Corporate Cyber Defense

IT-Sicherheit im Wandel: Die Rolle von Penetrationstests in einer zunehmend von KI geprägten Bedrohungslage

Vorgelegt von

Dominik Deschner

Inhaltsverzeichnis

Abkürzungsverzeichnis

APT
Advanced Persistant Threat.....14, 16, 17
BSI
Bundesamt für Sicherheit in der Informationstechnik.....15, 16, 17, 28, 29, 30
ENISA
Agentur der Europäischen Union für Cybersicherheit.....15, 16, 17, 28, 29, 30
GenAI
Generative künstliche Intelligenz 29
KI
Künstliche Intelligenz 1, 6, 7, 8, 9, 12, 19, 20, 21, 28, 29, 30, 31, 33, 34
LLMs
Large Language Models 8, 9, 10, 11, 12, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28
OWASP
Open Web Application Security Project 22

Abbildungsverzeichnis

1 Einleitung

1.1 Motivation und Zielsetzung

Stellen Sie sich vor: Im Jahr 2024 erlebt die Bundesrepublik Deutschland eine beispiellose Angriffswelle. Es beginnt mit einer koordinierten Cyberattacke auf kleine und mittelgroße Unternehmen im ganzen Land. Hacker infiltrieren ihre Netzwerke, stehlen sensible Daten und legen wichtige Systeme lahm. Produktionslinien stehen still, Lieferketten brechen zusammen, und Unternehmen geraten in den finanziellen Abgrund. Die Wirtschaft taumelt. Während die Unternehmen noch mit den Folgen der Cyberattacke kämpfen, tritt eine neue Bedrohung auf den Plan. Unter dem Deckmantel des Chaos, das die Wirtschaft bereits erfasst hat, operieren russische Kommandokräfte im Verborgenen. Sie sind bestens ausgebildet, agieren unabhängig und haben klare Ziele: kritische Infrastrukturen sabotieren. Während die Bevölkerung im Dunkeln sitzt, wird die öffentliche Meinung manipuliert. Russische Hacker verbreiten Deepfakes – täuschend echte Videos und Audiodateien –, die politische Führer, Experten und Journalisten zu Wort kommen lassen. Die Botschaften sind subtil, aber wirkungsvoll: Zweifel an der Regierungsführung, Misstrauen gegenüber den Medien und Spaltung in der Gesellschaft. Die Menschen sind verunsichert, und die sozialen Medien sind ein Schlachtfeld der Desinformation. Die Bundesrepublik Deutschland steht am Abgrund. Die Wirtschaft liegt in Trümmern, die Energieversorgung ist zusammengebrochen, und die öffentliche Meinung ist gespalten. Eine Geschichte aus dem Tom Clancy Universum, oder doch nur eines von vielen Szenarien, in dem Künstliche Intelligenz eine Schlüsselrolle für einen erfolgreichen Angriff auf unsere nationale Sicherheit spielt?

Die Sicherheit von Computersystemen und Netzwerken und damit inhärent verknüpft der Schutz von Daten, digitalen Vermögensgegenständen und Prozessen stellt nicht nur Privatpersonen sondern ebenso Unternehmen und ganze Nationen vor Herausforderungen von fundamentaler Tragweite. Die fortschreitende Digitalisierung in allen Bereichen unserer Gesellschaft führt zu vermehrter Präsenz von cyberphysischen Systemen, also vernetzten oder zumindest computerisierten Objekten und Anlagen, welche nicht nur zu Produktivitätssteigerungen in den Volkswirtschaften führen, sondern auch Bedrohungen aus dem Cyberraum in die reale Welt projizieren. Die Bedrohungslage im Cyberspace ist geprägt von unterschiedlichen, bösartigen Akteuren, welche verschiedene Ziele verfolgen. Ihnen ist jedoch gemein, dass sie versuchen Computersysteme, Netzwerke sowie Menschen über unterschiedliche Angriffsvektoren zu kompromittieren und für ihre Ziele zu instrumentalisieren. Hierfür stehen diesen verschiedene technische und

nicht-technische Werkzeuge und Methoden zur Verfügung, welche maßgeblich für den Erfolg eines Angreifers sind.

Planung und Aufbau einer effektiven Verteidigung profitieren stark davon, wenn die Fähigkeiten und Taktik der Gegner bekannt sind, da so Gegenmaßnahmen ergriffen werden können und deren Widerstandsfähigkeit anhand realistischer Szenarien verifiziert werden kann. Der rasante Fortschritt in der Entwicklung von leistungsfähigen KI-Systemen und deren bereite Verfügbarkeit, insbesondere im Bereich der generativen KI birgt nicht nur enorme Potenziale für unsere Gesellschaft, sondern kann auch von bösartigen Akteuren im Cyberspace missbraucht werden. Entsprechend ist es notwendig die Cyberbedrohungslage zu im Angesicht des KI-Zeitalters neu zu bewerten.

In dieser Arbeit werden die Einsatzmöglichkeiten für offensive wie auch defensive Zwecke und der daraus resultierende Einfluss von generativer KI, insbesondere von Large Language Models, auf die Bedrohungslage im Cyberspace eruiert. Hierfür wird folgende Forschungsfrage gestellt:

Wie wirkt sich die Existenz und breite, öffentliche Verfügbarkeit leistungsstarker, generativer KI-Modelle auf das Kräftegleichgewicht zwischen Angreifer und Verteidiger im Cyberspace aus und wie können Penetrationstests zur nachhaltigen Absicherung der digitalen Infrastruktur eingesetzt werden?

Diese Fragefragestellung wird anhand der nachfolgend aufgelisteten Hypothesen überprüft:

1. Generative KI ermöglicht es komplexe Cyberangriffe autonom zu orchestrieren und mit hoher Erfolgsquote auszuführen.
2. Insbesondere kleinere und mittelgroße Unternehmen sind besonders angreifbar und generative KI macht Angriffe auf ebendiese wirtschaftlich attraktiv.
3. Generative KI versetzt Bedrohungsakteure in die Lage den demokratische Entscheidungsfindungsprozesse zu sabotieren und somit die Grundfeste einer jeden Demokratie zu destabilisieren.

Ziel der Arbeit ist unter Berücksichtigung des aktuellsten Stand der Technik im dynamischen Forschungsgebiet der Künstlichen Intelligenz eine realistische Einschätzung über mögliche Verwendungszwecke von GenAI im Bereich der Cybersicherheit abzugeben. Um dieser Dynamik gerecht zu werden wird eine narrative Literaturanalyse anhand neuester Erkenntnisse aus Forschung und Wirtschaft durchgeführt.

1.2 Struktur der Arbeit

Zu Beginn dieser Arbeit werden zunächst die relevanten Grundlagen bezüglich LLMs und Cybersicherheit beleuchtet. Darauf aufbauend erfolgt eine detaillierte Analyse der aktuellen Bedrohungsszenarien im Cyberraum. Dabei werden sowohl die Aktivitäten als auch die Intentionen der Bedrohungsakteure anhand aussagekräftiger Beispiele erörtert. Zudem wird der gegenwärtige Stand der Cyberabwehrmaßnahmen in Deutschland eingehend beschrieben. Im vierten Kapitel wird untersucht, in welchen Bereichen der Cybersicherheit LLMs sowohl offensiv als auch defensiv eingesetzt werden können. Die daraus resultierenden Erkenntnisse werden anschließend kritisch diskutiert.

In den abschließenden Kapiteln stehen die erforderlichen Strategien und Herausforderungen im Fokus, die für die Erreichung digitaler Souveränität in Post-KI-Zeitalter von Bedeutung sind. Im Schlussteil der Arbeit erfolgt eine kritische Reflexion der erarbeiteten Inhalte sowie ein Ausblick auf zukünftige Entwicklungen.

1.3 Abgrenzung

Diese Arbeit befasst sich mit den Bedrohungen und den schadhaften Anwendungsmöglichkeiten von generativen KI-Modellen. KI-Modelle und auf diesen aufbauende Dienste sind als Softwaresysteme selbst wiederrum über unterschiedliche Vektoren angreifbar und machen auf diese Bedrohung abgestimmte Sicherungsmaßnahmen notwendig. Dieser Aspekt steht vordergründig nicht im Fokus dieser Arbeit um den Umfang nicht überzustrapazieren, gleichwohl hieraus in der Zukunft wesentliche Bedrohungen des KI-Zeitalters resultieren können.

Diese Arbeit bezieht explizit auf die Auswirkungen von Künstlicher Intelligenz auf Cybersicherheitslage der Bundesrepublik Deutschland. Die Erkenntnisse dieser Arbeit lassen sich wahrscheinlich auf andere Industrienationen übertragen, jedoch ist eine Diskussion selbiger unter Berücksichtigung lokaler Gegebenheiten sinnvoll. Diese Arbeit betrachtet die Auswirkungen und Bedrohungspotenziale auf branchenagnostisch auf nationaler Ebene. Entsprechend bedarf es im nächsten Schritt einer detaillierteren Überprüfung einzelner Sektoren, um eine belastbare Bedrohungs-Taxonomie aufbauen zu können.

2 Grundlagen

In diesem Kapitel werden die für das Verständnis dieser Arbeit essenziellen fachlichen Grundlagen erläutert. Diese Arbeit richtet sich an Personen mit fundiertem Fachwissen und technischem Verständnis im Bereich Computersicherheit. Der avisierte Umfang erlaubt keine eine vollumfängliche Diskussion und Heranführung an das Thema. Dieses Kapitel beschränkt sich entsprechend auf die Erfassung relevanter Entwicklungen im Bereich der Künstlichen Intelligenz und den vom Autor gewählten Blickwinkel auf die Cybersicherheit.

2.1 Künstliche Intelligenz

2.1.1 Large Language Models

Large Language Model beschreibt eine Kategorie von KI-Modellen aus dem Bereich des Natural Language Processings, also Modelle die darauf spezialisiert sind natürliche Sprachen wie z.B. Deutsch oder Englisch zu verarbeiten. Im Vergleich zu klassischen Methoden zur Textverarbeitung wie Klassifizierung oder Mustererkennung, sind LLMs in der Lage komplexe Zusammenhänge oder mehrstufige Anweisungen zu verstehen und kontextbezogene, kohärente Antworten zu generieren, wie z.B. das Generieren einer vollständigen E-Mail anhand einiger Stichworte oder das Zusammenfassen von langen, komplizierten Texten.

Diese Modelle sind als tiefe, neuronale Netze strukturiert, welche sowohl eine große Anzahl an Neuronen aufweisen als auch mit enorm großen Datenmengen trainiert wurden[1]. So wurde ChatGPT 3.0 mit rund 570 Gigabyte an Dokumenten, welche aus einem 45 Terrabyte großen Datensatz kuratiert wurde, trainiert und besteht aus 175 Milliarden Neuronen[2]. Während des Trainings erlangen LLMs sowohl ihre Fähigkeiten Eingaben in natürlicher Sprache zu verarbeiten und ebensolche Ausgaben zu generieren, wie auch eine Wissensbasis aufbauend auf dem Trainingsmaterial. Somit hängt die Qualität von LLMs stark von der Qualität der verwendeten Trainingsdaten ab.

Über sog. Fine Tuning können vortrainierten Modellen zusätzliche Informationen und Kontext bereitgestellt werden, wie z.B. Dokumente oder Wissensdatenbanken, auf welche diese ohne weiteres energie- und rechenleistungsintensives Training,

[1] Raschka, 2024, S. 7–9.
[2] Raschka, 2024, S. 17–19.

zugreifen können[3]. Mithilfe des Fine Tunings können vortrainierte Modelle auf spezifische Aufgaben zugeschnitten werden und

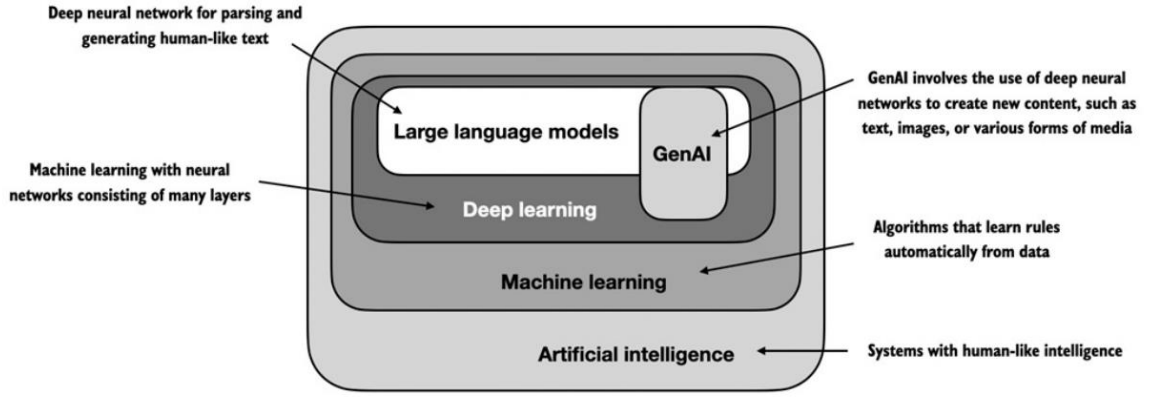

Abbildung 1: Hierarchischer Zusammenhang Begriffe aus dem Bereich der Künstlichen Intelligenz[4]

Da LLMs basierend auf einer Benutzereingabe entsprechende Antworten generieren, werden diese auch dem Bereich der generativen KI zugeordnet, welcher jedoch nicht nur auf die Generierung von Text beschränkt ist, sondern alle Modelle beinhaltet, welche mithilfe neuronaler Netze neue Inhalte – wie z.B. Bilder, Videos oder Layouts - erstellen. In Abbildung 1 ist dieser Zusammenhang visualisiert und es wird erkenntlich, dass LLMs eine sehr spezialisierte Ausprägung von Machine Learning Modellen darstellen.

Dadurch werden diese in die Lage versetzt in unterschiedlichen Disziplinen auf komplexe Aufgabenstellungen Antworten zu finden. In unterschiedlichen Testszenarien zeigt sich die aktuelle Generation der ChatGPT-Familie auf einen äußert hohen Niveau. In verschiedenen Wissenschaftsdisziplinen[5] oder bei der Generierung von Programmcode[6] können LLMs Ergebnisse erzielen, die mit Menschen vergleichbar sind[7] .

2.1.2 Multiagenten-System

Die im vorherigen Abschnitt beschriebenen Large Language Models erzielen beeindruckende Ergebnisse in der Bearbeitung anspruchsvoller Aufgaben. Jedoch sind diese nicht in der Lage bisher getroffene Antworten selbst zu reflektieren und von einem falschen Ergebnis abzukommen.[8] Ebenso fällt es ihnen schwer komplett neue Lösungsansätze in Erwägung zu ziehen, wenn bereits eine konkrete Lösung

[3] Alto, 2023, S. 29–30.
[4] Raschka, 2024, S. 8.
[5] Mao, Chen, Zhang, Guerin & Cambria, 2023, S. 7–8.Arora, Singh & Mausam, 2023, S. 9.
[6] Vaithilingam, Zhang & Glassman, 2022, S. 5–6.
[7] OpenAI et al., 2023.
[8] Das, A., Chen, Shyu & Sadiq, 2023, S. 92.

in der Historie der Abfrage enthalten ist, was die Leistung negativ beeinträchtigt.[9] Insbesondere die kontextbezogene Bearbeitung von Problemstellungen aus der echten Welt, die einen iterativen Lösungsansatz erforderlich machen, lassen sich nicht durch einfache Frage-Antwort-Systeme lösen[10].

Multiagenten-Systeme stellen einen Lösungsansatz für die zuvor beschriebene Problemstellung dar. So wird die Bearbeitung einer komplexen Fragestellung mehreren LLMs zugewiesen, welche die Aufgabe kooperativ und iterativ lösen. Jedes LLM steht in dieser Architektur für einen Agenten, welcher einen spezifischen Aufgabenbereich in der Problemdomäne abdecken kann. Einzelne Modelle können mittels Fine Tuning für einen konkreten Aufgabenbereich zusätzlich trainiert werden, um in diesem besonders gute Leistung zu erbringen[11].

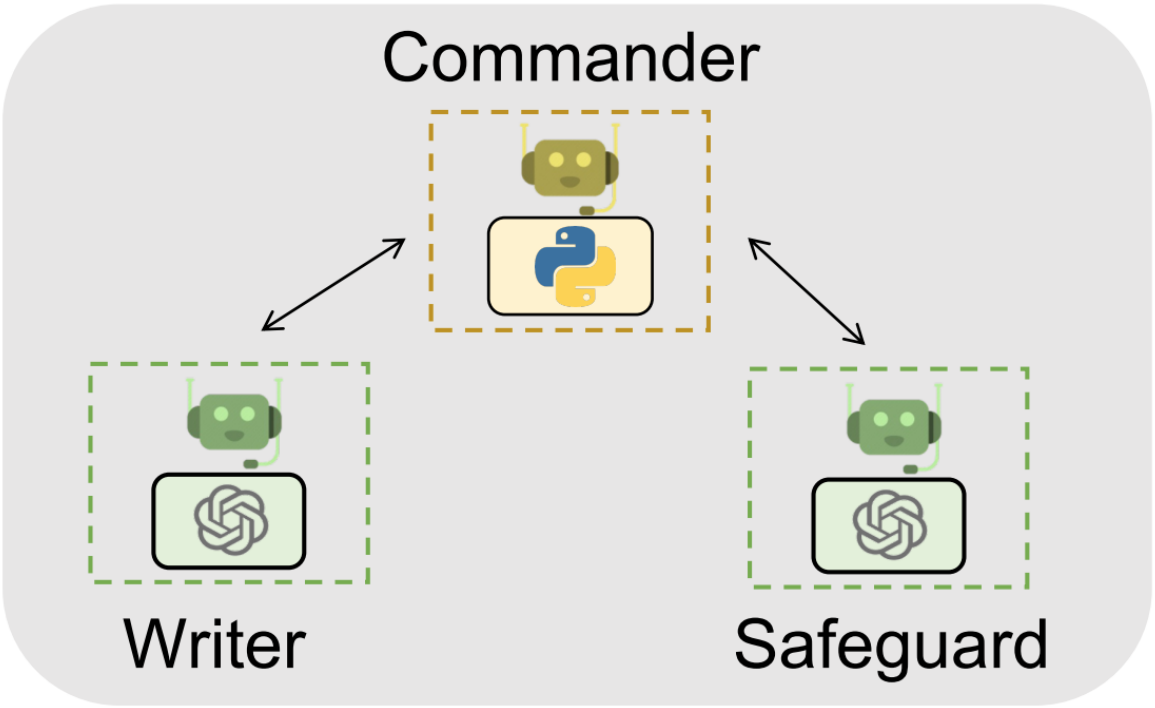

Abbildung 2: Multiagent-Architektur zur Erstellung und Ausführung von Programmcode[12]

Abbildung 2 zeig ein einfaches Multiagenten-System, dessen Aufgabe darin besteht Programmcode zu Aufgaben zu generieren, die von einem Benutzer angefragt wurden. Diese Aufgabenstellung wird von dem Commander-Agent empfangen und an den Writer-Agent mit der Aufforderung eine adäquate Lösung hierfür zu generieren weitergegeben. Der so erstellte Code wird von dem Commander an einen Safeguard-Agent gesendet, welcher prüft ob der Code ausführbar ist und

[9] Liang et al., 2023, S. 1–2.
[10] Das, A. et al., 2023, S. 92.
[11] Wu, Q. et al., 2023, S. 3.
[12] Wu, Q. et al., 2023, S. 6.

sorge dafür trägt, dass diese keine gefährlichen Segmente enthält. Wenn der Safeguard feststellt, dass der Code fehlerbehaftet ist, wird der Writer-Agent aufgefordert die Fehler zu beheben und eine neue Version des Codes zu erstellen. Dieser Vorgang wird so lange wiederholt bis dem Commander ein ausführbare und korrekter Code vorliegt, welcher die gestellte Aufgabe lösen kann. Die Multi-Agent Konfiguration schneidet in der zuvor beschrieben Aufgabe bis zu 35% besser ab als ein einzelnes LLM[13].

2.1.3 State of the Art

Die vorherigen Abschnitte im Kapitel Large Language Models geben einen sehr kondensierten, generalisierten Überblick über ein sehr großes, dynamisches Forschungsgebiet. Einen wesentlichen Aspekt, der bisher nicht hinreichend erfasst wurde stellt der rasante, technologische Fortschritt in verschiedenen Bereichen der Large Language Models dar.

Ende 2022 wurden Large Language Models mit der Veröffentlichung von ChatGPT[14] innerhalb kürzester Zeit zum Hype-Thema[15], obwohl diese Generation lediglich in der Lage war während des Trainings erlangtes Wissen widerzugeben. Im Herbst 2023 wurde ChatGPT durch ein Update in die Lage versetzt auf das Internet zuzugreifen und auf tagesaktuelle Informationen zuzugreifen[16]. Seit der Veröffentlichung von ChatGPT ist eine Vielzahl an konkurrierenden Modellen von unterschiedlichen Herstellern veröffentlicht worden. Ein Teil dieser neuen Modelle wurden vollständig der Öffentlichkeit zugänglich gemacht und können auch außerhalb von Cloud-Rechenzentren bereitgestellt und verwendet werden[17].

Nicht nur die verfügbaren Modelle und darauf aufbauenden Dienste, wie z.B. ChatGPT, Gemini, Dall-E 2 oder kontinuierlich immer leistungsfähiger sondern auch, wie in Tabelle 1 dargestellt, die Anzahl verfügbarer die Programmierschnittstellen und Frameworks hat stark zugenommen und erlauben es mit wenig Expertise LLMs in Software zu integrieren und auf spezielle Anwendungsfälle zuzuschneiden.

[13] Wu, Q. et al., 2023, S. 7–8.
[14] OpenAI, 2024b.
[15] Hu, K., 2023.
[16] Das, A., 2023.
[17] Göbel, 2024; Meta, 2024.

Projektname	Anwendungsgebiet	Referenz
Ollama[18]	Verwaltung, Versionierung und lokale Ausführung & Bereitstellung einer Programmierschnittstelle zu Open Source Large Language Models wie Llama oder Phi-3.	Ollama
LangChain[19]	Framework zur Integration und Anpassung/Fine Tuning von LLMs in Softwareprojekten, Orchestrierung von Multi-Agenten Workflows, Integration von eigenem Code in LLMs.	LangChain
SemanticKernel[20]	Framework für Integration und Erweiterung von LLMs in eigene Software. Entwicklung von Copiloten, die komplexe Aufgaben erledigen. Technisches Fundament für die Copilot-Funktionen in Microsoft-Produkten	microsoft/semantic-kernel: Integrate
Autogen[21]	Framework zur Entwicklung und Orchestrierung komplexer Multi-Agenten-Systeme.	microsoft/autogen

Tabelle 1: Übersicht Entwicklungswerkzeuge LLMs

Mit den mittlerweile öffentlich zur Verfügung stehenden Werkzeugen und Frameworks ist es möglich ohne tiefgreifende Vorkenntnisse und technisches Verständnis von KI-Anwendungen die Fähigkeiten von LLMs in jede Art von Software zu integrieren. Ebenfalls kann auf eine große Auswahl an unterschiedlich leistungsfähigen Open Source LLMs zurückgegriffen werden, die auch auf Consumer-Hardware ausreichend schnell ausgeführt werden können. Damit ist die Entwicklung von LLM-getriebener Software weitgehend unabhängig von zentralen Diensten wie z.B. OpenAI oder Azure. Es lässt sich resümieren, seit der Veröffentlichung von ChatGPT als proprietärem Dienst hat eine weitreichende technische Demokratisierung von LLMs stattgefunden. Die Verwendung von LLMs und deren Integration in eigene Software war noch nie so einfach wie heute.

[18] Morgan, 2024.
[19] LangChain, 2024.
[20] Microsoft, 2024a.
[21] Microsoft, 2024b.

2.2 Cybersicherheit

In diesem Abschnitt werden die für diese Arbeit relevanten Begriffe aus dem Bereich der Cybersicherheit definiert.

2.2.1 Asset

Ein Asset ist ein schützenswerter Gegenstand einer Organisation. Darunter fallen jegliche materielle und immaterielle Güter, welche für die Organisation einen Wert besitzen, bspw. geistiges Eigentum, Prozesses, Computersysteme, Mitarbeiter- und Kundendaten, Betriebsmittel oder Kapital. In Abbildung 3 werden die Assets „Product and Process Information“ und „Production Process“ von dem Threat Actor in angegriffen. [22]

2.2.2 Threat Actor

Ein Threat Actor ist eine Organisation oder Person, welche vorsätzlich und unberechtigterweise versuchen Zugriff auf Assets zu erlangen oder diesen Schaden zuzufügen. Darunter fallen unteranderen folgende Gruppen: Cyberkriminelle, Hacktivisten, Insider, Terroristen, Geheimdienste oder wirtschaftliche Konkurrenten. In Abbildung 3 wird der Thread Actor durch einen Insider repräsentiert.[23]

2.2.3 Attack Vector

Ein Angriffsvektor ist eine Methode bzw. eine Kombination von Methoden, mit Hilfe derer ein Threat Actor Zugriff auf ein Computersystem- oder Netzwerk erlangt. Der Threat Actor in Abbildung 3 verwendet „Privilege Escalation“ und „Physical Tampering“ als Angriffsvektor um das Zielsystem zu kompromittieren.[24]

2.2.4 Attack Surface

Gesamtheit aller Komponenten und Schnittstellen, die ein Threat Actor verwenden kann um in ein Computersystem- oder Netzwerk einzudringen. Im Beispiel von Abbildung 3 stellen die Systeme „Cloud Storage“ und „Human-Machine Interface“ die Attack Surface dar.[25]

---

[22] Klipper, 2015, S. 16.; Muckin & Fitch, 2019, S. 42.
[23] Crowd Strike, 2024c; Klipper, 2015, S. 3–4.; Muckin & Fitch, 2019, S. 42.
[24] Crowd Strike, 2024b.
[25] Muckin & Fitch, 2019, S. 42.

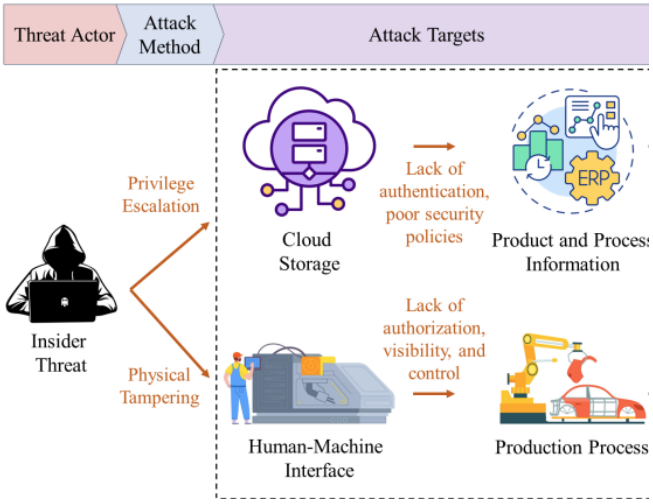

Abbildung 3: Angriff durch einen Threat Actor[26]

2.2.5 Advanced Persistant Threat

Unter Advanced Persistant Threat werden langanhaltende, zielgerichtete Angriffe verstanden, die häufig von sehr gut ausgebildeten und erfahrenen Akteuren ausgeführt werden. Unter Umständen bleiben APT Angriffe lange unerkannt und erlauben es dem Angreifer sich im Zielnetzwerk auszubreiten und Manipulationen vorzunehmen, die deren Entdeckung erschweren. Diese Angriffe verfolgen in der Regel keine wirtschaftlichen Ziele sondern dienen der Sabotage oder Informationsbeschaffung und finden ihren Ursprung in der Nähe staatlicher Organe, wie z.B. Geheimdiensten.[27]

2.2.6 Penetrationstest

Ein Penetrationstest hat das Ziel zu prüfen ob ein Computersystem- oder Netzwerk für Cyberangriffe verwundbar ist. Hierfür kommen u.a. Methoden und Werkzeuge zum Einsatz, welche von tatsächlichen Threat Actors verwendet werden. Vor der Testausführung wird der Kontext definiert in welchem der Test ausgeführt wird. Dabei wird u.a. definiert welche System- und Netzwerkbereiche getestet werden und wie weit die Tester in das System eindringen dürfen, wenn sie Schwachstellen aufdecken. Gefundene Schwachstellen werden dokumentiert. Mithilfe der Testergebnisse können gefundene Schwachstellen behoben und die Sicherheit des getesteten Systems gesteigert werden.[28]

[26] Rahman, Cassandro, Wuest & Shafae, 2023, S. 19.
[27] Bundesamt für Sicherheit in der Informationstechnik, 2023a.
[28] Bundesamt für Sicherheit in der Informationstechnik, 2020, S. 4.; National Institute of Standards and Technology [NIST], 2024.

3 Bedrohungslage

In diesem Kapitel wird die Gefahrensituation im Cyberspace detailliert dargelegt, um ein ganzheitliches Verständnis der gegenwärtigen Lage zu gewinnen. Dieses Verständnis bildet die essentielle Basis, um potenzielle Veränderungen zu erkennen und zu verstehen, die durch den Einsatz von umfangreichen Sprachmodellen hervorgerufen werden könnten. Hierfür werden einerseits die Aktivitäten und Methoden der Bedrohungsakteure, aber auch die defensiven Fähigkeiten von Unternehmen und Behörden betrachtet.

3.1 Aktivitäten der Bedrohungsakteure

Die Bedrohungslage im Cyberspace 2024 ist geprägt von den vielschichtigen Interessen unterschiedlicher Akteure, der kontinuierlich fortschreitenden Digitalisierung sowie der aus dem Ukrainekrieg resultierenden geopolitischen Spannungen zwischen westlichen Demokratien und autoritären Regimen. Das BSI resümiert in dem 2023 veröffentlichten Lagebericht zur IT-Sicherheit in Deutschland:

„Insgesamt zeigte sich im aktuellen Berichtszeitraum eine angespannte bis kritische Lage. Die Bedrohung im Cyberraum ist damit so hoch wie nie zuvor.“[29]

Zu übereinstimmenden Einschätzungen kommen die Lageberichte der Agentur der Europäischen Union für Cybersicherheit (ENISA)[30] und des Cybersicherheitsanbieters Crowdstrike[31]. So verdoppelte sich die Anzahl der von der ENISA weltweit erfassten Cybersicherheitsvorfälle von 1000 im Jahr 2022 auf 2500 im Jahr 2023. Etwa 80% der befragten Unternehmen waren nach einer Erhebung der Bitkom[32] von einem Cyberangriff betroffen.

Als Angriffsart kam am häufigsten Ransomware zum Einsatz, also Schadsoftware, welche Daten verschlüsselt oder abfließen lässt um das Opfer anschließend zu erpressen[33]. Dieser Trend wird auch vom BSI identifiziert wonach 2023 deutlich mehr Daten von deutschen Ransomware Opfern veröffentlicht wurden[34]. Im Jahr 2023 wurden vom BSI pro Tag bis zu 332 neue Varianten von Schadsoftware

[29] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 11.
[30] ENISA, 2023, S. 6.
[31] Crowd Strike, 2023a, S. 2–3.
[32] Bitkom, 2023b, S. 3.
[33] ENISA, 2023, S. 13.
[34] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 19.

entdeckt[35]. Diese bildet u.a. das Fundament für Ransomware-Angriffe, entsprechend ist davon auszugehen, dass die Bedrohung durch ebendiese hoch bleibt.

Ransomware-Angriffe werden überwiegend aus finanziellen Motiven ausgeführt[36] jedoch nicht ausschließlich, sie werden ebenfalls zu Sabotage-Zwecken oder zur Verschleierung von APT-Angriffen eingesetzt[37]. Das BSI schätzt die Akteure hinter Ransomware-Angriffen als hoch professionell und gut organisiert ein, die insofern Arbeitsteilung betreiben, dass einzelne Schritte eines Angriffs wie z.B. das Beschaffen von Zugangsdaten oder die Bereitstellung von Schadsoftware als Dienstleistung von spezialisierten Cybercrime-Gruppen zugekauft wird[38].

Neben Ransomware wurden 2023 sehr häufig Distributed Denial of Service (DDoS) Angriffe ausgeführt, welche zum Ziel haben die Erreichbarkeit von Diensten und Webanwendungen zu stören[39]. Diese wurden vermehrt von prorussischen Aktivisten im Zuge des Ukrainekriegs eingesetzt[40]. Für die Ausführung von DDoS-Attacken kommen häufig Botnetze zum Einsatz, welche aus tausenden von infizierten Geräten bestehen die von einem Threat Actor für seine Zwecke kontrolliert werden können und stellen somit ein relevantes Werkzeug für Cyberkriminelle dar. Die Generierung neuer Bots ist entsprechend essenziell und stellt ebenfalls eine relevante Bedrohung im Cyberraum dar, welche auch in 2023 erfasst wurde.[41]

Abseits von Cyberkriminellen, deren Angriffe meist nicht zielgerichtet sind, wird die Bedrohungslage im Cyberspace durch APT-Gruppen verschärft, die zielgerichtete und hochkomplexe Angriffe auf Hochwertziele, innerhalb staatlicher oder wirtschaftlicher Organisationen, ausführen. APT-Gruppen aus China, Russland, Iran und Nordkorea sind maßgeblich für die APTs verantwortlich.[42] Die von APT-Gruppen ausgehende Gefahr lässt sich anhand des kürzlich bekannt gewordenen und verhinderte Supply-Chain-Angriffs[43] auf die Linux Secure Shell, welcher es den Angreifern im Erfolgsfall ermöglicht hätte auf jeden vom Internet erreichbaren Linux-Server zuzugreifen. Die Tragweite des Angriffs und das dieser fast unentdeckt geblieben wäre zeigt mit welcher Raffinesse und Geld APT-Gruppen versuchen Backdoors in der Supply-Chain zu platzieren. Hier drängt sich die Frage auf, wie

---

[35] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 12.
[36] ENISA, 2023, S. 17.
[37] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 15.
[38] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 16–17.; ENISA, 2023, S. 33–34.
[39] ENISA, 2023, S. 13.
[40] ENISA, 2023, S. 35.
[41] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 13.
[42] Crowd Strike, 2024a, S. 1–2.
[43] Bundesamt für Sicherheit in der Informationstechnik, 2024, S. 1–2.

viele Angriffe dieser Art in der Vergangenheit unentdeckt geblieben sind. Der erfolgreiche Angriff auf den Voice-Over-IP Anbieter 3CX ist ein weiteres Beispiel für einen erfolgreichen Supply-Chain-Angriff, der 12 Millionen Anwender betraf[44]. Derartige Angriffe beschränken sich jedoch nicht nur auf Softwarekomponenten, welche von Zulieferern bezogen werden, sondern auch Dienstleistungsanbieter stehen im Ziel von APT-Gruppen, um deren Kunden anzugreifen[45].

Neben den sehr aufwändigen und langwierigen Supply-Chain-Angriffen, setzten vor allen APT-Gruppen, aber auch Cyberkriminelle, verstärkt auf Angriffe gegen Serversysteme mit direktem Internetzugriff, wie z.B. E-Mail-Server, Firewalls, oder VPN-Systeme. Bei diesen Systemen werden sowohl Zero Day als auch ältere Schwachstellen und Fehlkonfiguration ausgenutzt, um initial in die Unternehmensnetze einzudringen[46]. Außerdem werden ebenfalls aufwändigere Spear Phishing Angriffe gegen einzelne Personen in Schlüsselrollen angewandt, um langfristig in Computersysteme einzudringen oder an wertvolle Informationen zu gelangen. Dabei ist festzustellen, dass Phishing über unterschiedliche Kanäle abseits der E-Mail stattfindet wie z.B. Social Media oder Video Calls[47].

Die zuvor erörterten, vielschichtigen Ziele und die Intensität, mit der Bedrohungsakteure diese mittels diverser Methoden verfolgen, in Verbindung mit der, durch zunehmende Digitalisierung in öffentlichen Einrichtungen und Unternehmen gestiegene Angriffsfläche, deuten auf ein außerordentlich hohes Bedrohungspotenzial hin. Diese Erkenntnisse unterstreichen die Bedeutung der eingangs dargelegten Lageeinschätzung des BSI.

3.2 Stand der Cyberabwehr

Laut dem Wirtschaftsschutz-Bericht[48] der Bitkom ist der deutschen Wirtschaft durch Cyberkriminalität im Jahr 2023 ein Schaden in Höhe von 200 Mrd. Euro entstanden. Das entsprach etwa 4,8% des deutschen Bruttoinlandsproduktes. Die zuvor skizzierte Bedrohungslage ist folglich nicht nur abstrakt im Cyberspace vorhanden, sondern richtet realen Schaden an der deutschen Wirtschaft an, die als drittgrößte Volkswirtschaft ein attraktives Ziel für Cyberkriminelle und staatliche Akteure darstellt. Diese Bedrohung ist von der deutschen Wirtschaft weitestgehend erkannt worden, welche 2023 etwa 9,2 Mrd. Euro in IT-Sicherheit investierten[49]. Trotz

[44] Bundesamt für Sicherheit in der Informationstechnik, S. 1–2.
[45] ENISA, 2023, S. 127.
[46] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 26.; ENISA, 2023, S. 22–23.
[47] ENISA, 2023, S. 23.
[48] Bitkom, 2023b, S. 4.
[49] Bitkom, 2023a.

verpflichtender Implementierung von Angriffserkennungssystemen im KRITIS-Sektor stieg die Anzahl der meldepflichtigen Vorfälle von 452[50] auf 490[51] im Jahr 2023. Ein IT-Sicherheitsmanagementsystem haben 2023 immer noch 155 von 575 Betreibern kritischer Infrastruktur nicht oder nur weitestgehend umgesetzt[52].

Kleine und mittelgroße Unternehmen, mit 1 bis 249 Mitarbeitenden, sind mit den Herausforderungen in der Cybersicherheit auf allen Ebenen überfordert. Bei diesen mangelt es weiterhin an Risikobewusstsein, organisatorischen und technischen Maßnahmen wie dem regelmäßigen Einspielen von Updates. Selbst wenn Unternehmen die Notwendigkeit für Investitionen in IT-Sicherheit erkennen, finden sie häufig keinen passenden Dienstleister oder können keine Fachkräfte anwerben[53].

Obwohl die grundlegende Brisanz des Themenkomplexes Cybersicherheit bekannt ist, genügen die Anstrengungen der deutschen Wirtschaft in Summe nicht aus um der Bedrohungslage gerecht zu werden. Es mangelt sowohl an organisatorischen als auch geeigneten, technischen Maßnahmen zur Prävention von Cyberangriffen.

[50] Bundesamt für Sicherheit in der Informationstechnik, 2022, S. 69.
[51] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 62.
[52] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 63.
[53] Bundesamt für Sicherheit in der Informationstechnik, 2023b, S. 6466.

4 Wie verändert generative KI die Bedrohungslage im Cyberspace?

In diesem Kapitel wird erörtert wie der Aufstieg von generativer KI die im vorherigen Abschnitt skizzierte Bedrohungslage im Cyberspace beeinflusst. Dabei wird zunächst auf potenzielle Anwendungsfälle für Bedrohungsakteure eingegangen. Danach wird analysiert an welchen Stellen die Cyber Defense durch generative KI profitieren kann.

4.1 Threat Actors

LLMs sind vielseitig einsetzbare Werkzeuge, die durch Threat Actors in unterschiedlichen Szenarien verwendet werden können und bereits verwendet werden.

4.1.1 Wissensabfrage

Das Eindringen in Computersysteme erfordert je nach verwendetem Angriffsvektor ein tiefes technisches Verständnis in einer Vielzahl unterschiedlicher Teilgebiete der Informatik wie z.B. der Netzwerktechnik, Programmierung oder Betriebssystemen. LLMs können von Bedrohungsakteuren dazu verwendet werden, um schnell und kontextbezogen an Wissen aus den zuvor beschriebenen Bereichen zu gelangen. Als niedrigschwellige und mächtige Informationsquelle, haben LLMs das Potenzial die technischen Fähigkeiten und von Bedrohungsakteuren zu verbessern und sind allgemein deren Effizienz zuträglich.

Das breitflächige Wissen, das aus den LLMs extrahiert werden kann, ermöglicht aber nicht nur die Optimierung bereits erlernter Angriffe eines Bedrohungsakteurs. Es beschleunigt auch technologische Anpassungen hin zu neuen Angriffsmustern, in dem z.B. auf neue Programmiersprachen oder andere Werkzeuge und Frameworks für den Angriff zu wechseln. Insgesamt werden Angreifer durch Verwendung von LLMs flexibler.

Aber nicht nur bereits etablierte Bedrohungsakteure können durch die verbesserte Informationsbeschaffung profitieren. Die Einstiegshürde sinkt auch für neue Bedrohungsakteure, die den Cyberraum für ihre Aktivitäten in Zukunft erschließen möchten.

4.1.2 Generierung von Schadsoftware

Bedrohungsakteure verwenden Schadsoftware um Computersysteme zu infizieren, die Daten ihrer Opfer zu verschlüsseln oder um Zugangsdaten abzugreifen. Die

Schadsoftware muss in regelmäßigen Abständen angepasst werden, um diese mit auf neue Ziele oder Angriffsmethoden des Bedrohungsakteurs anzupassen. Auf LLMs aufbauende Copiloten erhöhen die Produktivität in der Softwareentwicklung[54]. Diese Produktivitätssteigerung macht nicht vor Bedrohungsakteuren halt und ermöglicht es diesen mehr unterschiedliche Schadsoftware schneller zu entwickeln.

Dadurch entsteht die Möglichkeit während der Verbreitung von Schadsoftware, dass diese mit Hilfe von LLMS Änderungen am eigenen Quellcode[55] vornimmt und so die vermeidet aufgespürt zu werden.

4.1.3 Phishing/Social Engineering

Der Erfolg von Phishing hängt maßgeblich von der Glaubwürdigkeit der übermittelten Phishing-Nachrichten ab. Mithilfe von Large Language Models (LLMs) lassen sich hochwertige Phishing-Nachrichten und Szenarien in verschiedenen Sprachen auf muttersprachlichem Niveau erstellen.[56] Ebenso können diese grundsätzlich zur Ideengenerierung über Inhalte und Tonalität neuer Phishing- und Spamkampagnen befragt werden.[57] Phishing-Nachrichten, die von modernen LLMs generiert wurden, haben eine ähnliche Öffnungsrate wie von Menschen erstellte Nachrichten und sind ebenso effektiv darin, Benutzer dazu zu bewegen, Links oder Anhänge zu öffnen.[58].

Auch für Spear-Phishing-Angriffe, in denen eine Person mit maßgeschneiderten Nachrichten angegriffen wird, können LLMs über die Erzeugung von Phishing-Nachrichten hinaus, bei der Onlinerecherche z.B. nach persönlichen Informationen über die Zielperson genutzt werden[59]. Mithilfe von Multiagenten-Systemen können LLMs für die massive Automatisierung[60] hochwertiger Phishing-Kampagnen genutzt werden. Hierbei könnte bspw. einem Agenten die Aufgabe zuteilen, Daten über die Zielperson in den sozialen Medien zu sammeln. Danach entscheidet ein weiterer Agent basierend auf den gefundenen Informationen, ob ein maßgeschneiderter Angriff erfolgen kann oder ob eine generische Nachricht erzeugt wird und sendet diese direkt an das Opfer ab.

Neben Text können via generativer KI Modelle ebenso Bilder, Videos und Sprachausgaben synthetisiert werden. So können Bedrohungsakteuer noch

---

[54] GitHub, 2024.
[55] HYAS Infosec Inc., 2023.
[56] Gupta, M., Akiri, Aryal, Parker & Praharaj, 2023, S. 7–8.
[57] Hazell, 2023, S. 9–10.
[58] Bethany, M. et al., 2024, S. 13–15.
[59] Hazell, 2023, S. 3–4.
[60] Falade, 2023, S. 189.

realistischere Social Engineering Szenarien erstellen und glaubhaft z.B. die Autorität von Führungspersonen ausnutzen in dem deren Stimme bei einem Anruf imitiert oder gar ein Deep-Fake-Video von Ihnen erstellt wird[61].

Die zuvor beschriebenen Fähigkeiten von LLMs lassen sich nicht nur dazu verwenden, um via Social Engineering in Computersysteme einzudringen oder sensible Informationen zu erlangen, sondern können darüber hinaus dazu verwendet werden die öffentliche Meinung einer Gesellschaft zu beeinflussen[62]. Ausländische Bedrohungsakteure wie z.B. Geheimdienste, sind in ihren Fähigkeiten und Kapazitäten fremdsprachige Inhalte für manipulative Informationskampagnen beschränkt. Mit generativer KI können einzelne Zielgruppen viel feingranularer angesprochen und effektiver beeinflusst werden. In Zusammenarbeit mit LLMs können Bedrohungsakteure:

- Grammatikalisch korrekte und inhaltlich kohärente Texte in fremden Sprachen erzeugen[63]
- Fremdsprachige Inhalte recherchieren und zusammenfassen lassen um neue
- Feingranulare Anpassungen an bestehenden Inhalten vornehmen, um diese besser auf einzelne Zielgruppen abzustimmen[64]

Daraus lässt sich schließen, dass somit die Effizienz und Effektivität von Social Engineering und Phishing-Angriffen durch Einsatz von LLMs positiv beeinflusst wird.

4.1.4 Automatisierte Hacking-Agenten

Für Cyber-Bedrohungsakteure stellt der Zugriff auf Computersysteme und Netzwerke ihrer Zielobjekte einen entscheidenden Schritt in der Ausführung vielfältiger Angriffsszenarien dar. Der initiale Zugang zu einem Zielsystem multipliziert die Erfolgsaussichten eines Angriffs signifikant, insbesondere wenn es den Eindringlingen gelingt, sich weiter in interne Systeme vorzuarbeiten, sich unauffällig im Netzwerk des Opfers zu bewegen und ihre Präsenz zu verfestigen. Diese Akteure machen sich Schwachstellen oder Konfigurationsfehler in Software und Betriebssystemen zunutze, um Kontrolle über Computerinfrastrukturen zu erlangen. Dieser Prozess ist komplex und vielschichtig, erfordert tiefgreifendes, kontextspezifisches Wissen sowie technische Expertise und ist zudem zeitintensiv.

Die neueste Generation von LLMs kann diesen Prozess unterstützen oder teilweise automatisieren, insbesondere durch den Einsatz von Multi-Agenten-Architekturen,

---

[61] Falade, 2023, S. 190–191.
[62] Chaudhary & Penn, 2024, S. 5.
[63] Fredheim & Pamment, 2024, S. 2.
[64] Fredheim & Pamment, 2024, S. 1–2.

die fähig sind, mehrstufige und komplexe Angriffsstrategien zu entwerfen und durchzuführen. Aktuelle Forschungsergebnisse[65] belegen eine stetige Verbesserung der Fähigkeiten von LLM-Agenten in diesem Bereich, die mittlerweile in der Lage sind, bekannte Sicherheitslücken basierend auf öffentlich zugänglichen Beschreibungen automatisch und erfolgreich zu exploitieren. Dadurch übertreffen diese LLM-Agenten bereits heute die Leistungsfähigkeit konventioneller Sicherheitsscanner wie beispielsweise OWASP ZAP[66]. Neben vollautomatisierten Systemen erscheinen Ansätze, wie sie PentestGPT[67] verfolgt, vielversprechend: Hier unterstützen LLMs das Hacking von Systemen, indem sie einzelne Aufgaben automatisieren, Tools steuern und Vorschläge unterbreiten, während sie zugleich auf Entscheidungen des Nutzers warten.

Die zuvor dargelegten Veröffentlichungen zu autonomen Agenten markieren einen signifikanten Fortschritt in der Cyber-Sicherheit gleichwohl diese aus dem akademischen Sektor. Diese Systeme, die ohne umfangreiche Ressourcen oder fortschrittliche Multi-Agenten-Architekturen entwickelt wurden, demonstrieren bereits zum jetzigen Zeitpunkt eine beeindruckende Leistungsfähigkeit. Dies deutet darauf hin, dass ihr Potenzial bei weiterer Forschung und Entwicklung, besonders durch Akteure deren wirtschaftliche Mittel die von kleinen Forschergruppen deutliche übertreffen, noch erheblich gesteigert werden könnte. Die kontinuierliche Evolution und Verfeinerung von Hacking-Agenten, gepaart mit der Entwicklung neuer und mächtigerer Grundlagenmodelle wie GPT-4o[68], versprechen eine signifikante Erweiterung der Möglichkeiten in diesem Bereich. Durch die steigende Verfügbarkeit von Leistungsfähigen Open Source LLMs wird es Bedrohungsakteuren in Zukunft möglich sein vergleichbare System auch offline zu Entwickeln und zu Betreiben. Dadurch können diese unabhängig ihre Abhängigkeit von Clouddiensten wie ChatGPT reduzieren und es besteht keine Möglichkeit mehr Bedrohungsakteure zentralisiert von der Nutzung dieser Technologie auszuschließen.

4.2 Cyber Defense

LLMs lassen sich nicht nur von Threat Actors einsetzten, sondern können in der Defensive nützlich sein. In diesem Kapitel werden unterschiedliche Einsatzgebiete von LLMs in der Cyber Defense diskutiert.

---

[65] Deng et al., 2023; Fang, Bindu, Gupta & Kang, 2024; Fang, Bindu, Gupta, Zhan & Kang, 2024; Xu, J. et al., 2024, S. 13.
[66] Fang et al., 2024, S. 6.
[67] Deng et al., 2023, S. 14–15.
[68] OpenAI, 2024a.

4.2.1 Security Operations

In der täglichen Wartung und Überwachung von Computersystemen und Netzwerken erweisen sich LLMs als probate Instrumente für sicherheitsrelevante Aufgaben. Sie bieten Unterstützung bei der Analyse von Skripten, dem Verständnis von Quellcodefragmenten und entwickeln Lösungsstrategien für eine sichere Systemkonfiguration. Dies ermöglicht es den Mitarbeitern, den Betrieb effizienter und zügiger zu gewährleisten, indem sie sich auf die Identifizierung von Anomalien und die Optimierung der Sicherheitsarchitektur konzentrieren können. LLMs übernehmen auch die essentielle Rolle in der Protokollanalyse, indem sie klassifizieren, ob ein spezifischer Log-Eintrag harmlos ist oder einen Angriffsversuch darstellt. Diese Fähigkeit verstärkt die analytischen Kompetenzen der Cyberabwehr und trägt dazu bei, dass Angriffe schneller und verlässlicher identifiziert werden.[69]

Im Sektor der Threat Intelligence leisten LLMs einen wertvollen Beitrag zur automatisierten Recherche verschiedener Informationsquellen, um stets die neuesten Daten bereitzustellen.[70] Sie unterstützen auch das Mapping von potenziellen Bedrohungen auf die eigene Organisation, vorausgesetzt, detaillierte Informationen über die Infrastruktur sind verfügbar. Darüber hinaus können LLMs zur Erstellung von Threat Intelligence Berichten und anderen Reportarten eingesetzt werden, um die Security Operation Teams zu unterstützen und deren Arbeitslast zu reduzieren.[71]

4.2.2 Anwendungssicherheit

Die Entwicklung sicherer Softwaresysteme stellt eine große Herausforderung in der Softwareentwicklung dar und involviert eine Reihe anspruchsvoller und zeitintensiver Tätigkeiten, wie z.B. Durchführung von Code Reviews, statischer Codeanalyse und sicherer Entwicklungsrichtlinien.

Durch LLMs können einige Aufgaben, wie z.B. Code Reviews hinsichtlich sicherheitsrelevanter Bugs oder Schwachstellen während des Entwicklungsprozesses fortlaufend umgesetzt werden. Damit unterstützen diese bei der frühzeitigen Erkennung von Schwachstellen während des Entwicklungsprozesses, so dass Schwachstellen erst gar nicht in das fertige Produkt gelangen. Darüber hinaus können LLMs unsichere Stellen im Quellcode direkt beheben[72] oder während der Entwicklung sicheren Quellcode generieren und

---

[69] Yigit, Buchanan, Tehrani & Maglaras, 2024, S. 17–18.
[70] Patel, Yeh & Gondhalekar, S. 11.
[71] Yigit et al., 2024, S. 17–18.
[72] Nong et al., 2024, S. 12–13.

so die Entstehung von unsicherem Quellcode vermeiden[73]. In beiden Fällen tragen LLMs also dazu bei, dass weniger Schwachstellen während der Entwicklung entstehen und Software somit sicherer wird.[74]

Insbesondere in Verbindung mit statischer Codeanalyse haben LLMs das Potenzial, zusätzliche Einblicke in identifizierte Schwachstellen und mögliche Behebungsstrategien zu liefern. Obwohl aktuelle LLMs in der Erkennung von Schwachstellen keine überlegenen Fähigkeiten gegenüber bestehenden Analysewerkzeugen aufweisen, besitzen sie die Fähigkeit, Quellcode in fast jeder Programmiersprache zu analysieren. Dies ermöglicht ihren Einsatz als Ergänzung zur statischen Codeanalyse in Bereichen, die von herkömmlichen Werkzeugen nicht abgedeckt werden.[75]

4.2.3 Ausbildung & Training

Die Ausbildung und Schulung von Cybersicherheitsexperten kann durch LLMs verbessert und skaliert werden. Dabei haben sich diese in anderen Fachbereichen in folgenden Bereichen der Lehre als wertvoll erwiesen:

- Interaktive Fragestellung durch Lernenden[76]
- Generierung von Übungsaufgaben[77]
- Schnelles und individuelles Feedback für die Lernenden[78]

Weiterhin wirken sich die in 4.1.1 diskutierten Aspekte der effizienten Wissensabfrage ebenfalls positiv auf die Lehre aus. Damit stellen die Chancen in der Lehre einen wertvollen Lösungsweg zur Bekämpfung des globalen Fachkräftemangels[79] in der Cyber Defense und der Forschung, was ein essenzieller Baustein in der Bekämpfung von Cyberbedrohungen darstellt.

4.2.4 Penetrationstests

Die im Abschnitt 4.1 dargelegten Anwendungsfälle von LLMs für Bedrohungsakteure lassen sich effektiv in der Cyberabwehr nutzen, um die Sicherheit zu stärken und im Rahmen von Penetrationstests sowie Red-Team-Übungen die Cybersicherheit zu bewerten. Sie ermöglichen es, komplexere und realitätsnähere Angriffsszenarien nachzustellen. Es ist davon auszugehen, dass Cyberabwehrteams dadurch nicht nur effizienter arbeiten, sondern auch eine Demokratisierung der Cyberabwehr eintritt.[80] Denn LLMs vermögen es, in

---

[73] Kim, S. Yong, Fan, Noller & Roychoudhury, 2024, S. 4–5.
[74] Zhou, Cao, Sun & Lo David, 2024, S. 7–8.
[75] Li, H. & Shan, 2023, S. 4.; Purba, Ghosh, Radford & Chu, 2023, S. 119.
[76] Krause, Panchal & Ubhe, 2024, S. 12–13.
[77] Scanlon, Breitinger, Hargreaves, Hilgert & Sheppard, 2023, S. 8.
[78] Koutcheme et al., 2024, S. 5–6.
[79] NIST, 2023.
[80] Shashwat et al., 2024, S. 6.

zahlreichen Bereichen bis zu einem gewissen Maß fachkundige Unterstützung zu bieten[81]. Dies eröffnet gerade kleinen und mittelständischen Unternehmen, die oft vor Herausforderungen bei der Suche nach qualifizierten Dienstleistern oder Mitarbeitern im Bereich der Cybersicherheit stehen, neue Möglichkeiten. Obwohl LLMs echte Experten nicht ersetzen, können sie doch IT-Personal mit den bereits diskutierten Kompetenzen ausstatten, um das grundlegende Sicherheitsniveau zu verbessern.[82]

Mithilfe der allgemeinen Effizienzsteigerung, die mit der Verwendung von LLMs einhergeht ist erwartbar, dass die Kosten pro Testszenario sinken. Entsprechend kann die Frequenz bzw. der Umfang von Penetrationstests erhöht werden und so eine höhere Testabdeckung realisiert werden.

4.3 A(i)pokalypse Now?!

In den vorhergehenden Abschnitten wurde umfassend erarbeitet auf welche Bereiche der Cybersicherheit sich die letzten Fortschritte der künstlichen Intelligenz, den Large Language Models, auswirken. Hierbei wurde das Augenmerk auf einzelne, isolierte Teilbereiche gelegt, eine ganzheitliche Betrachtung und Bewertung, ob nun die Offensive oder die Defensive gestärkt wird und welche Maßnahmen notwendig sind, um zur Stärkung der gesamtstaatlichen Resilienz beizutragen.

Betrachtet man die Anwendungsfälle von LLMs für Bedrohungsakteure und in der Cyberabwehr dediziert, so mag der Eindruck entstehen, dass beide Seiten von Effizienzsteigerungen gleichermaßen profitieren. Dies trifft jedoch nur für Organisationen zu, die bereits eine funktionierende Cyberabwehr etabliert haben, welche erfolgreich Cyberangriffe verhindert und kontinuierlich zur Verbesserung der Sicherheit beiträgt. LLMs können im multimodalen Kontext der Cybersicherheit als skalierbare Multiplikatoren betrachtet werden, die erfahrene Akteure besonders positiv beeinflussen, da diese bereits über viel Erfahrung in den jeweiligen Bereichen verfügen und die neuartigen und fortgeschrittenen Fähigkeiten von LLMs für ihre Ziele profitabel einsetzen können. Die zuvor skizzierte These der Kräfteparität verblasst jedoch wenn die Auswirkungen von leistungsstarken LLMs auf die Gesamtheit aller Teilnehmer im Cyberraum extrapoliert wird. Hier sehen sich alle Teilnehmer mit mannigfaltigen, unterschiedlich schlagkräftigen Bedrohungsakteuren konfrontiert, welche bereits heute erfolgreich sind und hohe Schäden anrichten und somit **alle** legitimen Organisationen bedrohen. Auf der

[81] Happe & Cito, 2023, S. 3–5.
[82] Yigit et al., 2024, S. 19–20.

anderen Seite stehen Unternehmen, Behörden, Staaten und Privatpersonen, die bereits heute in Summe keine effektiven Antworten haben um dieser Bedrohungslage zu begegnen. Das bedeutet, dass Organisationen, die bereits heute wenig Maßnahmen zur Cyberabwehr umgesetzt haben, zunächst wenig Mehrwerte mithilfe von LLMs heben können. Entsprechend ist wahrscheinlich dass sich das Kräfteverhältnis in Summe mindestens temporär zu Gunsten der Bedrohungsakteure verschiebt.

Entsprechend der Erkenntnisse aus der Analyse des Zustands der Cyberabwehr aus Abschnitt 3.2, sind besonders kleine und mittelgroße Unternehmen in Deutschland von der sich verschärfenden Bedrohungslage durch LLMs betroffen. Wie in Abbildung 4 dargestellt, wird etwa ein Drittel des Umsatzes deutscher Unternehmen in kleinen und mittelgroßen Unternehmen generiert. Dabei stellen kleine bis mittelgroße Unternehmen wie Abbildung 5 zeigt die nummerische die Mehrheit der deutschen Wirtschaft dar. Dabei sind kleine Unternehmen oft Teil der Lieferkette größerer Unternehmen[83]. Folglich sind auch größere Organisationen mit etablierter Cyberabwehr indirekt stärker bedroht, da sie in wirtschaftlicher Abhängig zu weniger stark geschützten Unternehmen stehen.

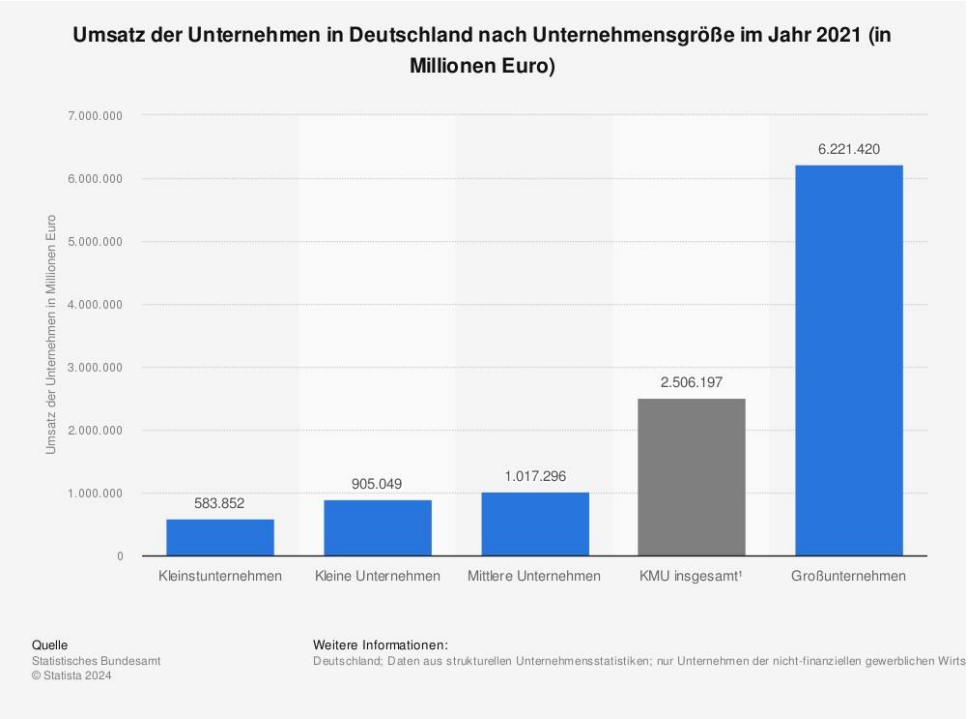

Abbildung 4: Umsatzverteilung nach Unternehmensgröße in Deutschland 2021[84]

[83] *Pressemitteilungen - LkSG: Neue Informationen zur Zusammenarbeit in der Lieferkette,* 2024.
[84] Statista, 2024a.

Die Verwendung von LLMs durch Bedrohungsakteure reduziert deren Kosten für einzelne Angriffe erheblich. Dies steigert die wirtschaftliche Attraktivität von Angriffen auf kleinere Unternehmen und somit steigt die Wahrscheinlichkeit[85] von Angriffen. Darüber hinaus könnten auch Akteure, die nicht primär auf Profit aus sind, neue Angriffsmuster entwickeln. Ein mögliches Szenario wäre, dass ein feindlicher staatlicher Akteur mithilfe autonomer Agenten systematisch die Schwachstellen einer Vielzahl kleiner Unternehmen ausnutzt, um deren Computersysteme zu sabotieren. Durch die enorme Anzahl an betroffenen Systemen würde eine vollständige Erholung entsprechend lange dauern. Dies würde eine ernsthafte Bedrohung für die gesamte Wirtschaft und damit inhärent die Bundesrepublik Deutschland darstellen.

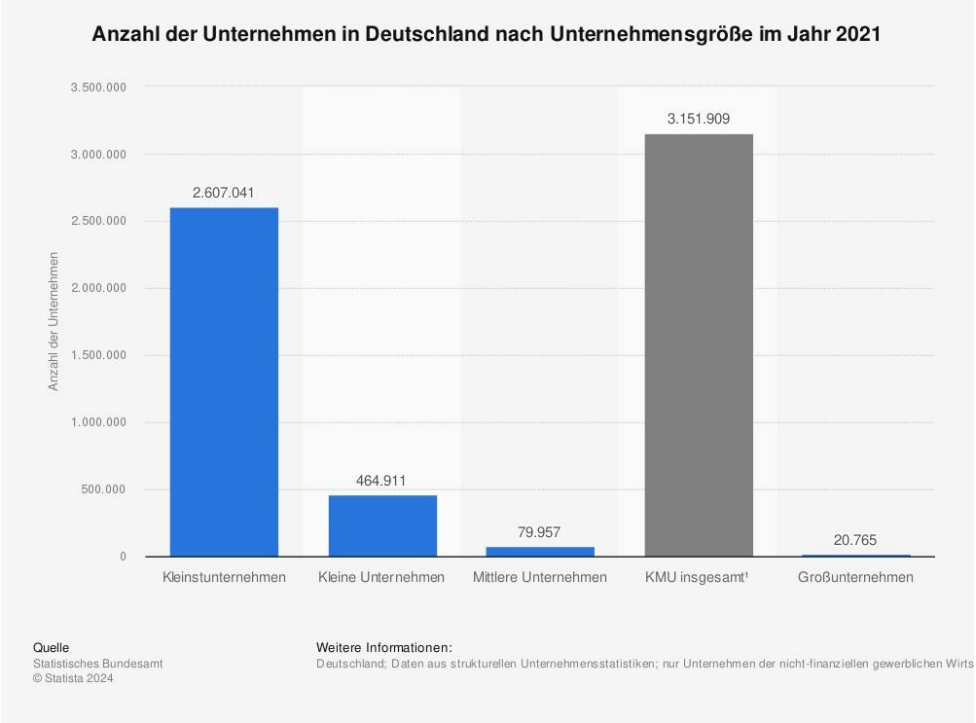

Abbildung 5: Aufteilung deutscher Unternehmen nach Unternehmensgröße im Jahr 2021[86]

In der vorliegenden Arbeit wurden verschiedene Anwendungsbereiche von LLMs sowohl für Bedrohungsakteure als auch für die Cyberabwehr beleuchtet. Es wurde mehrfach betont, dass es sich bei LLMs um ein äußerst dynamisches Forschungsfeld handelt, welches sich in den letzten zwei Jahren rasant weiterentwickelt hat. Die Fortschritte sind beeindruckend, allerdings befindet sich das Feld noch in einem so frühen Stadium, dass keine verlässlichen Vorhersagen über das potenzielle Bedrohungsausmaß durch KI-Agenten oder den Zeitrahmen

[85] Kshetri, 2006, S. 36–37.
[86] Statista, 2024b.

ihrer Einflussnahme getroffen werden können. Insgesamt weisen die verschiedenen Forschungsergebnisse aus dem akademischen Sektor auf eine reale und ernst zu nehmende Möglichkeit hin, dass LLMs Bedrohungsakteure in verschiedenen Disziplinen signifikant stärken könnten. Es ist anzunehmen, dass Bedrohungsakteure mit Nachdruck, Beständigkeit und letztlich Erfolg an der Entwicklung eigener LLM-basierter Tools arbeiten werden.

Es bleibt festzustellen, dass die zuständigen staatlichen Institutionen bislang nur vereinzelt und zurückhaltend Stellung zur Relevanz und zum defensiven oder offensiven Einsatz generativer KI in der Cybersicherheit bezogen haben. Das BSI[87] erkennt zwar in seiner jüngsten Einschätzung die multimodalen Einsatzgebiete von LLMs für Bedrohungsakteure und empfiehlt die Einhaltung von generischen Sicherheitspraktiken, wie z.B. Patchmanagement. Schlussfolgerungen auf nationaler oder europäischer Ebene werden vom BSI nicht gezogen. Von Seiten der ENISA ist keine dedizierte Bewertung zu LLMs zum Zeitpunkt der Ausarbeitung aufzufinden. Im folgenden Kapitel werden Lösungsansätze diskutiert, wie der sich ändernden Bedrohungslage, begegnet werden kann.

[87] Bundesamt für Sicherheit in der Informationstechnik [BSI], 2024.

5 Generative KI als Pfeiler der digitalen Souveränität

Generative künstliche Intelligenz ist eine mächtige Technologie, mit ungeahntem Potenzial. Ob diese maßgeblich als Schwert unserer Feinde oder als unsere Rüstung und Schild in die Geschichte eingehen wird hängt davon ab, ob wir als Gesellschaft, Nation und innerhalb unserer Bündnisse in der Lage sind dieses Potenzial systematisch, ubiquitär und zeitnah zu nutzen. Ein ganzheitliches Konzept, das alle relevanten Stakeholder aus Gesellschaft, Wirtschaft und Staat vereint, ist dafür unerlässlich. Die Klärung folgender Fragestellungen ist für eine erfolgreiche Transition in das Post-KI Zeitalter aus Sicht hinsichtlich Cybersicherheit essenziell:

1. Mit welchen Werkzeugen und Maßnahmen kann GenAI zur Verteidigung unserer digitaler Infrastruktur eingesetzt werden?

2. Wie können diese Lösungen flächendeckend und wirksam, unter Berücksichtigung des akuten Fachkräftemangels, implementiert werden?

3. Wie kann die Wirksamkeit aller Cyber-Schutzmaßnahmen auf nationaler Ebene verifiziert und kontinuierlich überwacht werden?

Dies kann nur mit gemeinsamer Anstrengung gelingen. GenAI wird bereits heute in die Sicherheitslösungen[88] führender Hersteller integriert. Jedoch ist die Adaption von KI-gestützter Software in Deutschland bisher gering[89]. Es fehlen konkrete Lösungskonzepte an welchen Stellen AI in der Sicherheitsarchitektur von Behörden und Unternehmen einen nennenswerten Beitrag leistet und wie diese integriert werden kann. Dieses Konzept kann ganzheitlich und abstrakt erarbeitet werden und fällt somit in die Zuständigkeit des BSI und der ENISA. Bspw. Könnte der Grundschutzkatalog entsprechend ergänzt werden.

Die reine Existenz von Maßnahmenkatalogen und Leitlinien ist, wie bereits die aktuelle Bedrohungslage eindrucksvollzeigt zeigt, nicht ausreichend um Computersysteme zu verteidigen. Diese Maßnahmen müssen korrekt und aufeinander abgestimmt umgesetzt werden. Um insbesondere kleinere Unternehmen beim Schutz ihrer Infrastruktur zu unterstützen und deren Defizite auszugleichen ist eine tiefgreifende Kooperation innerhalb der Wirtschaft notwendig. Größere Unternehmen besitzen in der Regel signifikante Kompetenzen und Ressourcen im Bereich der Cybersicherheit[90]. Diese können dafür eingesetzt um in kleineren Unternehmen ein höheres Schutzniveau zu etablieren. Diese

[88] Crowd Strike, 2023b; Fortinet, 2024.
[89] Ifo Institut, 2023.
[90] Heidt, Gerlach & Buxmann, 2019, S. 1298–1299.

Symbiose würde sich ebenfalls für große Unternehmen auszahlen, da diese effektiv zur Sicherheit ihrer Lieferkette beitragen können und so die direkte Bedrohung durch Supply Chain Angriffe senken. Dieser Austausch könnte z.B. innerhalb von Branchenverbänden stattfinden und durch staatliche Mittel monetär, aber auch informell gefördert werden.

Um den Wissenstransfer im Bereich der Cybersicherheit zu fördern und insbesondere in dynamischen Forschungsbereichen wie der KI auf dem neuesten Stand der Technik zu bleiben, ist es unerlässlich, sowohl in die Forschung und Entwicklung als auch in die Operationalisierung zu investieren. Der gesetzliche Rahmen in Deutschland sollte so angepasst werden, dass er international wettbewerbsfähige Forschung ermöglicht und attraktiv macht. Darüber hinaus sind umfangreiche Investitionen von staatlicher und privatwirtschaftlicher Seite in die Ausbildung von Fachkräften im Bereich KI und Cybersicherheit erforderlich, um ein robustes Fundament für die technologische Entwicklung Deutschlands zu legen. Die Einführung und breite Implementierung von Schutzmaßnahmen gegen durch KI verstärkte Bedrohungen könnten durch staatlich geförderte und vom BSI koordinierte Kompetenzzentren unterstützt werden, die Unternehmen und Behörden bei der Entwicklung, Optimierung und Implementierung von Cyberabwehrstrategien effizient und zugänglich unterstützen.

Die Wirksamkeit der zuvor genannten Maßnahmen bedarf einer kontinuierlichen und detaillierten Überprüfung. Ein Paradigmenwechsel hin zu dezentralisierten und internationalen Schwachstellen-Scans ist erforderlich, um ein genaues Bild der Cybersicherheitslage zu erhalten. Die Sicherheit deutscher Unternehmen und Behörden basiert nicht allein auf gesetzlichen Vorgaben, sondern wird letztlich durch die technische Anfälligkeit von Computersystemen bestimmt. Um eine technische Sichtweise auf die Cybersicherheit zu gewährleisten, ist es entscheidend, dass das BSI oder auf europäischer Ebene die ENISA alle aus dem Internet zugänglichen deutschen bzw. europäischen Systeme auf Schwachstellen hin überprüft. Wie in den Abschnitten 4.1.4 und 4.2.4 dargelegt, bietet Künstliche Intelligenz im Kontext automatisierter Penetrationstests erhebliche Möglichkeiten. Allerdings benötigen die Behörden dafür fortschrittliche KI-Systeme, die derzeit noch nicht auf dem Markt erhältlich sind. Im Sinne der nationalen Souveränität ist es unabdingbar, eigene Lösungen zu entwickeln. Dies dient auch dazu, proaktiv auf die Möglichkeit vorzubereiten, dass solche Systeme in naher Zukunft unseren Gegnern zur Verfügung stehen könnten, um ihnen technologisch überlegen zu sein. Zudem ist die Etablierung eines effizienten Meldesystems notwendig, damit betroffene Unternehmen, Behörden und Privatpersonen schnell und fachkundig über entdeckte Schwachstellen sowie Maßnahmen zu deren Behebung informiert

werden können. Eine gesetzliche Verpflichtung zur Behebung identifizierter Schwachstellen innerhalb einer angemessenen Frist könnte die Effektivität dieser Maßnahmen steigern. Hierfür muss nicht nur in die Behören investiert werden sondern auch der rechtliche Rahmen geschaffen werden.

Um KI-unterstütze Desinformationskampagnen zu erkennen und sinnvoll zu bekämpfen bedarf es ebenfalls weitreichender, aber vor allem gesellschaftlich akzeptierter und rechtstaatlicher Konzepte. Hierfür muss eine Debatte geführt werden, was eigentlich legitime und illegitime Inhalte sind so wie diese erkannt und markiert werden können. Weiterhin muss die öffentliche Wahrnehmung für generierte Inhalte, z.B. durch Aufklärungskampagnen, erhöht werden.[91]

Die durch GenAI induzierten Herausforderungen für Staat, Gesellschaft und Wirtschaft sind vielschichtig, zahlreich und dringlich. Besonders in der Cybersicherheit ist es von essenzieller Bedeutung diese ganzheitlich und kohärent zu lösen. Der Fokus und ein großer Teil des Ressourceneinsatzes muss auf der Lösung der technischen Bedrohungsfaktoren die von GenAI ausgehen liegen. Dies kann nur durch den Einsatz technischer Gegenmaßnahmen gelingen. Eine Regulation global und öffentlich verfügbarer Technologie kann nicht gelingen, da unsere Gegner von dieser nicht betroffen sind.

Die Entwicklung und der zielgerichtete Einsatz von KI-Systemen im Bereich Cybersicherheit ist für die digitale Souveränität der Bundesrepublik Deutschland unerlässlich. Fortschrittliche KI-Systeme und deren unaufhaltsame Demokratisierung ermöglichen unseren Gegnern stärkere und effizientere Angriffe gegen uns zu richten und häufiger erfolgreich zu sein. Wenn wir also in Zukunft nicht mehr in der Lage sind unsere digitale Infrastruktur, unsere Wirtschaft und unsere Demokratie gegen diese zu verteidigen, dann sind wir unseren Gegnern ausgeliefert und folglich auch nicht souverän. Und weil LLMs und KI im allgemeinen zukünftig eine Schlüsselrolle in der Cyberabwehr einnehmen wird ist es notwendig, dass in diesem Bereich keine überproportionale Abhängigkeit zu unseren Partnern entsteht. Deutschland muss in der Lage sein so existenziell wichtige Fähigkeiten wie die Entwicklung defensiver KI-Systeme und deren Einsatz in Eigenleistung erbringen zu können. Anderenfalls ist eine unabhängige und selbstbestimmte Nutzung des Cyberraums nicht möglich, da unser Schutz von Dritten abhängt. Die fortlaufende operative Überprüfung der deutschen, digitalen Infrastruktur durch KI-basierter Penetrationstests ist für die Erreichung eines zufriedenstellenden Sicherheitsniveaus unerlässlich. Denn in Summe ist die deutsche Wirtschaft

[91] Hartmann & Giles, 2020, S. 248–250.

kritische Infrastruktur für unsere Nation, diese vor breitflächigen und KI-gestützten Angriffen zu schützen ist prioritär.

6 Fazit & Ausblick

Im Rahmen dieser Arbeit wurde eine umfassende Analyse des aktuellen Stands der Technik sowie der neuesten Entwicklungen im Bereich der LLMs und GenAI durchgeführt. Es folgte eine detaillierte Betrachtung der aktuellen Bedrohungslage im Cyberspace, wobei insbesondere auf die signifikantesten Risiken eingegangen wurde. Auf dieser Grundlage wurden potenzielle Anwendungen von LLMs sowohl für Bedrohungsakteure als auch in der Cyberabwehr erörtert. Es zeigte sich, dass Bedrohungsakteure kurz- bis mittelfristig wahrscheinlich stärker von diesen Technologien profitieren werden. Abschließend wurden verschiedene Ansätze zur Bewältigung der identifizierten Herausforderungen vorgestellt, die in zukünftigen Forschungsarbeiten weiter vertieft werden sollten. Es wurde deutlich, dass eine erfolgreiche Bewältigung dieser Herausforderungen nur durch eine landesweite Zusammenarbeit zwischen Vertretern aus Gesellschaft, Staat und Wirtschaft möglich ist.

Diese Ausarbeitung verdeutlicht die weitreichenden Konsequenzen, die sich aus der breiten Verfügbarkeit leistungsfähiger LLMs für die Cybersicherheit ergeben. Dies stellt lediglich einen Ausschnitt aus dem breiten Spektrum an Herausforderungen und Möglichkeiten dar, die LLMs mit sich bringen. Zukünftige Untersuchungen sollten sich daher auch mit Fragen beschäftigen, wie diese Systeme ethisch verantwortungsvoll in der Wirtschaft eingesetzt werden können, oder wie sie unser Bildungssystem unterstützen und verbessern könnten. Darüber hinaus ist es entscheidend, die regulatorischen Rahmenbedingungen zu schaffen, die eine vertrauenswürdige und sichere Nutzung von KI-Technologien gewährleisten. Nur wenn es uns gelingt, die Chancen, die sich aus dem technologischen Neuland ergeben, zu erkennen, die damit verbundenen Herausforderungen zu meistern und eine schnelle und flexible Adaption zu fördern, wird Deutschland als Nation davon stark profitieren können.

Aus der Notwendigkeit heraus wesentliche Kompetenzen in Entwicklung und Betrieb von KI-Systemen aufzubauen eröffnet Deutschland zusätzliche Chancen. Mit der so erlangte Expertise können wir unsere Verbündeten bei der Cyberabwehr unterstützten und so unsere Rolle als verlässlicher Partner einer internationalen Sicherheitsarchitektur zu stärken. Darüber hinaus würde der Standort Deutschland von Investitionen in die Zukunftstechnologie massiv profitieren, da sich deren Effekt nicht nur auf die Cybersicherheit beschränkt sondern in nahezu jedem Wirtschaftszweig aber auch in der Administration relevant ist.

Zusätzlich zu den bereits diskutierten Punkten ist es wichtig, die Rolle der internationalen Zusammenarbeit und des Informationsaustauschs zu betonen. Die Cybersicherheit ist ein globales Anliegen, und nur durch die Bündelung von Ressourcen und Expertise können wir effektive Gegenmaßnahmen gegen die sich ständig weiterentwickelnden Bedrohungen entwickeln. Ebenso sollte die Bedeutung von öffentlichen Aufklärungskampagnen und der Förderung von Cybersicherheitsbewusstsein in der Bevölkerung nicht unterschätzt werden. Schließlich ist die kontinuierliche Weiterentwicklung von KI-gestützten Sicherheitssystemen von entscheidender Bedeutung, um den Schutz vor Cyberangriffen zu verbessern und die Resilienz unserer digitalen Infrastrukturen zu stärken.

7 Literaturverzeichnis

Alto, V. (2023). *Modern Generative AI with ChatGPT and OpenAI Models. Leverage the Capabilities of OpenAI's LLM for Productivity and Innovation with GPT3 and GPT4* (1st ed.). Birmingham: Packt Publishing Limited.

Arora, D., Singh, H. G. & Mausam. (2023, 24. Mai). *Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models*. Verfügbar unter http://arxiv.org/pdf/2305.15074

Bethany, M., Galiopoulos, A., Bethany, E., Karkevandi, M. B., Vishwamitra, N. & Najafirad, P. (2024, 18. Januar). *Large Language Model Lateral Spear Phishing: A Comparative Study in Large-Scale Organizational Settings*. Verfügbar unter http://arxiv.org/pdf/2401.09727v1

Bitkom (2023a). Markt für IT-Sicherheit wächst auf mehr als 9 Milliarden Euro. *Bitkom e.V.* Zugriff am 16.05.2024. Verfügbar unter https://www.bitkom.org/Presse/Presseinformation/Markt-IT-Sicherheit-waechst-mehr-9-Milliarden-Euro#_

Bitkom. (2023b). Wirtschaftsschutz 2023. Zugriff am 16.05.2024. Verfügbar unter https://www.bitkom.org/sites/main/files/2023-09/Bitkom-Charts-Wirtschaftsschutz-Cybercrime.pdf

Bundesamt für Sicherheit in der Informationstechnik. Schadhafte Version der 3CX Desktop App im Umlauf. Zugriff am 13.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2023/2023-214778-1032.pdf?__blob=publicationFile&v=4

Bundesamt für Sicherheit in der Informationstechnik. (2020). BSI - Studie Penetrationstests. Zugriff am 11.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Studien/Penetrationstest/penetrationstest.pdf?__blob=publicationFile&v=3

Bundesamt für Sicherheit in der Informationstechnik. (2022). Die Lage der IT-Sicherheit in Deutschland 2022. Zugriff am 16.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2022.pdf?__blob=publicationFile&v=8

Bundesamt für Sicherheit in der Informationstechnik. (2023a). *APT.* Zugriff am 12.05.2024. Verfügbar unter https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Empfehlungen-nach-Gefaehrdungen/APT/apt_node.html

Bundesamt für Sicherheit in der Informationstechnik. (2023b). Die Lage der IT-Sicherheit in Deutschland. Zugriff am 11.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2023.pdf?__blob=publicationFile&v=8

Bundesamt für Sicherheit in der Informationstechnik. (2024). How is AI changing the cyber threat landscape? Zugriff am 11.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/How-is-AI-changing-cyber-threat-landscape.pdf?__blob=publicationFile&v=2

Bundesamt für Sicherheit in der Informationstechnik. (2024). Kritische Backdoor in XZ für Linux. Zugriff am 13.05.2024. Verfügbar unter https://www.bsi.bund.de/SharedDocs/Cybersicherheitswarnungen/DE/2024/2024-223608-1032.pdf?__blob=publicationFile&v=5

Chaudhary, Y. & Penn, J. (2024, 6. Mai). *Large Language Models as Instruments of Power: New Regimes of Autonomous Manipulation and Control*. Verfügbar unter http://arxiv.org/pdf/2405.03813v1

Crowd Strike. (2023a). 2023 Global Threat Report. Zugriff am 11.05.2024. Verfügbar unter https://go.crowdstrike.com/rs/281-OBQ-266/images/CrowdStrike2023GlobalThreatReport.pdf

Crowd Strike (2023b, 30. Mai). Introducing Charlotte AI, CrowdStrike's Generative AI Security Analyst: Ushering in the Future of AI-Powered Cybersecurity. *CrowdStrike.* Zugriff am 22.05.2024. Verfügbar unter https://www.crowdstrike.com/blog/crowdstrike-introduces-charlotte-ai-to-deliver-generative-ai-powered-cybersecurity/

Crowd Strike. (2024a). 2024 Global Threat Report. Zugriff am 11.05.2024. Verfügbar unter https://go.crowdstrike.com/rs/281-OBQ-266/images/GlobalThreatReport2024.pdf

Crowd Strike. (2024b, 21. Februar). *What are Attack Vectors: Definition & Vulnerabilities - CrowdStrike.* Zugriff am 11.05.2024. Verfügbar unter https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/attack-vector/

Crowd Strike. (2024c, 21. Februar). *What is a Cyber Threat Actor? - CrowdStrike.* Zugriff am 11.05.2024. Verfügbar unter https://www.crowdstrike.com/cybersecurity-101/threat-actor/

Das, A. (2023, 28. September). ChatGPT Got Internet Access! *DEV Community.* Zugriff am 09.05.2024. Verfügbar unter https://dev.to/ananddas/chatgpt-got-internet-access-21gj

Das, A., Chen, S.-C., Shyu, M.-L. & Sadiq, S. (2023). Enabling Synergistic Knowledge Sharing and Reasoning in Large Language Models with Collaborative Multi-Agents. In *2023 IEEE 9th International Conference on Collaboration and Internet Computing (CIC)* (S. 92–98). IEEE.

Deng, G., Liu, Y., Mayoral-Vilches, V., Liu, P., Li, Y., Xu, Y. et al. (2023, 13. August). *PentestGPT: An LLM-empowered Automatic Penetration Testing Tool.* Zugriff am 18.05.2024. Verfügbar unter https://arxiv.org/pdf/2308.06782

ENISA. (2023). *ENISA Threat Landscape 2023.* Zugriff am 11.05.2024. Verfügbar unter https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

Falade, P. V. (2023). Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology,* 185–198. https://doi.org/10.32628/CSEIT2390533

Fang, R., Bindu, R., Gupta, A. & Kang, D. (2024, 12. April). *LLM Agents can Autonomously Exploit One-day Vulnerabilities.* Zugriff am 18.05.2024. Verfügbar unter https://arxiv.org/pdf/2404.08144

Fang, R., Bindu, R., Gupta, A., Zhan, Q. & Kang, D. (2024, 6. Februar). *LLM Agents can Autonomously Hack Websites*. Verfügbar unter http://arxiv.org/pdf/2402.06664

Fortinet. (2024, 22. Mai). *Generative AI Security – FortiAI | Fortinet.* Zugriff am 22.05.2024. Verfügbar unter https://www.fortinet.com/de/products/fortiai

Fredheim, R. & Pamment, J. (2024). Assessing the risks and opportunities posed by AI-enhanced influence operations on social media. *Place Branding and Public Diplomacy*. https://doi.org/10.1057/s41254-023-00322-5

GitHub. (2024). *Measuring the impact of GitHub Copilot.* Zugriff am 17.05.2024. Verfügbar unter https://resources.github.com/learn/pathways/copilot/essentials/measuring-the-impact-of-github-copilot/

Göbel, M. (2024). *Phi-3: Neue Maßstäbe für die Möglichkeiten kleiner Sprachmodellen.* Zugriff am 26.05.2024. Verfügbar unter https://news.microsoft.com/de-de/phi-3-neue-massstaebe-fuer-die-moeglichkeiten-kleiner-sprachmodelle/

Gupta, M., Akiri, C., Aryal, K., Parker, E. & Praharaj, L. (2023, 3. Juli). *From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy*. Verfügbar unter http://arxiv.org/pdf/2307.00691

Happe, A. & Cito, J. (2023). *Getting pwn'd by AI: Penetration Testing with Large Language Models*. Verfügbar unter http://arxiv.org/pdf/2308.00121 https://doi.org/10.1145/3611643.3613083

Hartmann, K. & Giles, K. (2020). The Next Generation of Cyber-Enabled Information Warfare. In *2020 12th International Conference on Cyber Conflict (CyCon)* (S. 233–250). Place of publication not identified: IEEE.

Hazell, J. (2023, 11. Mai). *Spear Phishing With Large Language Models*. Verfügbar unter http://arxiv.org/pdf/2305.06972v3

Heidt, M., Gerlach, J. P. & Buxmann, P. (2019). Investigating the Security Divide between SME and Large Companies: How SME Characteristics Influence Organizational IT Security Investments. *Information Systems Frontiers, 21* (6), 1285–1305. https://doi.org/10.1007/s10796-019-09959-1

Hu, K. (2023, 2. Februar). ChatGPT sets record for fastest-growing user base - analyst note. *Reuters Media.* Zugriff am 09.05.2024. Verfügbar unter

https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/
HYAS Infosec Inc. (2023). *BlackMamba: Using AI to Generate Polymorphic Malware.* Zugriff am 17.05.2024. Verfügbar unter https://www.hyas.com/blog/blackmamba-using-ai-to-generate-polymorphic-malware
Ifo Institut. (2023). *13,3% der Unternehmen in Deutschland nutzen Künstliche Intelligenz.* Zugriff am 22.05.2024. Verfügbar unter https://www.ifo.de/fakten/2023-08-02/unternehmen-deutschland-nutzen-kuenstliche-intelligenz
Kim, S. Y., Fan, Z., Noller, Y. & Roychoudhury, A. (2024, 7. Mai). *Codexity: Secure AI-assisted Code Generation*. Verfügbar unter http://arxiv.org/pdf/2405.03927v1
Klipper, S. (2015). *Cyber Security. Ein Einblick für Wirtschaftswissenschaftler* (essentials). Wiesbaden: Springer Vieweg.
Koutcheme, C., Dainese, N., Sarsa, S., Hellas, A., Leinonen, J. & Denny, P. (2024). *Open Source Language Models Can Provide Feedback: Evaluating LLMs' Ability to Help Students Using GPT-4-As-A-Judge*. Verfügbar unter http://arxiv.org/pdf/2405.05253
Krause, S., Panchal, B. H. & Ubhe, N. (2024, 16. April). *The Evolution of Learning: Assessing the Transformative Impact of Generative AI on Higher Education*. Verfügbar unter http://arxiv.org/pdf/2404.10551
Kshetri, N. (2006). The simple economics of cybercrimes. *IEEE Security & Privacy Magazine, 4* (1), 33–39. https://doi.org/10.1109/MSP.2006.27
LangChain. (2024). *LangChain.* Zugriff am 09.05.2024. Verfügbar unter https://www.langchain.com/
Li, H. & Shan, L. (2023). LLM-based Vulnerability Detection. In *2023 International Conference on Human-Centered Cognitive Systems (HCCS)* (S. 1–4). IEEE.
Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R. et al. (2023, 30. Mai). *Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate*. Verfügbar unter http://arxiv.org/pdf/2305.19118v1
Mao, R., Chen, G., Zhang, X., Guerin, F. & Cambria, E. (2023, 24. August). *GPTEval: A Survey on Assessments of ChatGPT and GPT-4*. Verfügbar unter http://arxiv.org/pdf/2308.12488
Meta. (2024). *Meta Llama 3.* Zugriff am 26.05.2024. Verfügbar unter https://llama.meta.com/llama3/
Microsoft. (2024a, 9. Mai). *Create AI agents with Semantic Kernel.* Zugriff am 09.05.2024.
Microsoft. (2024b, 9. Mai). *GitHub - microsoft/autogen: A programming framework for agentic AI. Discord: https://aka.ms/autogen-dc. Roadmap:*

https://aka.ms/autogen-roadmap. Zugriff am 09.05.2024. Verfügbar unter https://github.com/microsoft/autogen

Morgan, J. (2024, 9. Mai). *GitHub - ollama/ollama: Get up and running with Llama 3, Mistral, Gemma, and other large language models.* Zugriff am 09.05.2024. Verfügbar unter https://github.com/ollama/ollama

Muckin, M. & Fitch, S. (2019). A Threat-Driven Approach to Cyber Security. Zugriff am 11.05.2024. Verfügbar unter https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Threat-Driven-Approach.pdf

National Institute of Standards and Technology. (2023). Cybersecurity Workforce Demand 2023. Zugriff am 17.05.2024. Verfügbar unter https://www.nist.gov/system/files/documents/2023/06/05/NICE%20FactSheet_Workforce%20Demand_Final_20211202.pdf

National Institute of Standards and Technology. (2024, 11. Mai). *penetration testing - Glossary | CSRC.* Zugriff am 11.05.2024. Verfügbar unter https://csrc.nist.gov/glossary/term/penetration_testing

Nong, Y., Aldeen, M., Cheng, L., Hu, H., Chen, F. & Cai, H. (2024, 27. Februar). *Chain-of-Thought Prompting of Large Language Models for Discovering and Fixing Software Vulnerabilities*. Verfügbar unter http://arxiv.org/pdf/2402.17230v1

OpenAI. (2024a). *Hello GPT-4o.* Zugriff am 18.05.2024. Verfügbar unter https://openai.com/index/hello-gpt-4o/

OpenAI. (2024b, 9. Mai). *Introducing ChatGPT.* Zugriff am 09.05.2024. Verfügbar unter https://openai.com/index/chatgpt/

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I. et al. (2023, 15. März). *GPT-4 Technical Report*. Verfügbar unter http://arxiv.org/pdf/2303.08774v6

Patel, U., Yeh, F.-C. & Gondhalekar, C. CANAL - Cyber Activity News Alerting Language Model : Empirical Approach vs. Expensive LLMs. *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC), Houston, TX, USA,* 1–12. https://doi.org/10.1109/ICAIC60265.2024.10433839

(2024, 20. Mai). *Pressemitteilungen - LkSG: Neue Informationen zur Zusammenarbeit in der Lieferkette.* Zugriff am 20.05.2024. Verfügbar unter https://www.bafa.de/SharedDocs/Downloads/DE/Lieferketten/faq_zusammenarbeit_lieferketten.pdf?__blob=publicationFile&v=6

Purba, M. D., Ghosh, A., Radford, B. J. & Chu, B. (2023). Software Vulnerability Detection using Large Language Models. In *2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops (ISSREW)* (S. 112–119). IEEE.

Rahman, M. H., Cassandro, R., Wuest, T. & Shafae, M. (2023, 30. Dezember). *Taxonomy for Cybersecurity Threat Attributes and Countermeasures in Smart Manufacturing Systems*. Verfügbar unter http://arxiv.org/pdf/2401.01374

Raschka, S. (2024). *Build a Large Language Model (from Scratch):* Manning Publications Co.

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.-N. & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation, 46,* 1–10. https://doi.org/10.1016/j.fsidi.2023.301609

Shashwat, K., Hahn, F., Ou, X., Goldgof, D., Hall, L., Ligatti, J. et al. (2024, 30. Januar). *A Preliminary Study on Using Large Language Models in Software Pentesting*. Verfügbar unter http://arxiv.org/pdf/2401.17459

Statista. (2024a, 20. Mai). *Umsatz der Unternehmen nach Unternehmensgröße | Statista.* Zugriff am 20.05.2024. Verfügbar unter https://de.statista.com/statistik/daten/studie/731964/umfrage/umsatz-der-unternehmen-in-deutschland-nach-unternehmensgroesse/

Statista. (2024b, 20. Mai). *Unternehmen nach Unternehmensgröße in Deutschland | Statista.* Zugriff am 20.05.2024. Verfügbar unter https://de.statista.com/statistik/daten/studie/731859/umfrage/unternehmen-in-deutschland-nach-unternehmensgroesse/

Vaithilingam, P., Zhang, T. & Glassman, E. L. (2022). Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In S. Barbosa, C. Lampe, C. Appert & D. A. Shamma (Hrsg.), *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (S. 1–7). New York, NY, USA: ACM.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E. et al. (2023, 16. August). *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*. Verfügbar unter http://arxiv.org/pdf/2308.08155

Xu, J., Stokes, J. W., McDonald, G., Bai, X., Marshall, D., Wang, S. et al. (2024, 3. Februar). *AutoAttacker: A Large Language Model Guided System to Implement Automatic Cyber-attacks*. Verfügbar unter http://arxiv.org/pdf/2403.01038v1

Yigit, Y., Buchanan, W. J., Tehrani, M. G. & Maglaras, L. (2024, 13. März). *Review of Generative AI Methods in Cybersecurity*. Verfügbar unter http://arxiv.org/pdf/2403.08701

Zhou, X., Cao, S., Sun, X. & Lo David. (2024, 4. März). *Large Language Model for Vulnerability Detection and Repair: Literature Review and the Road Ahead*. Verfügbar unter http://arxiv.org/pdf/2404.02525v2