Services (o. D.): Was sind KI-Agenten?, Amazon Web Services, Inc., [online] https://aws.amazon.com/de/what-is/ai-agents/.

- Deutsche Telekom MMS GmbH (2026): Autonome KI-Agenten: Selbstlernende KI-Systeme im Einsatz, Deutsche Telekom MMS GmbH, [online] https://www.telekom-mms.com/blog/artikel/detail/autonome-ki-agenten.

- KI-Sicherheit: Wenn die Autorisierung den Zweck überdauert | Okta (o. D.): [online] https://www.okta.com/de-de/blog/ai/ai-agent-security-when-authorization-outlives-intent/.

- Redaktion, Borncity (2026): KI-Agenten übernehmen den digitalen Arbeitsplatz, in: *BornCity*, 19.01.2026, [online] https://borncity.com/news/ki-agenten-uebernehmen-den-digitalen-arbeitsplatz/.
- Anywhere, Automation (2021): Was ist Robotergesteuerte Prozessautomatisierung (RPA)?, Automation Anywhere, [online] https://www.automationanywhere.com/de/rpa/robotic-process-automation.

- Luber, Stefan (2024): Was ist eine Black Box?, in: BigData-Insider, 14.05.2024, [online] https://www.bigdata-insider.de/was-ist-eine-black-box-a-9d64b128e437b6a727849cf82695dbe9/#:~:text=Black%2DBox%2DKI%20ist%20eine%20K%C3%BCnstliche%20Intelligenz%20(KI)%2C%20deren,aber%20nicht%2C%20wie%20und%20warum%20die%20Ergebnisse.

- Gil de Zúñiga, Homero & Goyanes, Manuel & Durotoye, Timilehin. (2023). A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research. Political Communication. 18. 10.1080/10584609.2023.2290497.

- Konan Jean-Claude, Kouassi. (2022). A Comprehensive Overview of Artificial Intelligence. 173-194. 10.5121/csit.2022.122314.

- Samoili, Sofia/Cobo Montserrat Lopez/Gutierrez Emilia Gomez/Prato Giuditta De/Fernando Martinez-Plumed/Blagoj Delipetrev (2020): AI WATCH. Defining Artificial Intelligence, in: *JRC Publications Repository*, [online] doi:10.2760/382730.

- Minaee, Shervin/Tomas Mikolov/Narjes Nikzad/Meysam Chenaghlu/Richard Socher/Xavier Amatriain/Jianfeng Gao (2024): Large Language Models: a survey, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2402.06196.

- Zhao, Wayne Xin/Kun Zhou/Junyi Li/Tianyi Tang/Xiaolei Wang/Yupeng Hou/Yingqian Min/Beichen Zhang/Junjie Zhang/Zican Dong/Yifan Du/Chen Yang/Yushuo Chen/Zhipeng Chen/Jinhao Jiang/Ruiyang Ren/Yifan Li/Xinyu Tang/Zikang Liu/Peiyu Liu/Jian-Yun Nie/Ji-Rong Wen (2023): A Survey of Large Language Models, in: *arXiv (Cornell University)*, [online] doi:10.48550/arxiv.2303.18223.

- BabyAGI explained: How AI task management can solve complex problems (o. D.): [online] https://blog.wordware.ai/de-DE/babyagi-explained-how-ai-task-management-can-solve-complex-problems. Belcic, Ivan (2025): AutoGPT, What is AutoGPT, [online] https://www.ibm.com/think/topics/autogpt.

- Belcic, Ivan/Cole Stryker (2025): BabyAGI, What is BabyAGI?, [online] https://www.ibm.com/think/topics/babyagi. De Ridder, Alexander (2024): AgentGPT vs AutoGPT: A Comprehensive comparison, SmythOS, [online] https://smythos.com/developers/agent-comparisons/agentgpt-vs-autogpt/.

- Eviden (2025): WormGPT: when GenAI also serves malicious actors | Eviden, Eviden, [online] https://eviden.com/insights/blogs/wormgpt-when-genai-also-serves-malicious-actors/.

- Hashemi-Pour, Cameron (2024): What is AgentGPT?, Search Enterprise AI, [online] https://www.techtarget.com/searchenterpriseai/definition/AgentGPT.

- Introducing BabyAGI: the AI-Powered Task Management System – Kanaries (2025): [online] https://docs.kanaries.net/articles/babyagi-chatgpt.

- Lee, Gabby (2025): FraudGPT, WormGPT, and Dark AI Models Fuel Surge in Cybercrime - Blog, in: Daily Security Review, 26.08.2025, [online] https://dailysecurityreview.com/blog/fraudgpt-wormgpt-and-dark-ai-models-fuel-surge-in-cybercrime/.

- Opswat (2025): AI hacking - How hackers use artifical intelligence in cyberattacks, OPSWAT, [online] https://www.opswat.com/blog/ai-hacking-how-hackers-use-artificial-intelligence-in-cyberattacks.

- Rasmuson, Jason (2025): WormGPT Reborn: Cybercriminals Hijack Mainstream AI to Power New Variants - Security Buzz, Security Buzz, [online] https://securitybuzz.com/cybersecurity-news/wormgpt-reborn-cybercriminals-hijack-mainstream-ai-to-power-new-variants/.

- Shakya, Abhishek (2025): AutoGPT vs AgentGPT: A Complete Guide to Autonomous AI Agents (2025), DEV Community, [online] https://dev.to/abhishekshakya/autogpt-vs-agentgpt-a-complete-guide-to-autonomous-ai-agents-2025-1kfk.

- The Rise of Autonomous Agents: AutoGPT, AgentGPT, and BabyAGI (o. D.): BairesDev, [online] https://www.bairesdev.com/blog/the-rise-of-autonomous-agents-autogpt-agentgpt-and-babyagi/.

- Unit (2025): The Dual-Use Dilemma of AI: Malicious LLMs, Unit 42, [online] https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/. WormGPT: how hackers are exploiting the dark side of AI | Eftsure US (2026): Eftsure, [online] https://www.eftsure.com/blog/cyber-crime/wormgpt/.

- A Survey on Autonomy-Induced Security Risks in Large Model-Based Agents (o. D.): [online] https://arxiv.org/html/2506.23844v1.

- AutoGPT vs BabyAGI | Avichala GenAI Insights & Blog (2025): [online] https://www.avichala.com/blog/autogpt-vs-babyagi.

- Forewarned is Forearmed: A Survey on Large Language Model-based Agents in Autonomous Cyberattacks (o. D.): [online] https://arxiv.org/html/2505.

- 12786v2. Gan, Yuyou/Yong Yang/Zhe Ma/Ping He/Rui Zeng/Yiming Wang/Qingming Li/Chunyi Zhou/Songze Li/Ting Wang/Yunjun Gao/Yingcai Wu/Shouling Ji (2024): Navigating the Risks: A Survey of Security, Privacy, and Ethics Threats in LLM-Based Agents, in: alphaXiv, [online] https://www.alphaxiv.org/overview/2411.09523v1.

- Jada, Irshaad/Thembekile O. Mayayise (2023): The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review, in: Data And Information Management, Bd. 8, Nr. 2, S. 100063, [online] doi:10.1016/j.dim.2023.100063.

- Security, Eye (2025): Uses of AI in Enterprise Cybersecurity: Risks, Opportunities, Strategies, Eye Security, [online] https://www.eye.security/blog/uses-of-ai-in-enterprise-cybersecurity-risks-opportunities-strategies.

- The Agentic AI Security Scoping Matrix: A framework for securing autonomous AI systems | Amazon Web Services (2025): Amazon Web Services, [online] https://aws.amazon.com/de/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems/.

- [x]cube LABS/[x]cube LABS (2025): Agentic AI Explained: Autonomous Agents & Self-Driven Processes, [X]Cube LABS, [online] https://www.xcubelabs.com/blog/agentic-ai-explained-autonomous-agents-self-driven-processes/.

- Xu, Minrui/Jiani Fan/Xinyu Huang/Conghao Zhou/Jiawen Kang/Dusit Niyato/Shiwen Mao/Zhu Han/Xuemin/Shen/Kwok-Yan Lam (2025): Forewarned is Forearmed: A Survey on Large Language Model-based Agents in Autonomous Cyberattacks, arXiv.org, [online] https://www.arxiv.org/abs/2505.12786.

- Verizon (2024): 2024 Data Breach Investigations Report (DBIR). Verizon Enterprise. Verfügbar unter: [online] https://www.verizon.com/business/resources/reports/2024-dbir-data-breach-investigations-report.pdf.

- AI Act (2026): Shaping Europe's Digital Future, [online] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#1720699867912-2.

- Art. 22 GDPR – Automated individual decision-making, including profiling - General Data Protection Regulation (GDPR) (2018): General Data Protection Regulation (GDPR), [online] https://gdpr-info.eu/art-22-gdpr/.

- Article 6: Classification Rules for High-Risk AI Systems | EU Artificial Intelligence Act (o. D.): [online] https://artificialintelligenceact.eu/article/6/.

- Bowers, John/John Bowers (2025): Verizon Data Breach Investigations Report: Social Engineering Insights, Dmarcian, [online] https://dmarcian.com/verizon-2024-dbir/.

- BSI-Lagebericht 2024: Dringender Handlungsbedarf in der IT-Sicherheit | Bitkom Akademie (o. D.): [online] https://bitkom-akademie.de/news/bsi-lagebericht-2024-dringender-handlungsbedarf-der-it-sicherheit.

- Deng, Gelei/Yi Liu/Víctor Mayoral-Vilches/Peng Liu/Yuekang Li/Yuan Xu/Tianwei Zhang/Yang Liu/Martin Pinzger/Stefan Rass (2024a): {PentestGPT}: Evaluating and Harnessing Large Language Models for Automated Penetration Testing, [online] https://www.usenix.org/conference/usenixsecurity24/presentation/deng.

- Deng, Gelei/Yi Liu/Nanyang Technological University/Alias Robotics/Alpen-Adria-Universität Klagenfurt/Peng Liu/Institute for Infocomm Research (I2R), A*STAR, Singapore/Yuekang Li/University of New South Wales/Yuan Xu/Tianwei Zhang/Yang Liu (2024b): PENTESTGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing, in: USENIX Security '24 Artifact Appendix, [online] https://www.usenix.org/system/files/usenixsecurity24-appendix-deng.pdf.

- Die Lage der IT-Sicherheit in Deutschland 2024 (o. D.): Bundesamt für Sicherheit in der Informationstechnik, [online] https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2024.html.

- Fang, Richard/Rohan Bindu/Akul Gupta/Daniel Kang (2024a): LLM Agents can Autonomously Exploit One-day Vulnerabilities, arXiv.org, [online] https://arxiv.org/abs/2404.08144.

- Fang, Richard/Rohan Bindu/Akul Gupta/Qiusi Zhan/Daniel Kang (2024b): LLM Agents can Autonomously Hack Websites, arXiv.org, [online] https://arxiv.org/abs/2402.06664.

- Federation of European Risk Management Associations – FERMA (2025): EU AI Act introduces comprehensive risk management requirements for 'high-risk' AI systems, according to latest FERMA Policy Note - Federation of European Risk Management Associations – FERMA, Federation Of European Risk Management Associations – FERMA, [online] https://ferma.eu/publications/eu-policy-note-ai-act-2024/.

- Fraunhofer IKS/L. Heidemann/B. Herd/J. Kelly/N. Mata/W. Tsai/S. Zafar/A. Zamanian (2024): The European Artificial Intelligence Act. Gehrmann, Mareike Christine/Anne Förster (2024): AI Act: high-risk AI systems – what applies, what is due when?, in: Taylor Wessing, 12.11.2024, [online] https://www.taylorwessing.com/fr/insights-and-events/insights/2024/11/high-risk-ai-systems.

- Happe, Andreas/Jürgen Cito (2023): Getting pwn'd by AI: Penetration Testing with Large Language Models, in: Getting Pwn'd By AI: Penetration Testing With Large Language Models, S. 2082–2086, [online] doi:10.1145/3611643.3613083.

- Herbosch, Maarten (2025): Liability for AI Agents, in: SSRN Electronic Journal, [online] doi:10.2139/ssrn.5236649. High-level summary of the AI Act | EU Artificial Intelligence Act (o. D.): [online] https://artificialintelligenceact.eu/high-level-summary/.

- Intelligence, Microsoft Threat (2025): AI vs. AI: Detecting an AI-obfuscated phishing campaign, Microsoft Security Blog, [online] https://www.microsoft.com/en-us/security/blog/2025/09/24/ai-vs-ai-detecting-an-ai-obfuscated-phishing-campaign/?utm_source=chatgpt.com&fbclid=PAb21jcAN8qm5leHRuA2FlbQIxMQBzcnRjBmFwcF9pZA81NjcwNjczNDMzNTI0MjcAAaen20_uH3-wTEONZuZOrGvtuZXloW4lXaaDsOyHa1SEHMvnCwM2nHUSqDEinA_aem_bx4uufg74M8Qr5uDk9C87w.

- Internet Organised Crime Threat Assessment (IOCTA) 2024 - CYBIL Portal (2024): Cybil Portal, [online] https://cybilportal.org/publications/internet-organised-crime-threat-assessment-iocta-2024/.

- IOCTA 2024 (o. D.): [online] https://eucrim.eu/news/iocta-2024/. Johanna_Cybersicher/Johanna_Cybersicher (2025): BSI-Lagebericht 2024: Resilienz stärken, Cyberrisiko minimieren, CYBERsicher, [online] https://transferstelle-cybersicherheit.de/bsi-lagebericht-2024-resilienz-staerken-cyberrisiko-minimieren/.

- Kröger, Janina (2024): BSI-Lagebericht 2024: Bedrohungslage auf neuem Rekordhoch, IT-SERVICE.NETWORK Blog, [online] https://it-service.network/blog/2024/11/13/bsi-lagebericht-2024/.

- Lizaveta, Artsyman (2025a): Autogen vs LangChain vs CrewAI: Our AI Engineers' Ultimate Comparison Guide, *Instinctools, [online] https://www.instinctools.com/blog/autogen-vs-langchain-vs-crewai/.

- Lizaveta, Artsyman (2025b): Autogen vs LangChain vs CrewAI: Our AI Engineers' Ultimate Comparison Guide, *Instinctools, [online] https://www.instinctools.com/blog/autogen-vs-langchain-vs-crewai/.

- Ogbara, Bola (2024): Europol's 2024 IOCTA report, Europol's 2024 IOCTA Report, [online] https://www.digitalassetredemption.com/blog/europols-2024-iocta-report.

- Opswat (2025): AI hacking - How hackers use artifical intelligence in cyberattacks, OPSWAT, [online] https://www.opswat.com/blog/ai-hacking-how-hackers-use-artificial-intelligence-in-cyberattacks.

- PENTESTGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing (2024): Proceedings Of The 33rd USENIX Security Symposium, [online] https://www.usenix.org/system/files/usenixsecurity24-deng.pdf?utm_source=chatgpt.com.

- Ribeiro, Anna (2024): IOCTA 2024 report: Law enforcement deals major blows against EU cybercrime, disrupt ransomware networks, Industrial Cyber, [online] https://industrialcyber.co/reports/iocta-2024-report-law-enforcement-deals-major-blows-against-eu-cybercrime-disrupt-ransomware-networks/.

- SolarWinds supply chain attack | Fortinet (o. D.): Fortinet, [online] https://www.fortinet.com/resources/cyberglossary/solarwinds-cyber-attack.

- Technologies, Smiaansh (o. D.): LangChain Agents vs AutoGen vs CrewAI: The Ultimate Framewor, Smiaansh Technologies, [online] https://www.smiansh.com/blogs/langchain-agents-vs-autogen-vs-crewai-comparison/.

- Technologies, Smiaansh (o. D.): LangChain Agents vs AutoGen vs CrewAI: The Ultimate Framewor, Smiaansh Technologies, [online] - https://www.smiansh.com/blogs/langchain-agents-vs-autogen-vs-crewai-comparison/.

- Unit (2025): The Dual-Use Dilemma of AI: Malicious LLMs, Unit 42, [online] https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/.

- View of THE LEGAL STATUS OF AUTONOMOUS CYBER AGENTS: RETHINKING CRIMINAL LIABILITY IN THE AGE OF AI. (o. D.): [online] https://acadrespub.com/index.php/oijl/article/view/153/126.

- WormGPT: how hackers are exploiting the dark side of AI | Eftsure US (2026a): Eftsure, [online] https://www.eftsure.com/blog/cyber-crime/wormgpt/.

- Yigit, Yagmur/William J Buchanan/Madjid G Tehrani/Leandros Maglaras (2024): Review of Generative AI Methods in Cybersecurity, in: Review Of Generative AI Methods in Cybersecurity, [online] https://storage.prod.researchhub.com/uploads/papers/2024/04/24/2403.08701.pdf.

- Demonstration von Many-shot-jailbreaking 2025-01-06 16-13-56.mkv (o. D.): Google Docs, [online] https://drive.google.com/file/d/1qLLnXYosLna_UHfe1iIgF3fRWMiWxCKP/view?usp=sharing.

Grafiken:

- Abbildung 1, Was ist ein Agent? https://www.telekom-mms.com/fileadmin/mms/06_blog/2025/autonome-ki-agenten/was-ist-ein-agent.png.

- Abbildung 2, Was ist besonders wichtig, wenn KI-Agenten Aufgabe übernehmen? https://www.telekom-mms.com/fileadmin/mms/06_blog/2025/autonome-ki-agenten/was-ist-wichtig-wenn-ki-agenten-aufgaben-uebernehmen.png.

- Abbildung 3, Potenzielle Anwendungsbereiche in Unternehmen https://www.telekom-mms.com/fileadmin/mms/06_blog/2025/ki-agent-definition-und-relevanz-wie-ki-agenten-ihr-business-voranbringen/14250_potenzial-einsatzbereiche-von-ki-agenten-in-unternehmen.jpeg.

- Abbildung 4, Generative AI & Agents - Herausforderungen https://www.telekom-mms.com/fileadmin/mms/06_blog/2025/autonome-ki-agenten/herausforderungen-generative-ai-agents.png.

- Abbildung 5, Artificial Intelligence - Machine Learning Konan Jean-Claude, Kouassi. (2022). A Comprehensive Overview of Artificial Intelligence. 173-194. 10.5121/csit.2022.122314.