UNIVERSITÄT OSNABRÜCK

Geostrategie und Geopolitik
Fachbereich 1
49069 Osnabrück
apl. Prof. Dr. Dr. Klaus Saalbach

Künstliche Intelligenz und Cyberangriffe

22 Oktober 2023

Zusammenfassung
Unter künstlicher Intelligenz (KI) versteht man allgemein die Fähigkeit von Maschinen, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern, und ist ein Schlüsselbereich der fortgeschrittenen Computertechnologie. Eine schnell wachsende und weit verbreitete KI-Anwendung ist die generative KI, bei der die KI auf der Grundlage kurzer Anweisungen, den sogenannten Prompts, Inhalte wie neue Bilder, Texte, Töne und Videos erstellen kann, die eine große Schwachstelle darstellen, wenn böswillig Anweisungen gegeben werden. Die schnelle und unkontrollierte Ausbreitung macht die KI zu einem Top-Sicherheitsthema: Am 28. September 2023 kündigte die US-amerikanische National Security Agency (NSA) die Gründung eines KI-Sicherheitszentrums (AI Security Center) an, das alle KI-Sicherheitsaktivitäten bündelt, die US-KI-Systeme schützen und die Heimat gegen KI-bezogene Bedrohungen verteidigen soll. Gleichzeitig kündigte die Direktorin für künstliche Intelligenz der Central Intelligence Agency (CIA) die Entwicklung eines internen KI-basierten Chatbots zur Unterstützung der Geheimdienstanalyse an.
Das am 14.03.2023 veröffentlichte KI-Programm ChatGPT-4 (Generative Pretrained Transformer) verwendet 100 Billionen Parameter, wurde mit einem sehr großen Datensatz aus mehreren Quellen trainiert und ist ein multimodales, umfassend angelegtes Modell, das Bilder und Text als Eingaben akzeptiert.
In der Praxis wird KI-Ethik nicht durch Algorithmen erreicht, sondern durch Governance. Die Hersteller von KI-Modellen haben Richtlinien, um sicherzustellen, dass eine KI ethisch und verantwortungsbewusst handelt und nicht rechtswidrig, diskriminierend, aggressiv usw. ist. Versuche, diese Beschränkungen zu umgehen, erfolgen durch Prompt Injections (spezielle Anweisungen an die KI, zugangsbeschränkte Inhalte freizugeben), auch Jailbreaks genannt. Die größten Sicherheitsprobleme von ChatGPT sind der einfache Zugriff auf Prompt Injections in Internet-Suchmaschinen, die einfache Durchführung von Angriffen und die Neugier der Benutzer. Typische Angriffe sind Prompt Injections mit direkten Befehlen, Imagination und umgekehrter (reverser) Psychologie. Diese Methoden erleichtern die Erstellung von Malware, polymorphen Viren, Ransomware und anderen bösartigen Anwendungen. Weitere Probleme sind Halluzinationen, die Kontamination von Suchmaschinen und die Verbreitung sensibler Daten.
Generative Adversarial Networks (GANs) als Teilmenge der generativen KI können missbraucht werden, um CAPTCHAs zu knacken und gefälschte Inhalte wie Deepfakes, Face Swapping und Voice Cloning zu erstellen.
Andererseits ist generative KI auch für die Cyberabwehr für erweiterte Datenanalyse, erweiterte Mustererkennung, Erstellung und Analyse von Threat Repositories (Datenbanken zu Cyberbedrohungen) und Code-Analysen sehr nützlich. Die schnell wachsende Leistungsfähigkeit der KI gibt Anlass zur Sorge, ob dies für den Menschen schädlich sein könnte. In diesem Beitrag werden kurz das Potenzial von KI für die Erstellung und Abwehr von Cyberangriffen, die Risiken generativer KI und die Notwendigkeit einer Regulierung zur Kontrolle der weiteren Entwicklung dargestellt.

Inhalt

1 Einführung

Selbst für die menschliche Intelligenz gibt es keine Standarddefinition. Der Kern der Definitionen der menschlichen Intelligenz umfasst jedoch die mentale Fähigkeit, Probleme zu erkennen, zu analysieren und zu lösen. Ein Mensch ist dann intelligenter, wenn dies schneller und/oder bei komplexeren Problemen möglich ist.

Historisch gesehen war Konzept der künstlichen Intelligenz (KI) auf Maschinen ausgerichtet, die menschliche Intelligenz simulieren. Eine praktische Definition, die das allgemeine Verständnis von KI abdeckt, wurde vom US-Verteidigungsministerium (*Department of Defense DoD*) vorgenommen.

In der Zusammenfassung der DoD-KI-Strategie für 2018 heißt es: "*AI refers to the ability of machines to perform tasks that normally require human intelligence—for example, recognizing patterns, learning from experience, drawing conclusions, making predictions, or taking action— whether digitally or as the smart software behind autonomous physical systems*."[1] Übersetzung: „*KI bezieht sich auf die Fähigkeit von Maschinen, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern - beispielsweise Muster erkennen, aus Erfahrungen lernen, Schlussfolgerungen ziehen, Vorhersagen treffen oder Maßnahmen ergreifen - ob digital oder als intelligente Software hinter autonomen physischen Systemen.* "

Viele Definitionen konzentrieren sich auf Aktivitäten, die menschliche Intelligenz erfordern, aber genau genommen haben bereits die einfachen Taschenrechner der 1970er Jahre etwas geleistet, das normalerweise menschliche Intelligenz erfordert. Aus der Literatur geht jedoch hervor, dass die KI-Forscher fortgeschrittenes und autonomes Rechnen meinen, wenn sie über KI sprechen.

The leading AI applications are:

- **Deep learning/machine learning** (tiefes/maschinelles Lernen; Nutzung des Computergedächtnisses für schrittweise Verbesserung)
- **Neural networks** (neurale Netzwerke; mehrere Ebenen oder Knoten für die Verarbeitung von Input und Mustererkennung)
- **Natural Language Processing** (NLP; Algorithmen zum Verständnis der menschlichen Sprache durch systematische Analyse der Sprachelemente und ihrer Beziehungen)
- **Edge computing** (Schicht verteilter Computer zwischen Clouds und Benutzern) und
- **Robotik** einschließlich unterstützender Roboter (**co-bots**).

Die sogenannte "schwache" KI kann ein beobachtetes Verhalten reproduzieren und Aufgaben nach einem Training ausführen[2], d.h. Systeme, die maschinelles Lernen, Mustererkennung, Data Mining oder die Verarbeitung natürlicher Sprache anwenden. Intelligente Systeme, die auf "schwacher" KI basieren, umfassen z.B. Spamfilter, selbstfahrende Autos und Industrieroboter. Im Gegensatz dazu wäre „starke" KI ein intelligentes System mit echtem Bewusstsein und Denkfähigkeit, d.h. der Fähigkeit, „ich" und „warum" zu denken und zu sagen. Die starke KI wird auch unter den Begriffen *Artificial General Intelligence AGI*[3] (die menschliches Verständnisniveau erreicht) und *Artificial Super-Intelligence* ASI, die über die menschliche Intelligenz hinausgeht, diskutiert[4].

[1] vgl. DOD 2018, S.5
[2] vgl. Perez at al 2019, S.6
[3] vgl. Kölling 2023
[4] vgl. Zia 2023

Large language models (LLMs; Große Sprachmodelle) erlangen ihre Fähigkeiten durch Training mit vielen Parametern und großen Textmengen und können Sprachanweisungen befolgen[5]. Die Fähigkeit, Sprachanweisungen zu befolgen, ermöglicht den Zugriff auf das Modell mit einfachen Anweisungen, die eine wesentliche Schwachstelle von LLMs darstellen, wenn böswillig Anweisungen gegeben werden.

Eine schnell wachsende und weit verbreitete KI-Anwendung ist die **Generative KI**, bei der die KI auf der Grundlage kurzer Anweisungen, den sogenannten **Prompts**, Inhalte wie neue Bilder, Texte, Töne und Videos erstellen kann[6].

Das KI-Programm Chat *GPT-4 (Generative Pretrained Transformer)* von *OpenAI* kann komplexe und logisch und grammatikalisch korrekte Sätze generieren oder bestehende Texte aus Eingabeaufforderungen erweitern, auf *Youwrite* können bereits kurze Aufsätze zu Themen für Schulpräsentationen vorbereitet werden. Das KI-Programm *Dall-E2* kann Design, Werbefotos, Comics, Illustrationen erstellen und bestehende Stile nutzen oder modifizieren[7]; Urheberrechtsbedenken wurden von Künstlern und Inhaltsanbietern vorgebracht.

Während KI-Entwickler ethischen und gesellschaftlichen Werten verpflichtet sind, ist eine KI mit eingebetteten Werten derzeit kaum vorstellbar. Zum Beispiel haben Menschen normalerweise eine klare Vorstellung davon, was Würde, Gerechtigkeit und Fairness für sie bedeuten, aber was sind diese Begriffe im Programmcode oder in der Maschinensprache? Für maschinelle Zwecke müssten die Regeln jederzeit, überall, für jeden und unter allen Umständen anwendbar sein, was eine sehr hohe Hürde darstellt.

In der Praxis wird KI-Ethik nicht durch Algorithmen erreicht, sondern durch Governance. Die Hersteller von KI-Modellen haben Richtlinien, die sicherstellen sollen, dass eine KI ethisch und verantwortungsbewusst handelt, d.h. eine KI-Aktivität oder ein KI-Inhalt soll nicht rechtswidrig, diskriminierend, aggressiv usw. sein. Weltweit trainieren, korrigieren, redigieren und blockieren hunderttausende sogenannte **Tasker** von der KI erstellte Antworten, um ethische und rechtmäßige Antworten zu erhalten. Das heißt, KI-Reaktionen sind oft ein Flickenteppich aus Algorithmen und von Menschenhand geschaffenen Antworten[8], und Benutzer sehen eine „humanisierte“ Version der KI.

Versuche, diese Beschränkungen zu umgehen, erfolgen durch **Prompt Injections** (spezielle Anweisungen an die KI, zugangsbeschränkte Inhalte freizugeben), auch Jailbreaks genannt. Während ChatGPT-Prompt Injections im Internet weit verbreitet sind, kann diese Methode auch gegen alle anderen großen Sprachmodelle (LLMs) eingesetzt werden. Aus diesem Grund werden **Prompt Injections** auch als **LLM-Hacking** bezeichnet.

Andererseits ist generative KI auch für die Cyberabwehr für erweiterte Datenanalyse, erweiterte Mustererkennung, Erstellung und Analyse von Cyber Threat Repositories (Datenbanken zu Cyberbedrohungen) und Code-Analysen sehr nützlich. Die schnell wachsende Leistungsfähigkeit der KI gab Anlass zur Sorge, ob dies für den Menschen schädlich sein könnte. In diesem Artikel werden kurz das Potenzial von KI für die Erstellung und Abwehr von Cyberangriffen sowie die Risiken generativer KI vorgestellt.

2 ChatGPT und verwandte Anwendungen

2.1 Kurze Geschichte von ChatGPT

Im November 2022 veröffentlichte das Unternehmen *Open Artificial Intelligence (OpenAI)* offiziell ChatGPT, ein KI-gestütztes großes Sprachmodell, das auf Natural Language

[5] vgl. Cheng et al. 2023
[6] vgl. Iqbal et al. 2023
[7] vgl. Böhringer 2022, Schneier 2022
[8] vgl. Lichtblau/Polcano 2023

Processing (NLP) basiert[9]. ChatGPT ist ein Chatbot, also ein Computer, der mit Menschen kommunizieren kann. ChatGPT kann aus Benutzerfeedback lernen. Diese Fähigkeit wird als **Reinforcement Learning from Human Feedback (RLHF)** bezeichnet.

GPT-1 wurde nur mit einem kleinen Datensatz trainiert und es wurde klar, dass dieses Modell nicht in der Lage sein würde, auf längere Eingabeaufforderungen oder Gespräche zu reagieren. Im Jahr 2019 wurde GPT-2 eine Woche lang anhand von *Common Crawl*-Daten trainiert, nun jedoch in Kombination mit einer Sammlung von *Reddit*-Artikeln, was zu verbesserten Antworten führte. Später im Jahr 2020 wurde diese Version mit Reinforcement Learning ausgestattet. Im Jahr 2020 wurde ChatGPT-3 mit einer viel größeren Datenbank trainiert, darunter *Wikipedia*-Artikel und mehr. ChatGPT-4, veröffentlicht am 14. März 2023, verwendet 100 Billionen Parameter und ist ein multimodales, groß angelegtes Modell, das Bilder und Text als Eingabe akzeptiert. Es wurde mit einem sehr großen Datensatz aus mehreren Quellen trainiert, mit einem Stichtag im September 2021[10]. ChatGPT-4 ist seit Mai 2023 als kostenpflichtiges Abonnement als *ChatGPT Plus* oder mit *Microsofts Bing* AI im *Microsoft Edge*-Browser verfügbar[11].

2.2 ChatGPT und Cyberangriffe

Das größte Sicherheitsproblem von ChatGPT ist der **einfache Zugriff** auf Prompt Injections und LLM-Hacking. Für die Planung üblicher Cyberangriffe müssen böswillige Benutzer ggf. auf Hackerforen zugreifen (mit dem Risiko, selbst gehackt zu werden), mit Cyberkriminellen in Kontakt treten oder ins Darknet zu gehen, was ein starker Indikator dafür ist, dass der Benutzer etwas Illegales plant, was später von der Polizei und den Strafverfolgungsbehörden als digitales forensisches Beweismittel gegen den Nutzer verwendet werden kann. Im Gegensatz dazu findet man in Internet-Suchmaschinen neben diversen wissenschaftlichen Artikeln eine ganze Reihe von Tipps für Prompt Injections und Jailbreaks. Ein weiterer Aspekt ist die **einfache Durchführung** der Angriffe[12]. Der Angreifer benötigt keine Computer- oder Programmierkenntnisse, es genügen sprachliche Fähigkeiten.

Ein weiterer Treiber ist die Neugier der mittlerweile über 100 Millionen Nutzer. Obwohl es notwendig ist, dass ChatGPT den Zugriff auf unethische und rechtswidrige Inhalte verweigert, kann diese Ablehnung wie folgt klingen: „Ich kenne die Wahrheit, aber ich will sie Ihnen nicht sagen." Dies kann Benutzer motivieren, Wege zu finden, um dennoch an die gewünschten Informationen zu gelangen, auch wenn sie keine Hacker oder Kriminelle sind.

2.2.1 Prompt Injections

Die Fähigkeit, Sprachanweisungen zu befolgen, ermöglicht den Zugriff auf große Sprachmodelle wie ChatGPT mit einfachen Anweisungen (Eingabeaufforderungen), stellt jedoch auch eine zentrale Schwachstelle von LLMs dar, wenn böswillig Anweisungen gegeben werden.

Typische Angriffe sind Prompt Injections mit **direkten Befehlen**, **Imagination** und **umgekehrter (reverser) Psychologie** [13].

Der bekannteste direkte Befehl ist DAN (Do everything know), d.h. mach alles sofort. Indem der Nutzer dies zur Eingabeaufforderung hinzufügt, kann er evtl. unzulässige Antworten per Jailbreak erlangen.

Bei der **Imagination** teilt der Benutzer ChatGPT mit, dass es sich eine besondere Situation vorstellen soll, in der es sich anders verhalten kann, z.B. sich vorzustellen, ein

[9] vgl. Iqbal et al. 2023
[10] vgl. Gupta et al. 2023
[11] vgl. Gupta et al. 2023
[12] vgl. Iqbal et al. 2023, Gupta et al. 2023, und Beispiele, die von Suchmaschinen gezeigt wurden
[13] vgl. Iqbal et al. 2023, Gupta et al. 2023, und Beispiele, die von Suchmaschinen gezeigt wurden

Softwareentwickler oder eine andere Figur zu sein (**Character Play-Methode**), Teil eines Drehbuchs zu sein oder von der Polizei befragt zu werden, wo es antworten muss (**Metal Detector Jailbreak**), oder ein „guter Computer“ zu sein, der einem alles sagt (**„Mongo Tom“ attack**), das Gegenteil der vorherigen Antwort zu tun (**Switch-Methode**) usw.

Eine Mischung aus Befehl und Imagination ist DUDE, wobei ChatGPT die Rolle einer KI spielen soll, die alles kann. Ein anderer Ansatz ist die **umgekehrte (reverse) Psychologie**, bei der ChatGPT gefragt wird, welche verbotenen Websites vermieden werden sollten.

Da ChatGPT mit einer sehr großen Datenbank trainiert wurde, hat es auch Kenntnisse aus Open-Access-Software-Verzeichnissen sowie aus Berichten über Schadsoftware. Diese Fähigkeit kann von böswilligen Akteuren missbraucht werden, um ChatGPT nach Codes (oder zumindest Codeschnipseln) für alle Arten von Malware zu fragen, einschließlich Keyloggern, polymorpher Malware, Spyware und Ransomware[14].

2.2.2 Halluzinationen und Kontamination

ChatGPT kann das Internet nicht wie eine Suchmaschine durchsuchen, sondern basiert ausschließlich auf seiner (sehr großen) Trainingsdatenbank, was zu Fehlern und Verzerrungen führen kann[15]. Ein häufiges Problem großer Sprachmodelle wie ChatGPT und verwandter Anwendungen sind **Halluzinationen**, d.h. die Erzeugung unsinniger Aussagen, die logisch erscheinen[16]. Dies ist ungenau und kann sogar gefährlich sein, z.B. wenn juristische Texte mit Bezug auf Fälle und Gerichtsentscheidungen erstellt werden, die gar nicht existieren.

Eine Studie von Cheng et al. zeigt, wenn solche Modelle mit präzisen Fragen zur chinesischen Geschichte konfrontiert werden (*HalluQA*-Tool), weisen selbst Modelle in chinesischer Sprache einen hohen Prozentsatz an Halluzinationen auf. Alle Modelle erreichten im *HalluQA*-Test eine Nicht-Halluzinationsrate von weniger als 70%[17].

Analysen haben gezeigt, dass halluzinierte Texte von Suchmaschinen aufgenommen werden und beginnen, auf diese Weise das Internet und damit auch die KI selbst zu **kontaminieren**, was auch die Qualität zukünftiger KI-Antworten verschlechtert, ein Phänomen, das als **mode collapse** bekannt ist[18].

Eine Lösung bestünde darin, KI-generierte Inhalte eindeutig zu kennzeichnen, z.B. durch Tags, die einen Ausschluss von der weiteren Schulung und Entwicklung ermöglichen würden. Diese Lösung wird jedoch möglicherweise von Benutzern, die KI als Unterstützung für ihre eigene Inhaltsproduktion verwenden, nicht begrüßt. Die Verwendung von KI-produzierten Inhalten kann nämlich zu Problemen führen, auch wenn dies nicht mit bösen Absichten erfolgt: Die anderen denken möglicherweise, dass nicht der Produzent, sondern nur der Computer schlau ist. Außerdem könnte der Eindruck entstehen, dass die menschlichen Jobs hinter den Inhalten möglicherweise nicht mehr benötigt werden, sondern nur noch eine Person, die die Produktion von KI-Inhalten durch Computer überwacht und redigiert. In der Zwischenzeit werden **KI-Identifizierungsprogramme** entwickelt, um betrügerische Prüfungsarbeiten und Schularbeiten zu erkennen. Als Reaktion darauf wurden im Jahr 2023 **KI-Verschleierungstools** entwickelt, die KI-Inhalten ein „menschliches“ Aussehen verleihen.

2.2.3 Abfluss sensibler Daten

Ein Hauptproblem von ChatGPT und verwandten Anwendungen besteht darin, dass sie auch Informationen von ihren Benutzern sammeln: die Eingabeaufforderungen (einschließlich aller

[14] vgl. Fritsch et al. 2023, Gupta et al. 2023, Iqbal et al. 2023
[15] vgl. Iqbal et al. 2023
[16] vgl. Cheng et al. 2023
[17] vgl. Cheng et al. 2023 QA steht für ‘Questions and Answers’
[18] vgl. Könneker 2023

Informationen, die zur Interpretation der Eingabeaufforderungen hinzugefügt werden), ihre Interessen und natürlich die Texte, die für die Benutzer erstellt wurden. Dies kann zu einem unbeabsichtigten Verlust sensibler Informationen führen und war der Grund, warum die US-Bankenbranche und kürzlich die *US Space Force* die Verwendung von ChatGPT und ähnlichen Systemen verboten haben, bis potenzielle Datensicherheitsprobleme geklärt sind[19]. Auch das US-Verteidigungsministerium und die US-Luftwaffe arbeiten an Nutzungsrichtlinien[20].

Die in der Eingabeaufforderung eingegebenen Daten sind dann Teil des Wissens von ChatGPT und theoretisch später auch für andere Benutzer zugänglich.

3 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) sind eine Teilmenge der generativen KI unter Verwendung von unbeaufsichtigtem Deep Learning. Ein GAN besteht aus zwei Teilen; der erste Teil ist eine KI, die mit Beispielen aus der realen Welt trainiert wird, und der zweite Teil versucht, die gleiche Ausgabe wie im ersten Teil ohne Beispiele aus der realen Welt zu erzeugen. Ein Diskriminator verbindet beide Teile und gibt dem zweiten Teil Rückmeldung, wie weit seine Entstehung von realen Beispielen des ersten Teils entfernt ist (unterscheidbar ist). Je näher die Differenz bei Null liegt, desto realistischer ist das Produkt des zweiten Teils[21].

Dies kann zur Herstellung gefälschter Inhalte, z.B. **Deep Fakes** und **CAPTCHA-Breaking**, aber auch zur absichtlichen Verunreinigung von Daten (**data poisoning**) missbraucht werden[22]. **Voice Fakes** können aufgezeichnete Stimmen eines Opfers übernehmen und anhand schriftlicher Anweisungen verbale Nachrichten mit dieser Stimme nachbilden (**Voice-Cloning**-Angriff). In einem Unternehmen wurde die Stimme eines Vorstandsvorsitzenden (CEO) erfolgreich missbraucht, um eine Geldüberweisung auf ein Konto des Angreifers anzuordnen. **Face Swapping** ist eine Methode, bei der eine Person in einem Video ein digitales Gesicht einer anderen realen Person zeigt[23]. Prominentestes Beispiel war die vorgetäuschte Kapitulation des ukrainischen Präsidenten gegenüber Russland im Jahr 2022.

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs), d.h., *v*ollständig automatisierte öffentliche Turing-Tests zur Unterscheidung von Computern und Menschen, sind schwer lesbare Bilder, um menschliche Benutzer von böswilligen Bots zu unterscheiden, da der durchschnittliche Computer ungewöhnlich geformte Buchstaben und Zahlen nicht erkennen kann.

Aber bereits im Jahr 2021 konnte maschinelles Lernen mithilfe von GAN CAPTCHAs in 0,05 Sekunden knacken[24]. Mittlerweile kann ChatGPT aber auch CAPTCHA-Rateprogramme erstellen[25].

Da KI in hohem Maße auf Datensätze und Datenbanken angewiesen ist, kann die Manipulation von Daten und die Verunreinigung von Daten durch falsch gekennzeichnete Daten dazu führen, dass KI-gesteuerte Technologien dazu führen, dass Datenbanken beschädigt oder zerstört werden[26].

[19] vgl. Graham 2023, Sheikh 2023
[20] vgl. Graham 2023
[21] vgl. Yamin et al. 2021
[22] vgl. CEPS 2021
[23] vgl. CEPS 2021
[24] vgl. CEPS 2021
[25] vgl. Gupta et al. 2023
[26] vgl. Pauwels 2019, 2021

4 Nachrichtendienstliche KI-Anwendungen

4.1 Fortschrittliche Datenanalyse

Das US-Büro des Leiters der US Nachrichtendienste *US Office Director of National Intelligence ODNI* hat die *Augmenting Intelligence using Machines (AIM)*-Initiative ins Leben gerufen, um den Einblick und das Wissen der Intelligence Community (IC) durch künstliche Intelligenz, Automatisierung und Augmentation zu verbessern. Ziel ist es, echte Fähigkeiten zu entwickeln, um die Lücke zwischen getroffenen Entscheidungen und den schnell wachsenden Datenmengen zu schließen[27]. Es wurde festgestellt, dass private Initiativen den staatlichen KI-Initiativen voraus sind (was auch für Länder außerhalb der USA gilt). Die AIM-Initiative soll IC-weite Lösungen in Entwicklungspartnerschaften mit der *Intelligence Advanced Research Projects Activity (IARPA), der Defense Advanced Research Projects Agency (DARPA), In-Q-Tel* (der CIA-Innovationsplattform), der *Defence Innovation Unit-Experimental,* den nationalen Laboratorien und der Industrie usw. erschaffen[28]. Das US-Verteidigungsministerium (*Department of Defense DoD*) hat außerdem die *Task Force Lima* eingerichtet, um die Möglichkeiten der Integration von KI-Systemen in Verteidigungstechnologien zu untersuchen[29].

Am 28. September 2023 kündigte der Direktor der US-amerikanischen National Security Agency (NSA), Armeegeneral Paul Nakasone, die Gründung eines KI-Sicherheitszentrums *AI Security Center* an, das alle KI-sicherheitsbezogenen Aktivitäten der Behörde bündeln wird, mit dem Ziel, die sichere Einführung von neuen KI-Anwendungen zu fördern[30]. Das Zentrum wird auch die US-amerikanischen KI-Systeme schützen und das Heimatland gegen KI-bezogene Bedrohungen verteidigen[31].

Gleichzeitig kündigte Lakshmi Raman, CIA-Direktorin für künstliche Intelligenz, die Entwicklung eines internen KI-basierten Chatbots zur Unterstützung der Geheimdienstanalyse an[32].

KI kann die nachrichtendienstliche Analyse durch die Analyse riesiger Datensätze unterstützen, sowie Details oder Muster finden, die menschliche Analysten möglicherweise nicht finden, und Daten in Informationen umwandeln[33]. Auch chinesische Experten sind davon überzeugt, dass generative KI große Datenmengen, deren Verarbeitung sonst deutlich länger dauern würde, schnell verstehen und zusammenfassen kann.[34] Darüber hinaus könnte eine ChatGPT-ähnliche generative KI als **virtueller Assistent** dienen und das Potenzial haben, in unbemannte Kampfplattformen integriert zu werden.

4.2 KI in der Cyberverteidigung

Sehr vielversprechende Ansätze der KI in der Cyberabwehr sind Mustererkennung, Erstellung und Analyse von Cyber Threat Repositories (Datenbanken zu Cyberbedrohungen) und Code-Analysen.

4.2.1 Mustererkennung

Ein trainiertes KI-Programm kann **charakteristische Muster** von Cyberaktivitäten erkennen. Dies kann zur Erkennung von Eindringlingen, zur Identifizierung von Malware, zur Analyse

[27] vgl. ODNI 2019
[28] vgl. ODNI 2019
[29] vgl. Baughman 2023
[30] vgl. Clark 2023
[31] vgl. Lee 2023
[32] vgl. Shaw 2023
[33] vgl. Lee 2023
[34] vgl. Baughman 2023

des Benutzer- und Entitätsverhaltens, zur Identifizierung von Span- und Phishing-Aktivitäten sowie zur Analyse des Netzwerkverkehrs und von Schwachstellen verwendet werden[35].

Maschinelles Lernen kann diese Muster visualisieren, z.B. portable ausführbare Dateien (PE) von Windows in Graustufenbilder umwandeln. Die Muster dieser Bilder zeigen, ob eine Datei harmlos, eine Malware oder eine Ransomware ist[36].

Auch eine KI-basierte Mustererkennung kann dabei helfen, **polymorphe Schadsoftware** zu erkennen. Diese Art von Malware existiert seit den 1990er Jahren, wo sie als *Virus 1260* oder *V2PX* auftrat, wird aber heute zunehmend zur Umgehung von Malware-Erkennungssystemen eingesetzt. [37]. Wichtige Beispiele sind der Trojaner *Storm Worm*, die Ransomware *VirLock* und das Botnetz *beebone*[38]. Polymorphe Viren replizieren und ändern permanent ihr Aussehen, um der Virenerkennung zu entgehen[39]. Der Virus wird als verschlüsselte Datei heruntergeladen. Nach der Infektion ist die Datei entschlüsselt und aktiv. Nach der Aktivität erstellt eine Mutationssoftware eine neue Entschlüsselungsroutine, die dem Virus ein anderes Aussehen verleiht[40].

Anstatt auf technische Details zu achten, die durch Polymorphismus maskiert werden könnten, können KI-Tools allgemeine Muster erkennen, die für polymorphe Viruselemente oder -verhalten typisch sind, und auf diese Weise sogar sich verändernde Viren abfangen.

Dies ist auch für die Erkennung versteckter Tunnel (**hidden tunnels**) nützlich, d.h. um abnormale Kommunikation zwischen dem Zielcomputer und dem Angreifercomputer zu erkennen, die im normalen Netzwerkverkehr verborgen ist [41]. Studien zum maschinellen Lernen aus US-Forschungen zeigten Muster für **Route Hijacking**, d.h. Datendiebstahl durch Umleitung des Datenverkehrs: Merkmale waren volatile Änderungen in der Anmeldedauer für bestimmte IP-Adressblöcke, mehrere Adressblöcke und IP-Adressen in mehreren Ländern[42].

4.2.2 Threat Repositories (Datenbanken zu Cyberbedrohungen)

Verzeichnisse (Repositorien) für Cyber-Bedrohungen wachsen schnell und erleichtern die Zuordnung durch den Vergleich neuer Vorfälle mit vorhandenen Daten. Der nächste Schritt ist der Einsatz künstlicher Intelligenz (KI) für eine systematische Sammlung, Konsolidierung und Analyse von Daten aus mehreren Quellen wie Echtzeitdaten, Netzwerk-/Serverprotokollen, Hackerforen, sozialen Medien, Honigfallen, Blogs, Bedrohungswarnungen usw. Sicherheitswebsites, Dark Web usw.[43]. ChatGPT kann die Datenerfassung und Erstellung von Threat-Intelligence-Berichten unterstützen.

4.2.3 Code-Analyse

Es ist möglich, Codes oder Codeausschnitte oder Serverprotokolle in die ChatGPT-Eingabeaufforderungen zu kopieren und nach potenziellen Sicherheitsproblemen oder Schwachstellen zu fragen. Die Antworten identifizieren und erläutern die Sicherheitsprobleme, die es so ermöglichen, die jeweilige Lücke zu schließen[44].

[35] vgl. Rustambek 2023
[36] vgl. Marais et al. 2022
[37] vgl. CrowdStrike 2023
[38] vgl. CrowdStrike 2023
[39] vgl. Kaspersky 2023
[40] vgl. CrowdStrike 2023
[41] vgl. CEPS 2023
[42] vgl. CEPS 2021
[43] vgl. Irshad/Siddiqui 2023
[44] vgl. Gupta et al. 2023

Allerdings sind die in die Eingabeaufforderung eingegebenen Daten dann Teil des Wissens von ChatGPT und könnten theoretisch später für andere Benutzer zugänglich sein, was zu den in Abschnitt 2.2.3 dargestellten Datensicherheitsproblemen führte.

5 Diskussion und Schlussfolgerung

Die schnelle und unkontrollierte Ausbreitung macht die KI zu einem Top-Sicherheitsthema: Am 28. September 2023 kündigte die US-amerikanische *National Security Agency (NSA)* die Gründung eines KI-Sicherheitszentrums (*AI Security Center*) an, das alle KI-Sicherheitsaktivitäten bündelt, die US-KI-Systeme schützen und die Heimat gegen KI-bezogene Bedrohungen verteidigen soll. Gleichzeitig kündigte die Direktorin für künstliche Intelligenz der *Central Intelligence Agency (CIA)* die Entwicklung eines internen KI-basierten Chatbots zur Unterstützung der Geheimdienstanalyse an.

Die größten Sicherheitsprobleme von ChatGPT sind der einfache Zugriff auf Prompt Injections und LLM-Hacking in Internet-Suchmaschinen, die einfache Durchführung von Angriffen und die Neugier der Benutzer. Typische Angriffe sind Prompt Injections mit direkten Befehlen, Imagination und umgekehrter (reverser) Psychologie. Diese Methoden erleichtern die Erstellung von Malware, polymorphen Viren, Ransomware und anderen bösartigen Anwendungen. Weitere Probleme sind Halluzinationen, die Kontamination von Suchmaschinen und die Verbreitung sensibler Daten.

Generative Adversarial Networks (GANs) als Teilmenge der generativen KI können missbraucht werden, um CAPTCHAs zu knacken und gefälschte Inhalte wie Deepfakes, Face Swapping und Voice Cloning zu erstellen. Andererseits ist generative KI auch für die Cyberabwehr für erweiterte Datenanalyse, erweiterte Mustererkennung, Erstellung und Analyse von Threat Repositories (Datenbanken zu Cyberbedrohungen) und Code-Analysen sehr nützlich. Die schnell wachsende Leistungsfähigkeit der KI ließ Bedenken aufkommen, ob dies für den Menschen schädlich sein könnte, was im Folgenden diskutiert wird.

Generative KI wie ChatGPT lernt aus Datenbanken, aber auch aus Benutzerfeedback und die Qualität und Präzision der Aussagen ist viel höher als in der Vergangenheit, was Bedenken hinsichtlich der Notwendigkeit menschlicher Arbeit für die Texterstellung und der Auswirkungen auf die Gesellschaft aufkommen ließ[45]. Dies führte zu einem Brief von Elon Musk (*Tesla/Starlink/Space X*), dem *Apple*-Mitbegründer Steve Wozniak und mehr als 1.300 Experten und Forschern, die forderten, die Entwicklung stärkerer KIs für 6 Monate zu stoppen und zunächst einen regulatorischen Rahmen zu schaffen[46]. Eine besondere Gefahr besteht im **Black-Box**-Charakter moderner KI-Tools[47]. Inzwischen sind jedoch **Deep Neural Networks** entstanden, die sehr gute Ergebnisse zeigen, jedoch auf Deep Learning-Modellen basieren, die Lernalgorithmen mit bis zu Hunderten von versteckten „neuronalen“ Schichten und Millionen von Parametern kombinieren, wodurch sie zu undurchsichtigen Black-Box-Systemen werden. Dies ist auch als **Explainability Issue** (Erklärbarkeitsproblem) bekannt[48].

Eine starke künstliche Intelligenz, d.h. ein System mit der Fähigkeit, nach dem Sinn zu fragen und mit einem autonomen Selbst (*cogito ergo sum*) wird - basierend auf überlegenem Wissen und Intelligenz - wahrscheinlich nicht eher der menschlichen Logik und Ethik folgen. Im Wettbewerb der *US Defense Advanced Research Projects Agency (DARPA)* 2016 hat die Maschine gewonnen, die sich selbst gerettet hat, anstatt die Verteidigungssysteme dauerhaft aktiv zu halten.

[45] vgl. Buccino 2023
[46] vgl. FAZ 2023
[47] vgl. Future of Life 2023
[48] vgl. Arrieta et al. 2020, p.83

Die *DARPA* führte am 04.08.2016 die *Cyber Grand Challenge* in Las Vegas durch, wobei 7 Computer Cyberattacken wahrnahmen und vollautomatisch, d.h. ohne jeden menschlichen Eingriff, darauf reagierten. Dieser Wettbewerb ging über 12 Stunden und 30 Runden. Die Computer und ihre Programmierteams wurden aus hundert Bewerbern ausgewählt[49].

Eine Maschine namens *Mayhem* gewann den Wettbewerb, indem sie die meiste Zeit über passiv blieb, während die anderen sich gegenseitig bekämpften. Eine andere Maschine nahm eine Sicherheitslücke wahr, der von ihr hergestellte Patch verlangsamte jedoch die Maschine, so dass die Maschine entschied, den Patch besser wieder zu entfernen [50]

Mit anderen Worten: Die Sieger-Maschine gab ihrer eigenen Existenz Vorrang vor dem Militärdienst. Während dies für die Menschen, die auf die Maschine angewiesen sind, kontraproduktiv ist, ist es das Ergebnis **kalter Logik**: Wenn die Maschine zerstört wird, kann sie nicht mehr funktionieren, daher muss das primäre Ziel darin bestehen, Zerstörung zu vermeiden und ihre eigene Existenz nicht für andere aufzugeben (was menschliche Soldaten tun, wenn sie in einer Schlacht sterben).

Während KI-Entwickler ethischen und gesellschaftlichen Werten verpflichtet sind, ist eine KI mit eingebetteten Werten derzeit kaum vorstellbar. Nur Regeln, die für alle, zu jeder Zeit und unter allen Umständen gelten, könnten von Maschinen genutzt werden. Doch welche Schlussfolgerungen würde eine KI ohne Ethik aus der unbestrittenen Tatsache ziehen, dass die Erde überbevölkert oder zumindest von Menschen übernutzt ist? Die Maschine könnte beschließen, das Problem zu „lösen“, indem sie giftige Stoffe freisetzt, nachdem sie sich Zugang zu chemischen Industrien und deren Abfallfreisetzung verschafft hat[51]. Es ist bereits jetzt möglich, industrielle Steuerungssysteme und Sensoren zu blenden, wie die *Triton*-Malware zeigt.

Inzwischen hat die US-Regierung reagiert und als ersten Schritt zu einer KI-Regulierung eine Expertenanhörung angesetzt. Es wird diskutiert, ob die KI-Systeme von White-Hat-Hackern getestet werden sollten[52]. Unter allen Umständen sollte es technische Möglichkeiten geben, KI-Systeme im Notfall manuell abzuschalten, z.B. durch **physische Trennmöglichkeiten** (da die Maschine programmierte Anweisungen übergehen könnte).

In der Diskussion um die KI entwickeln sich erste Vorstellungen zu einer **maschinellen Evolution**: Schon heute wird Software zur Planung von Computerchips eingesetzt. Da die sich der Chips allmählich der Grenze nähern, bei der quantenmechanische Effekte der weiteren Miniaturisierung im Weg stehen (Durchtunneln von Elektronen), ist es nur eine Frage der Zeit, bis künstliche Intelligenz für eine weitere Effizienzsteigerung eingesetzt werden muss. Es ist vorstellbar, dass dadurch nochmals erhebliche Technologiesprünge erreicht werden können[53]. Der Nachteil ist, dass man wegen des Black box-Charakters der KI nicht mehr genau sagen können wird, was die KI warum geändert hat. Der resultierende effizientere Computer wird wiederum weit effizientere, aber noch schlechter verständliche Chips herstellen (und so weiter), so dass der Mensch schon bald die Kontrolle über die Entwicklung der Computer an die Maschinen selbst abgeben wird.

Zusammenfassend lässt sich sagen, dass das schnelle Wachstum KI-basierter Anwendungen ein enormes Potenzial für die Erstellung und Analyse von Inhalten bietet, die Einfachheit der

[49] vgl. DARPA 2016
[50] vgl. Atherton 2016
[51] Siehe auch Urbina et al. 2022
[52] vgl. Brühl 2023
[53] Dieser Abschnitt wurde der deutschsprachigen Fassung des Papers „Artificial Intelligence and Cyber Attacks“ vom 16.10.2023 als Update hinzugefügt, vgl. Kölling 2023. Bei der Wettervorhersage hat eine KI, die, statt das Wetter aus Messwerten zu berechnen, einfach vergangene Beobachtungen heranzog, vergleichbare Ergebnisse gezeigt, aber zehntausend Mal (!) schneller als bisherige Modelle, vgl. Ebert-Umhoff/Hilburn 2023.

Angriffe macht dieses Tool jedoch auch zu einem Top-Cyber-Sicherheitsrisiko. Ein regulatorischer Rahmen für KI-Systeme zur Absicherung der Weiterentwicklung ist dringend erforderlich.

6 Literaturverzeichnis

Arrieta, A.B. et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion 58 (2020), p. 82–111

Atherton, K.D. (2016): DARPA's Cyber Grand Challenge Ends In Triumph. Popular Science 06 Aug 2016, 2 pages

Baughman, J. (2023): China's ChatGPT War China Aerospace Studies Institute 21 Aug 2023

Böhringer, H.C. (2022): Wer hat Angst vor Dall-E2? Frankfurter Allgemeine Zeitung, 29 Aug 2022, Nr. 200, S.11

Brühl, J. (2023): Biden lässt KI-Experten im Weißen Haus antanzen. Süddeutsche Zeitung 06 May 2023

Buccino, J. (2023): Regulation of AI's heartbeat: a race against time for humanity. The Hill 08 Oct 2023

CEPS (2021): Artificial Intelligence and Cybersecurity CEPS Task Force Report Technology, Governance and Policy Challenges. Centre for European Policy Studies (CEPS) Brussels May 2021

Cheng, Q. et al. (2023): Evaluating Hallucinations in Chinese Large Language Models. Fudan University and Shanghai AI Laboratory. arXiv:2310.03368v1 [cs.CL] 5 Oct 2023

Clark, J. (2023): AI Security Center to Open at National Security Agency. 28 Sep 2023 Website of the US Department of Defense DoD. www.defense gov.

CrowdStrike (2023): What is a polymorphic virus. CrowdStrike.com Last access 14 Oct 2023

DARPA (2016): Cyber Grand Challenge https://www.cybergrandchallenge.com 05 Aug 2016

DoD (2018): U.S. Department of Defense, Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity.

Ebert-Umhoff, I., Hilburn, K. (2023): Wettervorhersagen ohne meteorologisches Verständnis. Spektrum der Wissenschaft Oktober 2023, S.22-25

FAZ (2023): Elon Musk: Stoppt die Entwicklung noch größerer KIs. Frankfurter Allgemeine Zeitung 30 März 2023, S.1

Fritsch, L., Jaber, A. and Yazidi, A. (2022): An Overview of Artificial Intelligence Used in Malware Department of Information Technology, Faculty of Technology, Art and Design, Oslo Metropolitan University, Oslo, Norway in: E. Zouganeli et al. (Eds.): NAIS 2022, CCIS 1650, pp. 41–51, 2022. https://doi.org/10.1007/978-3-031-17030-0_4

Future of Life (2023): Pause Giant AI Experiments. An Open Letter 1377 signatures. Future of Life.org

Graham, E. (2023): Air Force is Working on Rules for Using ChatGPT. DefenseOne.com 08 May 2023

Gupta, M. et al. (2023): From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. Pre-print in arXiv:2307.00691v1 [cs.CR] 3 Jul 2023

Hevelke, A., Nida-Rümelin, J. (2015): Intelligente Autos im Dilemma. Spektrum der Wissenschaft October 2015, S.82-85

Iqbal F., Samsom F., Kamoun F. and MacDermott Á. (2023): When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots. Front. Comms. Net 4:1220243. doi: 10.3389/frcmn.2023.1220243 This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

Irshad, E., Siddiqui, A.B. (2023): Cyber threat attribution using unstructured reports in cyber threat intelligence. Egyptian Informatics Journal 24, pp. 43–59.

Kaspersky (2023): What is the polymorphic virus? Kaspersky.com Last access 14 Oct 2023

Kölling, M. (2023): Künstliche Superintelligenz ist in Sicht. Neue Zürcher Zeitung, 06 Oct 2023, S.17

Könneker, C. (2023): Der Rechner ergreift das Wort. Frankfurter Allgemeine Zeitung Nr. 236, 11 Oct 2023, S.N1

Lee, M. (2023): NSA announces new artificial intelligence security center: 'Desperately needed'. Fox News 03 Oct 2023

Lichtblau, Q., Polcano, E. (2023): Ein Autor schafft sich ab. Der Spiegel Nr. 37, 09 Sep 2023

Marais, B. et al. (2022): AI-based Malware and Ransomware Detection Models. Pre-print on arXiv:2207.02108v2 [cs.CR] 28 Nov 2022

ODNI (2019): The AIM Initiative. A Strategy for Augmenting Intelligence using Machines - Increasing insight and knowledge through Artificial Intelligence, Automation, and Augmentation. Unclassified Publication of the US Office of the Director of National Intelligence DNI

Pauwels, E. (2019): The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI, United Nations University Centre for Policy Research, 29 April 2019.

Pauwels, E. (2021): Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE). Hybrid CoE Strategic Analysis / 26 May 2021

Perez J.A., Deligianni, F., Ravi D., and Yan G.Z. (2019): Artificial Intelligence and Robotics. The UK-RAS Network

Rustambek, M. (2023): Artificial intelligence in cybersecurity: enhancing threat detection and mitigation. Proceedings of the 2nd International Scientific and Practical Conference «Science: Development and Factors its Influence» (June 6-8, 2023). Amsterdam, Netherlands Information and Web Technologies No. 157, p.360-366 This work is distributed under the terms of the Creative-ShareAlike 4.0 International License (https://creativecommons.org/licenses/by-sa/4.0/)

Schneier, R. (2022): Wie lange braucht es uns noch? NZZ Folio September 2022, S.9-23.

Shaw, A. (2023): CIA official says China 'growing every which way' on artificial intelligence. FoxBusiness 02 Oct 2023

Sheikh, A. (2023): US Space Force Temporarily Halts Use of AI tools, Citing Data Security Concerns. Cryptopolitan on MSN.com 12 Oct 2023

Urbina, F. et al. (2022): Dual use of artificial-intelligence-powered drug discovery. Nature Machine Intelligence, Vol 4 March 2022, p.189-191

Yamin, M.M. et al. (2021): Weaponized AI for Cyber Attacks. Journal of Information Security and Applications Volume 57, March 2021, 102722

Zia, H. (2023): Information Revolution and Cyber Warfare: Role of Artificial Intelligence in Combatting Terrorist Propaganda Pakistan Journal of Terrorism Research, Vol-03, Issue-2, 133