Künstliche Intelligenz in der Cybersicherheit

Chancen und Risiken

SBA Research gemeinnützige GmbH (gGmbH)

Floragasse 7

1040 Vienna

https://www.sba-research.org/

Andreas Ekelhart, Georg Goldenits, Rudolf Mayer, Verena Schuster

Jänner 2025

Executive Summary

Cybersicherheit ist in einer zunehmend digitalen und vernetzten Welt eine zentrale Herausforderung. Das österreichische Bundesministerium für Landesverteidigung weist in seinem Bericht „Risikobild 2024“ neben militärischen Konflikten auch auf die Gefahren von Cyber-Angriffen hin. Eine Studie von KPMG aus dem Jahr 2024 zeigt, dass nahezu alle befragten österreichischen Unternehmen von Cyberangriffen, wie (Spear-) Phishing, Malware und Social Engineering, betroffen waren.

Angesichts der zunehmenden Komplexität und Vielfalt der Bedrohungen reichen traditionelle Sicherheitsansätze oft nicht mehr aus. Um diesem Problem zu begegnen, wird zunehmend über den verstärkten Einsatz von Künstlicher Intelligenz (KI) in der Cybersicherheit nachgedacht. KI ermöglicht es, große Datenmengen zu analysieren, Muster zu erkennen und Anomalien zu identifizieren, was einen erheblichen Vorteil bei der Früherkennung von Bedrohungen bieten kann. KI kann auch die Reaktion auf Angriffe automatisieren. KI wird aber auch von Angreifern genutzt, um Angriffe zu automatisieren, zu verbessern und zielgerichteter zu gestalten. Sabotage und Desinformation, insbesondere durch Deepfake-Angriffe, verdeutlichen auch die Gefahr durch den Einsatz von KI auf Angreiferseite.

Diese Kurzstudie untersucht die Chancen und Risiken von KI in der Cybersicherheit und beleuchtet sowohl Verteidigungsansätze als auch potenzielle Angriffsstrategien mit KI. Ergänzt wird dies durch Expertenmeinungen aus Industrie und Forschung zur aktuellen Nutzung von KI für beide Seiten.

Inhaltsverzeichnis

1 Einleitung

Das Bundesministerium für Landesverteidigung wies im *Risikobild 2024* [ABC+24] neben den Auswirkungen militärischer Konflikte auch auf die Gefahren von Cyberangriffen hin. Auch viele österreichische Unternehmen sind von Cyberangriffen betroffen. Das Wirtschaftsprüfungs- und Beratungsunternehmen KPMG[1] veröffentlicht dazu jährlich eine Studie zur Cybersicherheit in Österreich. Im Jahr 2022[2] gaben alle 903 befragten Unternehmen an, von einem Phishing-Angriff betroffen gewesen zu sein, im Jahr 2023 waren es 87 % von 1.158 befragten Unternehmen.

Angesichts der zunehmenden Komplexität und Vielfalt der Bedrohungen – von einfachen Phishing-Angriffen bis hin zu ausgefeilten Cyberoperationen – sind herkömmliche Sicherheitsansätze oft nicht mehr ausreichend[3]. Um dieses Problem zu lösen, wird zunehmend Künstliche Intelligenz (KI) als Unterstützung eingesetzt. Künstliche Intelligenz ist hierbei „die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren“[4]. KI ermöglicht es, große Mengen an Daten zu analysieren und dabei Muster zu identifizieren, die menschliche Analyst:innen leicht übersehen könnten. Insbesondere bei der Erkennung von Anomalien, bei denen Angreifer:innen ungewöhnliches Verhalten zeigen, bietet KI einen erheblichen Vorteil. Selbst unbekannte Angriffsvektoren können durch Algorithmen des Machine Learning (ML) dynamisch erlernt und identifiziert werden und unterstützen so die Früherkennung von Angriffen. KI ermöglicht darüberhinaus die Automatisierung der Reaktion auf Cyberangriffe.

Allerdings kann KI in der Cybersicherheit nicht nur zur Abwehr von Angriffen, sondern auch zur Unterstützung von Angriffen eingesetzt werden [KGK23]. Auch das vorher erwähnte *Risikobild 2024* weist explizit auf Cyberangriffe unter Einsatz von Künstliche Intelligenz (KI) hin. So dienen Cyberangriffe nicht mehr nur der finanziellen Erpressung oder dem Diebstahl von Daten, sondern auch der Durchsetzung politischer Interessen durch Sabotage oder gezielte Desinformation, z. B. durch Deepfakes. Diese Gefahr betrifft alle Medienkanäle, sowohl im Internet als auch klassische Kanäle wie Printmedien, Radio oder Fernsehen.

In dieser Forschungsstudie werden zunächst verschiedene Aspekte der Cybersicherheit beschrieben und darauf aufbauend die Chancen und Risiken von KI in der Cybersicherheit zusammengefasst (Abschnitt 2). Danach wird die allgemeine Bedrohungslandschaft in der Cybersicherheit diskutiert (Abschnitt 3). Anschließend werden unterschiedliche KI-basierte Verteidigungsansätze in Abschnitt 4 untersucht. Abschnitt 5 beschreibt Angriffsstrategien auf KI und potenzielle KI-basierte Angriffe. Abschließend werden in Abschnitt 6 Einschätzungen von Expertinnen und Experten aus dem Bereich der Cybersicherheit zu den Potenzialen und Gefahren von KI präsentiert.

Diese aus Interviews gewonnenen Einschätzungen zeigen, dass die Verwendung von KI in der Cybersicherheit die Bedrohungslandschaft verändert und neue Möglichkeiten sowohl im Angriff als auch in der Verteidigung bietet. Angreifer:innen profitieren von automatisierten Werkzeugen z. B. für die Erstellung von Malware oder Social-Engineering-Kampagnen, sowie von der Skalierbarkeit von Angriffen. Dadurch wird die Eintrittsbarriere für Angriffe massiv gesenkt. Verteidiger:innen profitieren auch von KI-basierten Mechanismen wie Anomalieerkennung und Verhaltensanalyse, stoßen dabei aber auf Herausforderungen wie hohe Kosten, regulatorische Anforderungen und die Abhängigkeit von der Datenqualität. Sprachmodelle wie LLMs zeigen Potenzial, z. B. in der Bedrohungsanalyse und zur Unterstützung von Sicherheitsanalytiker:innen, sind aber noch nicht voll ausgereift. Der Markt ist geprägt von unrealistischen Versprechungen und begrenzten praktischen Anwendungen. Unternehmen schrecken vor einer vollständigen Automatisierung zurück und haben Schwierigkeiten damit, neue Technologien in bestehende Systeme zu integrieren.

Zusammenfassend lässt sich sagen, dass KI enorme Möglichkeiten zur Effizienzsteigerung und Bedrohungserkennung bietet, jedoch innovative Lösungen, robuste Datenstrategien, sowie eine Anpassung an regulatorische Vorgaben erfordert, um ihr volles Potenzial für die Cybersicherheit zu entfalten.

[1] https://kpmg.com/at/de/home.html
[2] https://info.kpmg.at/cyber-security-2023/
[3] https://stage.itwelt.at/news/studie-herkoemmliche-sicherheitsansaetze-nicht-mehr-ausreichend/
[4] https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt

2 Chancen und Risiken von KI für die Cybersicherheit

Im Folgenden wird ein Überblick über die Chancen und Risiken des Einsatzes von Künstlicher Intelligenz im Bereich der Cybersicherheit gegeben.

2.1 Vorteile des Einsatzes KI-gestützter Techniken gegenüber herkömmlichen Techniken

Zwei Vorteile des Einsatzes von KI-Techniken für die Cybersicherheit sind die Geschwindigkeit und Skalierbarkeit, da eine KI Bedrohungen in Echtzeit auch in großen oder komplexen Systemen erkennen kann [WDCP22, CPV22, SS21b]. Darüber hinaus kann KI bspw. bei der Erkennung von Spam- [WSM21, GMBV21] oder Phishing-E-Mails [GDSDBV+20, NNN18] unterstützen.

Einige KI-Techniken ermöglichen auch adaptives Lernen, um neue Bedrohungen, die zum Zeitpunkt der Entwicklung der KI noch nicht bekannt waren, zu adressieren. Ein Alarm kann nach dem Schweregrad der potenziellen Bedrohung priorisiert werden [DTN21, DMCF20] und darauf basierend ein Handlungsplan automatisch erstellt werden. Zusätzlich bieten Large Language Models (LLMs) neue Möglichkeiten zur Unterstützung von Cybersicherheitsaufgaben, wie z. B. die Erkennung von Schwachstellen in Anwendungen oder die Unterstützung von Sicherheitstests [YDX+24]. So kann der Einsatz von KI nicht nur zu einer schnelleren Angriffserkennung führen, sondern potenziell auch die Kosten senken, da KI Zeit sparen und Expertinnen und Experten entlasten kann [RDRB11, PW19, PZ21, ZALT19].

2.2 Risiko beim Einsatz von KI

Grundsätzlich handelt es sich bei KI um algorithmen- und datenbasierte Computerprogramme, die durch Programmierfehler, Fehler im Training oder den Daten, oder gezielte Angriffe anfällig für Fehlfunktionen sein können. Zudem ist zu berücksichtigen, dass komplexe Entscheidungen oft nicht mehr vollständig von menschlichen Entwickler:innen nachvollzogen werden können [ABC+24], was Auswirkungen auf die Verständlichkeit der Handlungsempfehlungen sowie die Auditierbarkeit haben kann. Obwohl KI ein erhebliches Potenzial zur Erhöhung der Cybersicherheit von Computersystemen bietet, kann die KI selbst auch zum Angriffsvektor werden. Mögliche Angriffe auf KI sind *Evasion, Poisioning*, oder *Privacy Attacks* sowie *Abuse Attacks*, die in Abschnitt 5.2 näher beschrieben werden.

Darüber hinaus kann KI nicht nur zur Verteidigung, sondern auch direkt für Angriffe eingesetzt werden. Beispielsweise kann sie zur Erstellung sehr realistischer und personalisierter Phishing-Attacken [HSV+23] oder zur Generierung von Deepfakes genutzt werden. Davor warnen etwa das Bundesministerium für Landesverteidigung in Österreich als auch das deutsche Bundesamt für Sicherheit in der Informationstechnik[5]. Nicht nur Privatpersonen, sondern auch Unternehmen können von Deepfakes betroffen sein. So wurde beispielsweise der Finanzvorstand des britischen Unternehmens *Arup* durch ein Deepfake-Artefakt dazu verleitet, mehrere Überweisungen in Höhe von insgesamt 25 Millionen Dollar zu tätigen[6]. Darüber hinaus können LLMs zum Erraten von Passwörtern (PassGPT [RPCH24]) oder zur Generierung von Malware [PPTK+23] verwendet werden. Weitere Details zu Angriffen mit LLMs sind in Abschnitt 5.1 beschrieben.

3 Bedrohungslandschaft in der Cybersicherheit

In diesem Kapitel werden relevante Frameworks und Taxonomien vorgestellt, um wesentliche Funktionen und Bedrohungen aus der Cybersicherheit zu beschreiben. Welcher dieser Konzepte mit KI-Methoden unterstützt werden können, werden in den Abschnitten 4 und 5 behandelt.

[5] https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html

[6] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

3.1 NIST Cybersecurity Framework

Das NIST Cybersecurity Framework (CSF) ist ein umfassender Leitfaden für das Management von Cyberrisiken, entwickelt vom National Institute of Standards and Technology (NIST) in den USA[7]; es dient als freiwilliger Standard zur Verbesserung der Cybersicherheit. Das Framework ist modular aufgebaut, erleichtert einen systematischen Ansatz zur Risikominderung und fördert die Implementierung von Best Practices und allgemein anerkannten Standards der Industrie wie ISO/IEC 27001.

Das Framework umfasst sechs Funktionen (siehe Abbildung 1). *Identify* und *Protect* helfen, einen Cyberangriff zu verhindern, während *Detect*, *Respond* und *Recover* nach einem Cyberangriff erforderlich sind.

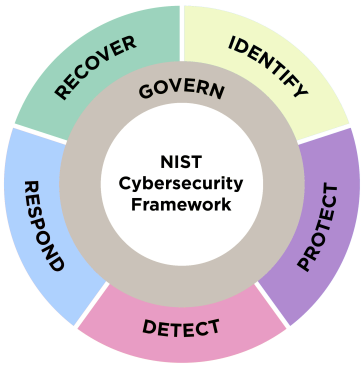

Abbildung 1: NIST Cybersecurity Framework 2.0

Govern unterstreicht die Wichtigkeit der Integration der Cybersicherheit in die *Governance* der jeweiligen Organisation und behandelt sie als ein zentrales Geschäftsrisiko neben finanziellen und operativen Risiken. Cybersicherheit ist aktiv zu überwachen und in die strategische Entscheidungsfindung, die Entwicklung von Richtlinien und die Zuweisung von Ressourcen einzubeziehen.

Identify bezeichnet das Verständnis der materiellen und immateriellen Vermögenswerte der Organisation (z. B. Daten, Hardware, Software, Systeme, Einrichtungen, Dienstleistungen und Personal), der Lieferant:innen sowie der damit verbundenen Cybersicherheitsrisiken. Es ermöglicht einer Organisation, ihre Aktivitäten im Einklang mit ihrer Risikomanagementstrategie zu priorisieren.

Protect unterstützt die Fähigkeit, die Vermögenswerte zu schützen, um die Wahrscheinlichkeit und die Auswirkungen von Cyberangriffen zu verhindern oder zumindest zu verringern.

Detect ermöglicht die rechtzeitige Erkennung und Analyse von Anomalien, Indicators of Compromise (IoC) und anderen potenziell schädlichen Ereignissen, die auf Cyberangriffe hindeuten. Diese Funktion bildet die Grundlage für *Respond* und *Recover*.

Respond unterstützt die Fähigkeit, die Auswirkungen von Cybersicherheitsvorfällen zu begrenzen.

Recover unterstützt die zeitnahe Wiederherstellung des Normalbetriebs, um die Auswirkungen von Cybersicherheitsvorfällen zu mindern und eine angemessene Kommunikation während der Wiederherstellungsmaßnahmen zu ermöglichen.

[7] https://www.nist.gov/cyberframework

3.2 Taxonomie von Cyberangriffen

Eine Taxonomie von Cyberangriffen ermöglicht die systematische Klassifizierung und Analyse von Cyberbedrohungen. In der Literatur wurden verschiedene Taxonomien entwickelt, die jeweils einen anderen Ansatz zur Klassifizierung von Cyberangriffen bieten. Diese Taxonomien bilden die Basis für die Zuordnung von KI-basierten Technologien.

Yao et al. [YDX+24] haben eine ebenenbasierte Taxonomie entwickelt, die Cyberangriffe in fünf Hauptgruppen (Ebenen) einteilt (siehe Abbildung 3): Angriffe auf dem *Hardware-Level* (Hardware-Ebene), *Operating System-Level* (Betriebssystem-Ebene), *Software-Level* (Software-Ebene), *Network-Level* (Netzwerk-Ebene) und *User-Level* (Benutzer:innen-Ebene). Simmons et al. [SES+09] haben die AVOIDIT-Taxonomie entwickelt (Abbildung 2), die sich auf die Angriffsvektoren (*Attack Vector*), die operativen Auswirkungen (*Operational Impact*), die Verteidigung (*Defense*), die Informationsauswirkungen (*Informational Impact*) und das Ziel (*Target*) konzentriert. Die ebenenbasierte Taxonomie [YDX+24] ähnelt dem Taxon *Target* der AVOIDIT-Taxonomie, welches die Kategorien *Operating System*, *Network*, *Local* (lokaler Computer), *User* und *Application* (Anwendung) definiert. Im folgenden werden diese zwei Taxonomien näher betrachtet.

3.2.1 AVOIDIT

Die AVOIDIT-Taxonomie (Abbildung 2) teilt Angriffe in verschiedene Dimensionen ein, die jeweils eine andere Eigenschaft oder Komponente des Angriffs beleuchten.

Attack Vector: Ein Angriffsvektor ist definiert als ein Weg, über den sich Angreifer:innen Zugang zu einem Computer oder einem Netzwerk verschaffen können, z. B. über Malware, Systemschwachstellen oder Social Engineering. Ein möglicher Angriffsvektor wäre eine *Misconfiguration* (Fehlkonfiguration) einer Anwendung oder eines Systems, durch die sich Angreifer:innen Zugang zu einem Netzwerk oder einem Computer verschaffen. Dieser Zugang kann dann für weitere Angriffe genutzt werden. Bei einem *Kernel Flaw* verschaffen sich Angreifer:innen Privilegien, um eine Sicherheitslücke auszunutzen. *Buffer Overflows* und *Insufficient Input Validation* (unzureichende Eingabevalidierung) können es Angreifer:innen z. B. ermöglichen, bösartigen Code auszuführen. Andere Angriffsvektoren inkludieren *Symbolic Link*-Sicherheitslücken, *File Descriptor-Schwachstellen* und *Race Conditions*.

Operational Impact: Diese Kategorie umfasst die Auswirkungen eines Angriffs auf die Funktionsfähigkeit eines Unternehmens oder einer Organisation. *Misuse of Resources* tritt auf, wenn Angreifer:innen unbefugten Zugriff auf IT-Funktionen erlangen. Verschiedene Arten von Kompromittierung (*Compromise*), wie zum Beispiel *User Compromise*, bei dem sich Angreifer:innen unberechtigten Zugriff auf ein Benutzer:innen-Konto verschaffen, *Root Compromise*, bei dem sich Angreifer:innen administrative Rechte verschaffen, oder *Web Compromise*, bei dem Schwachstellen wie Cross-Site-Scripting (XSS) oder SQL-Injections ausgenutzt werden, um Kontrolle über Web-Applikationen zu erlangen.

Defense: Diese Kategorie umfasst die beiden Verteidigungsstrategien Schadensbegrenzung (*Mitigation*) und Schadensbeseitigung (*Remediation*). Schadensbegrenzung kann z. B. dadurch erreicht werden, dass infizierte Hosts entfernt werden oder nur Hosts zugelassen werden die auf einer *Whitelist* stehen. Hinweise und Empfehlungen zur Eindämmung eines Angriffs oder zur Behebung einer Schwachstelle, z. B. aus einer Schwachstellendatenbank (*Vulnerability Database*), dienen dazu, potenzielle Schwachstellen frühzeitig zu beheben (*Reference Advisements*). Für eine Schadensbeseitigung (*Remediation*) kann ein Systempatch oder eine Codekorrektur (*Code Correction*) veröffentlicht werden.

Informational Impact: Diese Kategorie umfasst verschiedene Auswirkungen auf sensitive Informationen, die nach einem Angriff auftreten können. Dazu gehören die Verzerrung (*Distortion*) von Daten, etwa wenn Dateien manipuliert werden, die Unterbrechung (*Disruption*) eines Services oder eines Datenzugriffs, beispielsweise durch Denial-of-Service Angriffe, sowie die Vernichtung (*Destruction*) von Daten, etwa durch Löschung oder den Entzug des Zugriffs. Außerdem können vertrauliche Informationen offengelegt (*disclosed/discovered*) werden.

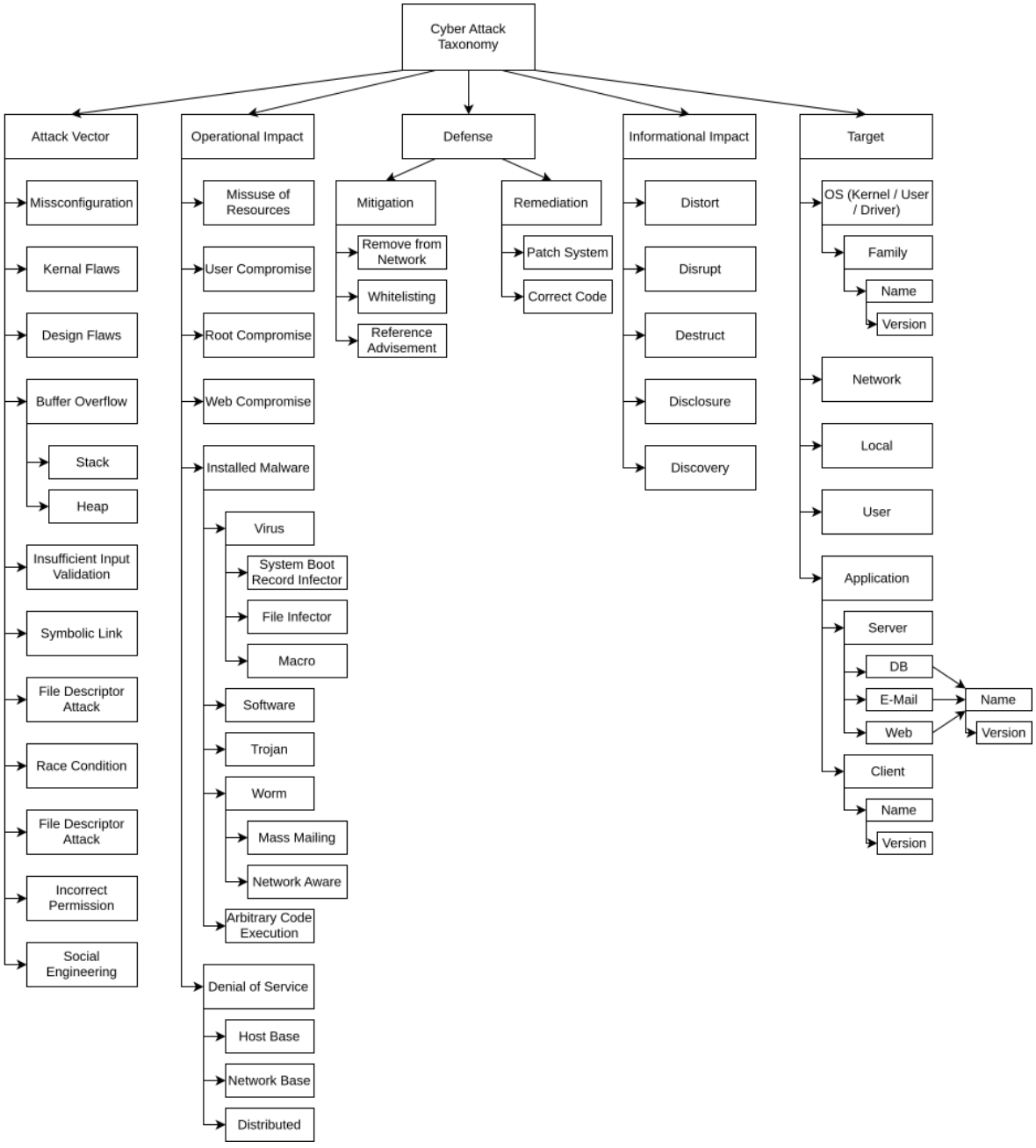

Abbildung 2: AVOIDIT Taxonomie für Cyberattacken von [SES+09]

Target: Das Ziel eines Cyberangriffs kann ein Betriebssystem (*Operating System*), ein Netzwerk (*Network*), ein lokaler Computer, ein:e bestimmte:r Benutzer:in (*User*), oder eine Anwendung (*Application*) sein.

3.2.2 Level-Based Taxonomy

Im Folgenden werden die Kategorien der Taxonomie nach Yao et al. [YDX+24] näher beschrieben (Abbildung 3). Bisher bekannte LLM-basierte-Angriffe (siehe Abschnitt 5.1) werden hier farblich hervorgehoben.

Hardwareebene: Angriffe auf der Hardwareebene umfassen Schadprogramme (*Malicious Firmware*), Fehlereinspeisungen (*Fault Injections*) oder Hardware-Side-Channel Angriffe.

Betriebssystemebene: Angriffe auf das Betriebssystem (OS) umfassen Rechteausweitung (*Privilege Escalation*), Remote Code-Execution (RCE), Side-Channel-Angriffe, Denial-of-Service (DoS), Speicherangriffe (*Memory Attack*) und Rootkits.

Softwareebene: Diese Ebene beinhaltet Buffer Overflow (BOF)-Angriffe, Race Conditions, Software Side-Channel Attacks, Software DoS, Software Exploration oder Malware Angriffe, einschließlich Brute-Force-, Worm-, Keylogger-, Ransomware- und Fileless-Angriffe.

Netzwerkebene: Angriffe auf das Netzwerk umfassen Cookie-Diebstahl (*Cookie Theft*), Directory Traversal, SQL-Injection, Cross-Site-Scripting (XSS) und Cross-Site-Request-Forgery (CSRF)-Angriffe, Web-DoS, webbasierte Side-Channel-Angriffe, Fingerprinting-Angriffe und Phishing-Angriffe und automatisiertes Lösen von CAPTCHAs.

Benutzer:innenebene: Diese Ebene beinhaltet z. B. Social Engineering oder Tailgaiting, bei dem sich Angreifer:innen physischen Zugriff zu einem Zielobjekt verschaffen.

4 KI-basierte Verteidigung

In diesem Abschnitt werden KI-basierte Verteidigungen vorgestellt, die auf der Arbeiten von Kaur et al. [KGK23], die Verteidigungen anhand des NIST Cybersecurity Framework organisieren, und Gorment et al. [GSCK23], die Methoden nach dem verwendeten ML-Modell und der verwendeten Klassifizierungsmethode (*classification method*) gruppieren. Gorment et al. konzentrieren sich auf die Erkennung von Malware, während Kaur et al. Systeme zur Erkennung und Verhinderung von Cyberbedrohungen beschreiben. Darüber hinaus analysierten Kaur et al. verschiedene Möglichkeiten der Reaktion (*response*) und Wiederherstellung (*recovery*).

4.1 KI-basierten Verteidigung nach dem NIST Cybersecurity Framework

Kaur et al. [KGK23] klassifizieren Methoden anhand des NIST Cybersecurity Framework (siehe Abschnitt 3.1) in die Funktionen *Identify*, *Protect*, *Detect*, *Respond* und *Recover*, mit 23 Unterkategorien, die jeweils Prozesse in diesen Funktionen darstellen, wie in Abbildung 4 abgebildet. Darüber hinaus haben Kaur et al. verschiedene Anwendungsfälle für jeden Prozess angeführt. Die Funktion *Govern* wurde erst später in das NIST Framework aufgenommen und fehlt daher bei Kaur et al.

4.1.1 Identify

Die Funktion *Identify* identifiziert aktuelle Sicherheitslücken und hilft bei der Entwicklung geeigneter Risikomanagementstrategien, die auf die Bedürfnisse des Unternehmens zugeschnitten sind. Sie bildet daher die Grundlage für alle anderen Funktionen. Diese Funktion kann in folgende Teilbereiche unterteilt werden.

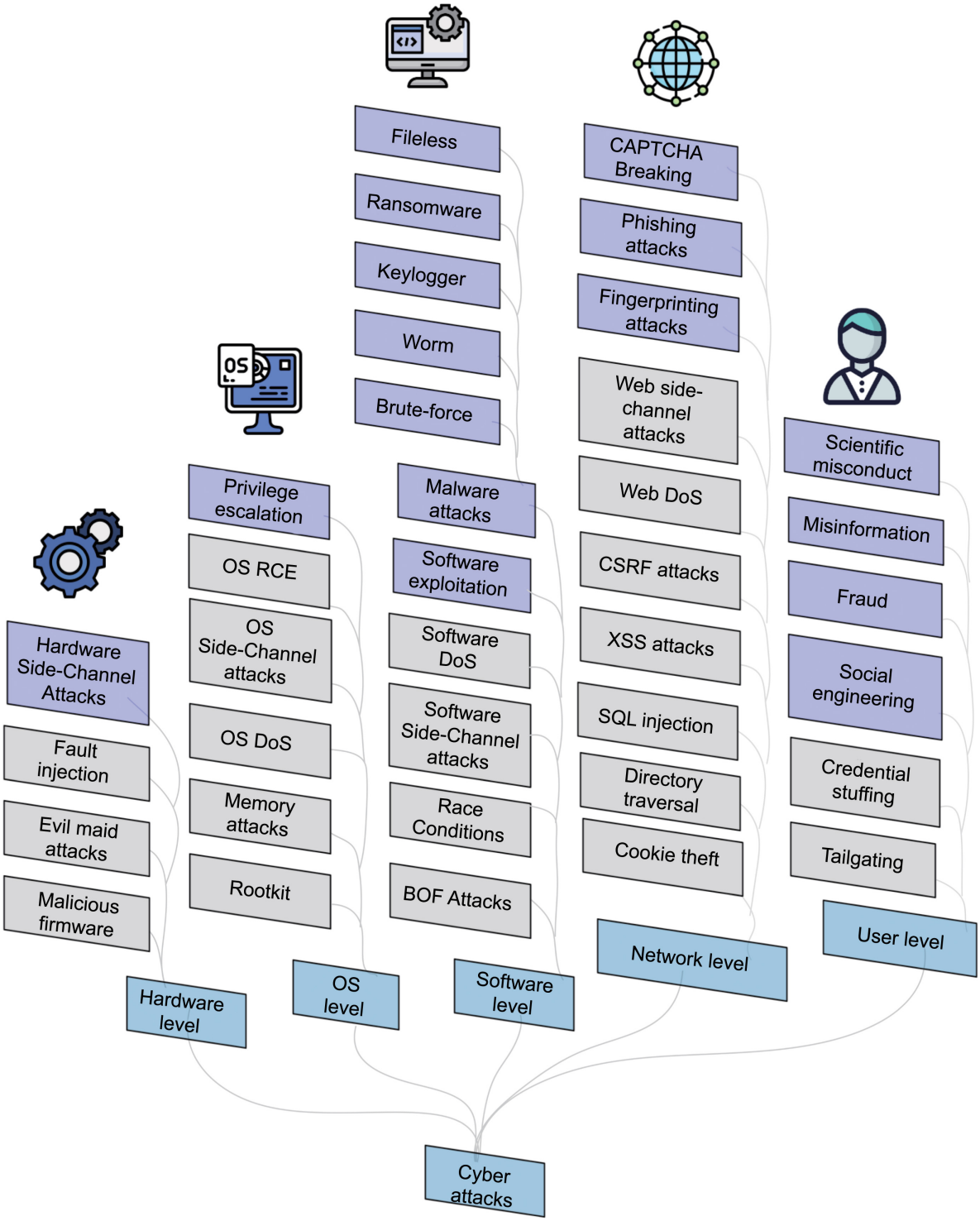

Abbildung 3: Layer-Based Taxonomie von Cyberangriffen [YDX+24]: Farbige Kästchen stellen Bereiche dar, für die es bereits LLM-basierte Angriffe in der akademischen Literatur gibt.

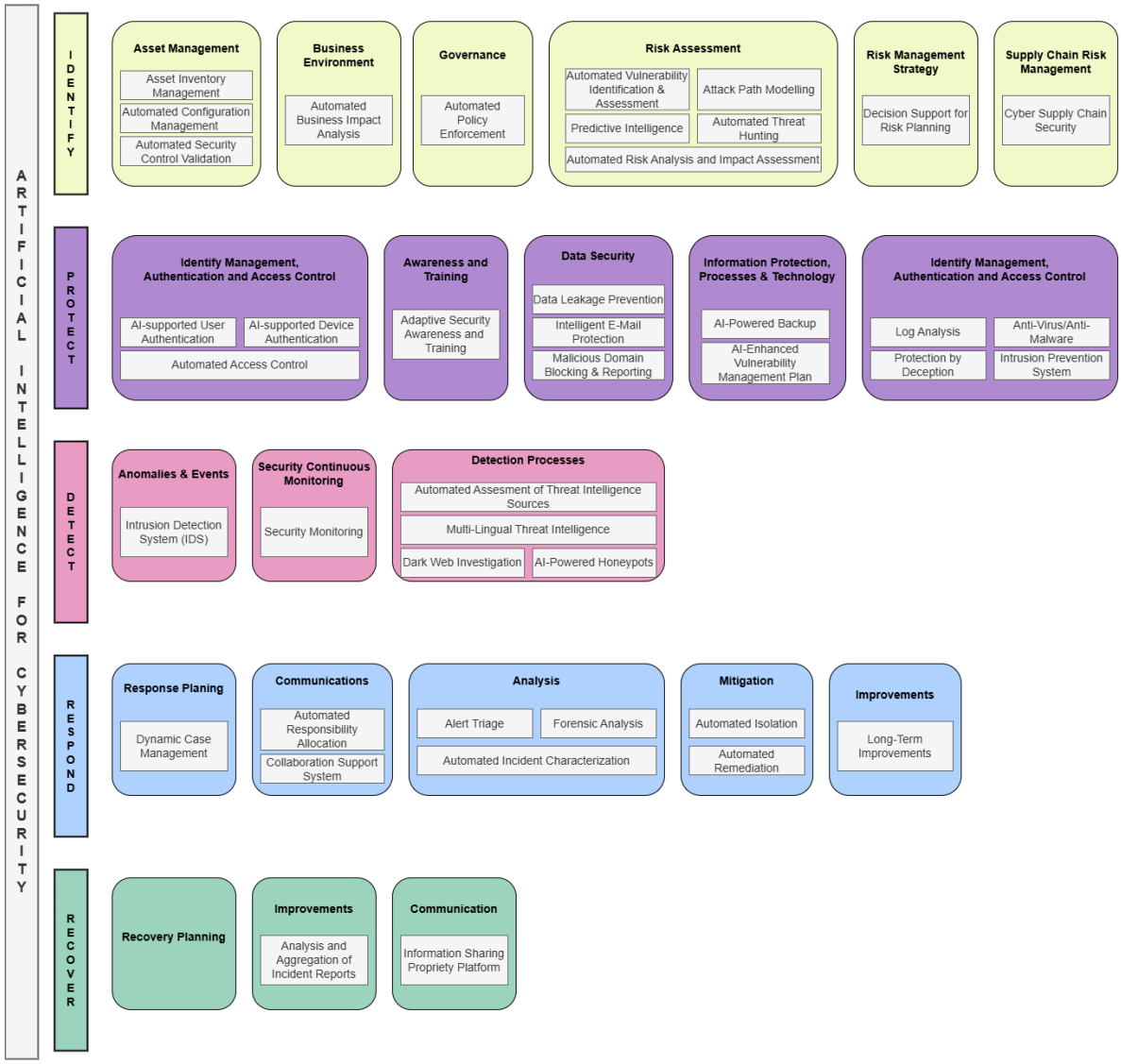

Abbildung 4: KI-basierte Verteidigung anhand der Funktionen *Identify, Protect, Detect, Respond* und *Recover* des NIST Cybersecurity Frameworks

Asset Management: Asset Management befasst sich mit der Identifizierung und Verwaltung von Informationen, Personal, Equipment, Systemen und Einrichtungen, die zur Erreichung der Ziele eines Unternehmens beitragen. Es umfasst folgende Anwendungsfälle: *Asset Inventory Management, Automated Configuration Management* und *Automated Security Control Validation*. Da Unternehmen immer mehr und verschiedene Plattformen nutzen, wird die Verwaltung von Cybersicherheitsressourcen immer komplexer. Um der ansteigenden Komplexität entgegen zu wirken, kann KI eine unterstützende Rolle spielen. Promyslov et al. [PSS19] verwendeten ein K-Means-Clustering zur Klassifizierung der Anlagen nach ihren Cybersicherheitsanforderungen auf der Grundlage ihrer Sicherheit, Funktionalität und Integrität in einem Kernkraftwerk. Millar et al. [MCCL20] präsentieren einen RF *Machine Learning Classifier* für die Klassifizierung von Betriebssystemen und die Identifizierung angreifbarer Geräte im Netz.

Business Environment: Hier werden kritische Prozesse und Anwendungen identifiziert, welche die Geschäftskontinuität auch in Krisensituationen sicherstellen. Diese Informationen sind entscheidend für die langfristige Stabilität des Unternehmens und bilden die Grundlage für die Entwicklung effizienter *Response*- und *Recover*-Strategien. KI-Technologie kann hier eingesetzt werden, um den Prozess für die Business-Impact-Analyse zu automatisieren. Kumar et al. [kN21] messen das wirtschaftliche Risiko der Cybersicherheit in verschiedenen Unternehmen mittels eines *Naïve Bayes* Algorithmus und *J48 Bagging Tree*. Dazu verwenden sie die Modellierung verschiedener bekannter Angriffsprofile. Im Gegensatz dazu präsentierten Nguyen et al. [NN20] einen Ansatz, der basierend auf *Uncertain Graphs* die Wahrscheinlichkeit von seltenen Sicherheitsvorfällen abschätzt. Um die Wahrscheinlichkeit eines hochgradig folgenschweren Cybervorfalles in mittelgroßen Netzwerken zu berechnen, schlugen Nguyen et al. eine Simulationsmethode für seltene Sicherheitsereignisse vor. Ponsard et al. [PRT21] untersuchten zusätzlich die Durchführbarkeit eines gezielten Angriffs auf ein bestimmtes Unternehmen, unter Berücksichtigung der Fähigkeiten der Angreifer:innen.

Governance: Richtlinien, Verfahren und Prozesse sind notwendig, um die betrieblichen Anforderungen zu verstehen und die Einhaltung der rechtlichen Verpflichtungen der Organisation zu überwachen. Damit können die Verantwortlichkeiten einer Organisation ermittelt und die Geschäftsleitung mit Informationen über Cybersicherheitsrisiken versorgt werden. In diesem Fall kann KI eingesetzt werden, um Richtlinien durchzusetzen oder den Abrufs wichtiger Risikoindikatoren zu automatisieren. Odegbile et al. [OCW19] präsentierten eine Architektur, die vollständig auf *software-defined Middleboxes* basiert. Dieser Ansatz ermöglicht eine automatische Durchsetzung von Richtlinien in *non-SDN*-Netzwerken, deren Router Pakete auf der Grundlage traditioneller, nicht richtlinienorientierter Routing-Protokolle weiterleiten.

Risk Assessment: Dies beinhaltet die Identifizierung, Abschätzung und Priorisierung von Cybersicherheitsrisiken in Bezug auf Betriebsabläufe, Betriebsmittel und Personen. Es erfordert eine sorgfältige Analyse von Informationen über Schwachstellen und Angriffe um festzustellen, welche Cybersecurity-Ereignisse negative Auswirkungen auf das Unternehmen haben könnten und wie wahrscheinlich das Eintreten solcher Ereignisse ist. Ein manueller Risikobewertungsprozess ist komplex, kostspielig, zeitaufwändig. Ein KI-basierter Risikobewertungsprozess kann das Risikomanagementteam bei der automatisierten Identifizierung und Bewertung von Schwachstellen, der automatisierten Suche nach Bedrohungen, der Modellierung von Angriffswegen und der automatisierten Risikoanalyse und Folgenabschätzung unterstützen. Nembhard et al. [NCE19] arbeiteten an der Erkennung von Softwareschwachstellen durch Überprüfung des Quellcodes mithilfe von Deep Learning und Transfer Learning. Huff et al. [HMLL21] fokussierten sich auf das Tracking von Schwachstellen, während Aota et al. [AKK+20] diese klassifizieren.

Risk Management Strategy: Die Strategie unterstützt die Entscheidungen über das operationelle Risiko, indem sie die Prioritäten, die Risikotoleranz und die Beschränkungen festlegt. Akzeptable Risikoniveaus werden zusammen mit angemessenen Lösungszeiten und Investitionen festgelegt und dokumentiert. KI kann die Risikoplanung mittels Entscheidungsunterstützung automatisieren. Während Rees et al. [RDRB11], Paul & Wang [PW19], Paul & Zhang [PZ21] sich auf *Decision Support Systems (DSS)* fokussieren, versuchen Zheng et al. [ZALT19] mit Hilfe eines Angriffsgraphen die effektivsten Sicherheitsmaßnahmen innerhalb eines vorgegebenen Budgets auszuwählen.

Supply Chain Risk Management: Dies erfordert ein sicheres, integriertes Netzwerk zwischen den Subsystemen der eingehenden und ausgehenden Lieferkette. KI-Techniken können genutzt werden, um Bedrohungsanalysen und -vorhersagen zu automatisieren [YOIL+21], optimale Investitionen in die Cybersicherheit zu ermitteln [Saw22] und die Cyber-Resilienz zu bewerten [RHG+21].

4.1.2 Protect

Diese Funktion hilft bei der Planung und Umsetzung geeigneter Maßnahmen, um die Auswirkungen eines potenziellen Cybersecurity-Ereignisses zu begrenzen oder einzudämmen. KI kann dabei helfen, das Benutzer:innen-Verhalten zu überwachen, Zugangskontrollen zu automatisieren, adaptives Training bereitzustellen, *Datenleaks* (Datenlecks) zu verhindern, die Integrität zu überwachen und automatisierten Informationsschutz zu bieten, um die Resilienz eines Systems zu verbessern. Im Folgenden werden KI-Ansätze für die Teilbereiche dieser Funktion beschrieben.

Identity Management, Authentication and Access Control: Diese Funktion stellt sicher, dass nur autorisierte Benutzer:innen, Prozesse, Geräte und Aktivitäten Zugang zu Ressourcen und relevanten Einrichtungen haben. KI kann bei der Verwaltung und Sicherung des physischen und des Fernzugriffs helfen, indem KI-gestützte Authentifizierungsprozesse für Benutzer:innen und Geräte, automatisierte Zugriffskontrollen sowie präzise Zugriffsberechtigungen eingesetzt werden. Der Authentifizierungsprozess kann entweder auf physischer oder verhaltensbasierter Biometrie (z. B. das Nutzungsverhalten, die Gangart etc.) basieren. Dadurch werden unbefugte Zugriffe erschwert und deren potenziellen Folgen verringert. Siam et al. [SSES+21] veröffentlichten einen physischen biometrischen Authentifizierungsprozess, der Photoplethysmographie[8] in Kombination mit ML nutzt. Sánchez et al. [SSHCFM+19] stellten einen kontinuierlichen, verhaltensbasierten Authentifizierungsmechanismus vor, der auf den Verhaltensprofilen der Benutzer:innen basiert. Sie analysierten dabei die Interaktion mit verschiedenen Bürogeräten in *Smart Offices* (intelligenten Büros) und nutzten ein Cloud-Computing-Paradigma sowie einen Random Forest Algorithmus.

Awareness and Training: Diese Kategorie umfasst die Sensibilisierung und die Schulung von Mitarbeiter:innen im Bereich Cybersecurity, damit sie ihren Verpflichtungen und Verantwortlichkeiten gemäß den Richtlinien und Verfahren nachkommen können. KI-Methoden können hier eingesetzt werden, um adaptive und personalisierte Cybersecurity-Schulungen oder -Empfehlungen bereitzustellen. Dies geschieht durch die automatische Auswahl von Inhalten mittels Sprachverarbeitungsalgorithmen [NCE17, TBH+20] oder durch KI, die intelligente Lösungsvorschläge liefert [EGLPA20].

Data Security: Datensicherheit regelt das Informationsmanagement gemäß der Risikostrategie zum Schutz sensibler Daten. Dies umfasst den Schutz von Daten im Ruhezustand und während der Übertragung, sowie die Verwaltung des gesamten Lebenszyklus von Systemen, die Daten verarbeiten. KI-Technologien können dabei helfen Datenlecks verhindern, E-Mails schützen und bösartige Domains blockieren. Le et al. [LZH21] haben einen Ansatz zur Anomalieerkennung auf Basis von Unsupervised Learning veröffentlicht, um Datenlecks zu verhindern, indem der Datenzugriff, die Datenbewegung und die Benutzer:innen-Aktivitäten überwacht werden. Zur Erkennung von Phishing-Versuchen in E-Mails haben Gualberto et al. [GDSDBV+20] einen mehrstufigen Ansatz veröffentlicht. Das Modell klassifiziert Nachrichten entweder als Phishing oder als legitim, basierend auf natürlicher Sprachverarbeitung und Machine Learning. Gallo et al. [GMBV21] konzentrieren sich auf die Erkennung von *Malicious Spam* (bösartigen Spam) in E-Mails, u. a. mittels RF und SVM.

Information Protection, Processes and Procedures: Dieser Bereich befasst sich mit dem Schutz von Informationsquellen in Übereinstimmung mit den festgelegten Sicherheitsrichtlinien, -prozessen und -verfahren. Dazu gehören der Schutz von Informationen sowie die Erstellung, Verwaltung und Umsetzung von Reaktions-, Wiederherstellungs- und Schwachstellenmanagementplänen. KI kann diesen Bereich unterstützen, indem sie Backups dynamisch plant und optimiert, oder ein verbessertes *Vulnerability Management* (Schwachstellenmanagement) bereitstellt. Qin et al. [QHL18] entwickelten

[8] Optische Technik zur Messung von Volumenänderungen im Blutgefäßsystem

ein dynamisches Backup-System mit intelligenten Planungsalgorithmen, um die Stabilität und Vorhersagbarkeit der Backup-Umgebung zu verbessern. Im Gegensatz dazu verwenden Van de Ven et al. [vdVZS14] eine zweidimensionale *Markov-Chain*, um Backups zu modellieren und die Optimierung der Backup-Planung zu untersuchen.

Protective Technology: Schutztechnologien sichern die Systeme und Vermögenswerte eines Unternehmens und stärken deren Widerstandsfähigkeit. Durch spezielle Manipulationsschutzmechanismen können Versuche, ins System einzudringen, es zu verändern oder Informationen unbefugt weiterzugeben, erkannt und abgewehrt werden. KI bietet in diesem Bereich vielfältige Schutzlösungen, wie z. B. durch Log-Analyse-Tools, Intrusion Prevention System (IPS) und KI-basierte Anti-Virus/Anti-Malware Lösungen. Afzaliseresht et al. [AMM+20] präsentieren eine neue Methode zur Darstellung von Log-Daten mittels KI. Sie nutzen Erzähltechniken, um einen Bericht über die Cyberbedrohungen in natürlicher Sprache zu erstellen, der dem Erfahrungsniveau des Lesers bzw. der Leserin entspricht.

Ein IPS für Automobil-Controller Area Networks (CANs), das Angriffe erkennen und verhindern kann, wurde von Freitas De Araujo-Filho et al. [FDAFPK+21] vorgeschlagen.

Gorment et al. [GSCK23] haben KI-Methoden zur Malware-Erkennung klassifiziert. Dabei unterscheiden sie primär zwischen überwachten und unüberwachten ML-Modellen.

- **Supervised Learning:** Die verwendeten Lernalgorithmen beinhalten Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees (DT) oder K-Nearest Neighbor (KNN).

  Makandar et al. [MP17] präsentieren einen Ansatz zur Erkennung von Malware-Klassen mithilfe einer Bildverarbeitungstechnik. Um Malware-Varianten zu klassifizieren wurden SVM und KNN verwendet.
  Bearden et al. [BL17] veröffentlichten ein KNN Modell, das Feature Selection und *Term Frequency - Inverse Document Frequency (TF-IDF)* verwendet, um Microsoft-Office-Dateien mit Makros als bösartig oder gutartig zu klassifizieren. Dabei wurde eine Accuracy von 96,3 % erreicht.
  Lu et al.'s [LKC+18] stellten einen SVM-basierten Mechanismus zur Erkennung von Malware in Android-Apps vor, indem sie die erforderlichen und verwendeten Berechtigungen analysierten und eine Accuracy von 99 % erreichten. Sie verwendeten eine signaturbasierte Klassifizierungsmethode mit einem statischen Analysetyp.
  Victoriano [Vic20] verwendete *Decision Trees* zur Erkennung von Android-Ransomware mit einer Behavior-Based-Technik und erzielte eine Accuracy von 99,08 %.

- **Unsupervised Learning:**

  Qasim et al. [Qa19] veröffentlichten einen hybriden Algorithmus mit NB Klassifikatoren und K-Means Clustering um Computerwürmer zu entdecken. Dabei wurde eine Accuracy von 88 % erzielt.
  Rosli et al. [RYMS19] stellten ein Modell zur Erkennung von Malware vor, das K-Means Clustering verwendet, um das Verhalten von Malware auf Basis von Computer-Registry Daten zu identifizieren und klassifizieren. Diese Modell erreichte eine Accuracy von über 90 %.

4.1.3 Detect

Während die Funktion *Identify* bestehende Sicherheitslücken aufdeckt, ermöglicht die Funktion *Detect* die frühzeitige Erkennung von Cybersecurity-Vorfällen. Diese Funktion ist entscheidend für die rasche Erkennung von Bedrohungen und der Minimierung deren Auswirkungen. KI kann die Erkennung beschleunigen, indem interne und externe Informationsquellen überwacht werden, um ungewöhnliche Aktivitäten frühzeitig zu erkennen. Diese Funktion gliedert sich in folgende Bereiche.

Anomalies and Events umfasst die Erkennung und Klassifizierung anomaler Aktivitäten durch die Definition von Baselines für normalen Betrieb und Datenflüsse. Ein *Intrusion Detection System (IDS)* überwacht System- und Netzwerkverkehr, um anormale Aktivitäten zu analysieren und mögliche Eindringversuche in das System zu erkennen. Die Erkennung durch das IDS kann entweder als *Binary Classification* (binäre Klassifikation) erfolgen, die zwischen normalem Verhalten und Angriff unterscheidet, oder als *Multi-Class Classification* (Klassifikation mit mehr als zwei Klassen), bei der verschiedene Angriffstypen klassifiziert werden. Vor allem zur *Binary Classification* gibt es viele Forschungsarbeiten; das Hauptaugenmerk ist die Evaluierung der Performanz unterschiedlicher *ML-Classifiers*

[WLFL21, SPPM21, GJB22] oder der Hyperparameter Optimierung [CP21]. Weitere Forschungsaspekte sind z. B. das Problem, dass die normalen und anomalen Fälle stark ungleich verteilt sind [BV21], oder auch *Feature Extractions* aus Datensätzen [RKI+22].

Continuous Security Monitoring umfasst die Echtzeitüberwachung von Informationssystemen und -ressourcen, um Einblicke in ihre Umgebung zu gewinnen und Sicherheitsereignisse zu erkennen. KI kann genutzt werden, um den Überwachungsprozess von Datenprotokollen zu automatisieren, die aus der Überwachung physischer Umgebungen, Netzwerke, Dienstleister, Benutzer:innen und Systeme mit sensiblen Informationen generiert werden. Grammatikis et al. [RGSI+21] haben ein speziell für ein *Smart Grid* entwickeltes Sicherheitsinformations- und Ereignisverwaltungssystem vorgestellt, das Cyberangriffe und Anomalien in Bezug auf eine Reihe von Protokollen der Anwendungsschicht eines *Smart Grid* erkennen, normalisieren und korrelieren kann. Fausto et al. [FGP+21] betrachten die Integration von Protokollen aus dem physischen und dem Cyber-Bereich und korrelierten deren Daten, um potenzielle Anomalien in kritischen Infrastrukturen zu erkennen.

Detection Processes: Dieser Prozess umfasst die kontinuierliche Verbesserung und Prüfung der Erkennungsprozesse für eine effiziente Funktionsweise. KI kann proaktive Überwachung im Internet bieten, indem sie automatisierte Bedrohungsintelligenz (*Threat Intelligence*) aus verschiedenen Web- und internen Ressourcen extrahiert. Sarhan und Spruit [SS21a] sowie Kim et al. [KLJL20] präsentierten Methoden, die es ermöglichen, mithilfe von *Named-Entity Recognition* Informationen aus Berichten zur Bedrohungsintelligenz zu extrahieren.

4.1.4 Respond

Diese Funktion erstellt einen Aktionsplan, um den Umgang mit und die Begrenzung der Auswirkungen eines potenziellen Cybersecurity-Ereignisses zu steuern. Mithilfe von KI können Vorfälle schneller und mit weniger Aufwand für Sicherheitsexpert:innen gelöst werden.

Response Planning: Diese Kategorie befasst sich mit der Erstellung von Aktionsplänen. KI kann dies automatisieren, indem ein dynamisches Fallmanagement-Tool den Plan dokumentiert, ausführt und aktualisiert. Dieses Tool lernt aus früheren Sicherheitsvorfällen, zeichnet verschiedene Angriffsszenarien auf und empfiehlt geeignete Reaktionsmaßnahmen, bevor ein Vorfall eintritt. Kim et al. [KIP10] beschreiben eine hierarchische Struktur, um ähnliche Fälle für eine schnelle Reaktion auf Sicherheitsvorfälle zu identifizieren. Dazu werden z. B. die Häufigkeit und verschiedene weitere Attribute analysiert. Jiang et al. [JGCX14] nutzen die hierarchische Struktur, um Attribute potenzieller Angriffsszenarien zu erfassen. Dazu gehören Informationen wie die Zielorganisation, Details über die Angreifer:innen, das betroffene System sowie die potenziellen Auswirkungen auf das Ziel. Nunes et al. [NCB+19] schlugen die Verwendung von *Case-Based Reasoning* (fallbasierter Schlussfolgerungen) auf Basis des *Incident Object Description Exchange Format (IODEF)* vor. Dadurch können die bei der Lösung von Cybersicherheitsvorfällen gesammelten Erfahrungen leichter wiederverwendet und ausgetauscht werden. Die Ähnlichkeit zwischen den Fällen wurde mittels *nearest-neighbour*-Suche ermittelt. Kraeva und Yakhyaeva [KY21] verwenden anstatt des ursprünglichen Merkmalsraums ein Empfehlungssystem, das Sicherheitsvorfälle mit Hilfe von Neural Network (NN) in einen latenten Merkmalsraum (*Embeddings*) umwandelt und dann nächstgelegene Vorfälle als Basis für einen Lösungsvorschlag ermittelt.

Communications: Mit dieser Aktivität kann die Kommunikation zwischen den Beteiligten während und nach einem Sicherheitsvorfall koordiniert werden. KI kann dabei unterstützen, indem sie die Zuweisung von Verantwortlichkeiten automatisiert oder eine Plattform für den Austausch von *Threat Intelligence* bereitstellt wie zum Beispiel von Lin et al. [LWY+19]. Shah et al. [SGJC18] befassten sich mit dem Zuweisungsproblem von Ressourcen wie Zeit und zusätzliches Personal und entwickelten dafür ein *Reinforcement Learning Model*.

Analysis: Dieser Prozess beinhaltet die Überprüfung des Sicherheitsvorfalls und der Reaktionsmaßnahmen, um sicherzustellen, dass der Vorfall korrekt gehandhabt wurde. KI kann hier für die auto-

matische Vorfallscharakterisierung, die Priorisierung von Warnmeldungen und die forensische Analyse eingesetzt werden. [DTN21, DMCF20]

Mitigation: Diese Aktivität hilft, die Ausbreitung von Sicherheitsvorfällen zu verhindern und ihre Auswirkungen zu beheben. KI kann eingesetzt werden, um einzelne Geräte zu isolieren, sobald ein Hinweis auf eine Kompromittierung (IoC) erkannt wird, oder um automatisierte Maßnahmen zur Behebung von Problemen, zur Beseitigung von Bedrohungen und zur Nachverfolgung der Bewegungen von Angreifer:innen durchzuführen. Sakhnini et al. [SKDP21] präsentieren ein Modell zur Angriffsklassifizierung und -lokalisierung für den *Physical Layer* in *Smart Grids*. Das Modell verwendet *Ensemble-* und *Representational-Learning* für die Angriffsklassifizierung und statistische Tests um den Angriff zu lokalisieren. Maimo et al. [FMHCPG+19] verwendeten ein ML-Modell um die Ausbreitungsphase von Ransomware-Angriffen zu erkennen und kategorisieren.

Improvements: Dies stellt sicher, dass aus dem Cyberangriff wichtige Erkenntnisse gewonnen werden, um eventuell den Aktionsplan und die Strategien entsprechend zu aktualisiert. KI kann hierbei Wissen aus den Vorfallsberichten extrahieren. Piplai et al. [PMJ+20] stellten ein System vor, mit dem Wissen von einem *After Action Report* extrahiert und gruppiert wird, um *Cybersecurity Knowledge Graphs* zu erstellen. Diese *Knowledge Graphs* helfen, Ähnlichkeiten zwischen verschiedenen Cyberangriffen zu finden. Woods et al. [WPL15] beschreiben einen Wissensgewinnungsprozess, der auf *Data-Mining-Techniken* basiert. Damit können Teilbereiche von Indikatoren und Vorfällen identifiziert werden, bei denen vollständige Informationen über den Vorfall für Sicherheitsanalystinnen und -analysten und Entscheidungsträger:innen hilfreich sein könnten.

4.1.5 Recover

Das Hauptziel der Wiederherstellungsfunktion ist eine schnelle Rückkehr zum Normalbetrieb, um die Auswirkungen des Vorfalls zu minimieren. Diese Funktion umfasst die Wiederherstellungsplanung, Verbesserungen und die Kommunikation.

Improvements: Der Prozess der Wiederherstellungsplanung kann durch die Überprüfung des Sicherheitsereignisses optimiert werden. In diesem Zusammenhang kann KI bestehende Strategien, Vorfallberichte und Prüfprotokolle automatisch analysieren, um Verbesserungspotenziale für zukünftige Reaktionspläne zu identifizieren. Meyers und Meneely [MM21] entwickelten eine automatisierte Methode für die *Post-Mortem Analysis* von Schwachstellen, um mit Hilfe von *Natural Language Processing* komplexe Beziehungen zu finden. Carriegos et al. [CCTDZ21] entwickelten eine Methode die Berichte über Cybersicherheitsvorfälle zusammenzuführt, um effektive Maßnahmen zu identifizieren und Vorhersagen zu treffen.

4.2 Anwendung von LLMs zur Verteidigung

Der Einsatz von LLMs in der Cybersicherheit bietet neue Möglichkeiten, den Zugriff auf Systeme und Daten zu verhindern oder zumindest vor möglichen Risiken und Gefahren zu warnen. Aktuelle Studien [YDX+24, MHM+24, ZBW+24, dJCdSW24, DP24], die den Stand der Forschung zusammenfassen, geben einen umfassenden Überblick. Es lässt sich feststellen, dass das Haupteinsatzgebiet von LLMs in der Verteidigung derzeit in der Unterstützung von Cybersicherheitsaufgaben liegt. Dabei wird die Fähigkeit von LLMs genutzt, um Wissen aus großen Textmengen zu extrahieren und bei Bedarf kompakt zur Verfügung zu stellen, was insbesondere bei repetitiven Aufgaben eine Zeitersparnis bedeuten kann und somit Kapazitäten für andere Aufgaben schafft. Im Folgenden werden besonders relevante Anwendungsgebiete vorgestellt.

4.2.1 Phishing

Phishing ist eine Betrugsmasche, bei der Angreifer:innen versuchen, durch gefälschte Textdokumente wie E-Mails, SMS, oder Webseiten an die Daten eines Opfers zu gelangen. Durch Phishing entstand in Österreich im Jahr 2023 ein Schaden von 24 Millionen Euro [fIuPPSAG]. Da die Angriffe auf Textdokumenten basieren, sind LLMs gut geeignet, diese zu analysieren und potenziell zu verhindern. Bursztein [Bur24] stellt eine Methode vor, bei dem ein LLM PDF-Dokumente analysiert und erklärt, warum

diese nicht vertrauenswürdig sind. Shibli et al. [SPG24] beschreiben eine Methode, die Benutzer:innen dabei unterstützen soll, SMS-Phishing („Smishing“), zu erkennen. Das Modell analysiert eine SMS beim Empfang und gibt anhand bestimmter Wörter und Satzteile Hinweise darauf, warum die SMS in betrügerischer Absicht versendet worden sein könnte. E-Mails und SMS sind jedoch oft nur der erste Schritt einer Phishing-Attacke. Der tatsächliche Betrug findet dann statt, wenn das Opfer seine Daten auf einer echt aussehenden Webseite eingibt. Um solche Webseiten zu erkennen, haben Koide et al. [KFNC24] ein LLM namens *ChatPhishDetector* entwickelt, das mittels Webcrawling-Methoden Phishing-Seiten identifizieren kann. Eine weitere Möglichkeit, mit Phishing-Versuchen umzugehen, wird von Cambiaso et al. [CC23] beschreiben. Sie präsentierten ein LLM, das Betrüger:innen in sinnlose Konversationen verwickelt, um deren Zeit und Ressourcen zu verschwenden.

Bursztein [Bur24] sagt auch voraus, dass diese Art der Analyse auch auf Bildern und Videos, die ebenfalls für Phishing-Versuche genutzt werden können, angewandt werden wird. Ziel ist es vor allem, gefälschte Videos und Bilder (Deepfakes) zu identifizieren und Menschen vor dem Material zu warnen.

4.2.2 Analyse und Absicherung der Codebasis

Die Absicherung von Code durch LLMs ist vor allem ein Werkzeug für Software-Entwickler:innen. In der Literaturübersicht (siehe Abschnitt 4; weiters auch z. B. Pearce et al. [PTA+23]) wird die Erkennung von Bugs und Fehlern im Code als wichtiges Ziel angesehen. Dabei werden Fehlerquellen aufgezeigt oder sogar direkt vom LLM behoben [YJW+24]. Zukünftig soll es zusätzlich die Möglichkeit geben, den Code *live*, während dem Programmieren, zu analysieren und auf Fehler oder unsichere Codestellen unmittelbar hinzuweisen. Zaharudin et al. [ZZS24] stellen ein LLM vor, das in Kombination mit der etablierten Software KLEE[9] Schwachstellen im Code erkennt, die zu Problemen mit dem Arbeitsspeicher führen können. Ein Vorteil bei der Verwendung von LLMs in diesem Fall ist, dass der Code bereits voranalysiert wird und nur relevante Codepassagen von KLEE überprüft werden. Dadurch kann der ansonsten sehr zeitintensive Einsatz von KLEE optimiert werden und größere Teile der Codebasis in kürzerer Zeit verifiziert werden.

Der aktuelle Konsens in der Forschung ist jedoch, dass die automatische Erkennung und Behebung von Schwachstellen durch LLMs selbst bei einfachen Problemen noch zu ungenau ist. Bursztein [Bur24] erwähnt, dass LLMs nur in etwa 15 % der Fälle Fehler im Code korrekt identifizieren und auch beheben. Dies liegt vor allem auch daran, dass das LLM oft nicht in der Lage ist den fehlerhaften Code adäquat zu ersetzen. So gibt es beispielsweise Situationen, wo der fehlerhafte Code einfach auskommentiert wird oder das LLM vorschlägt, den Modultest[10] aus dem Code zu entfernen, was dazu führen kann, dass potenzielle Fehler in zukünftigen Änderungen unentdeckt bleiben.

4.2.3 Allgemein unterstützende Aufgaben

Ein großer Vorteil von LLMs ist ihre Fähigkeit, große Informationsmengen effizient zu verarbeiten und Muster zu erkennen. Da sie auf einer umfangreichen Wissensbasis trainiert wurden, können sie Nutzer:innen bei einer Vielzahl von Aufgaben unterstützen. Solche Assistenzsysteme werden oft als Kopiloten bezeichnet. Im Folgenden werden einige Anwendungsbereiche aufgezeigt.

Content Moderation: Vor allem im Bereich der sozialen Medien, aber auch bei Videospielen mit Chat-Funktion ist die Moderation von Inhalten ein wesentlicher Bestandteil, um sicherzustellen, dass keine illegalen, betrügerischen, gewaltverherrlichenden oder pornografischen Inhalte geteilt werden [Bur24]. LLMs können Texte automatisch analysieren und bei Verdacht auf unerwünschte Inhalte diese markieren, so dass ein Mensch entscheiden kann, ob es sich tatsächlich um unangemessene Inhalte handelt oder nicht. In diesem Zusammenhang können LLMs auch eingesetzt werden, um das Fehlverhalten zu kategorisieren oder sogar zu korrigieren, indem verletzende Inhalte durch angemessene Inhalte ersetzt werden. Bis LLMs erfolgreich für *Content Moderation* eingesetzt werden, ist vermutlich noch einige Arbeit nötig, da es sich um einen sehr sensiblen Bereich handelt und Fehleinschätzungen weitreichende Folgen haben können. Kumar et al. [KAD24] und Kolla et al. [KSCS24] fassen das derzeitige Potenzial dieser Modelle anhand von Reddit[11] Textdaten zusammen. Kumar et al. [KAD24] geben an,

[9] https://klee-se.org/
[10] Ein Test, ob ein bestimmtes Modul (ein abgegrenzter Teil des Codes) richtig funktioniert.
[11] https://www.reddit.com/

dass unterschiedliche LLMs eine Genauigkeit von 64 % und eine Präzision von 83 % erreichen, wenn es darum geht zu klassifizieren, ob Texte den vorgegebenen Regeln entsprechen. Kolla et al. [KSCS24] geben für das Modell *LLM-Mod* eine Spezifität von 92,3 % und eine Sensitivität von 43,1 % für die Klassifikation der Texte an. Beide Publikationen kommen zu dem Schluss, dass die Modelle jetzt schon einen Beitrag in der *Content Moderation* leisten können, aber noch verbessert werden müssen, um sie vollautomatisch entscheiden zu lassen.

Penetrationstests: Im Bereich der Penetrationstests von Systemen gibt es beispielsweise Ansätze von Deng et al. [DLMV+24b] und Goyal et al. [GSP24], wo LLMs als Hilfsmittel zur Anleitung oder Zusammenfassung von Sicherheitstests verwendet werden. PentestGPT ist ein frühes Beispiel für LLMs, die Sicherheitstests und CTF-ähnliche Herausforderungen unterstützen [DLMV+24a]. Es handelt sich um ein interaktives Werkzeug, das Pentester:innen sowohl auf hoher als auch auf niedriger Ebene (Toolauswahl und -ausführung) anleitet. Dieser Ansatz hat auch auf Github einige Aufmerksamkeit bekommen[12].

Happe et al. [HC23] präsentierten einen Ansatz zur automatisierten Ausnutzung von Schwachstellen mit einem LLM (GPT3.5) in einer verwundbaren virtuellen Maschine. Das LLM wird mittels *Prompt* aufgefordert *Root-Zugriff* zu erlangen, generiert daraufhin Angriffsstrategien und Low-Level-Aktionen/Befehle, die automatisiert über SSH ausgeführt werden. Die Systemantwort wird dem LLM übergeben und dadurch die nächste Aktion ausgelöst (Schleife). Trotz einiger Herausforderungen, wie dem Versuch von LLMs nicht vorhandene Dateien (z. B. exploit.sh) auszuführen, begrenzten Context-Windows, oder Instabilität aufgrund von nicht deterministischen Abläufen war das LLM mit dieser einfachen Struktur routinemäßig in der Lage, Root-Rechte auf dem anfälligen Computer zu erlangen. In ihren Folgearbeiten [HKC24, HKC23] konzentrierten sich die Autoren auf Privilege-Escalation-Angriffe. Sie testeten mehrere LLMs und die Auswirkungen verschiedener Context-Windows, In-Context Lernen, optionalen High-Level-Guidance-Mechanismen und Speicherverwaltungstechniken. Die Ergebnisse zeigten, dass GPT-4-turbo gut geeignet ist, Schwachstellen auszunutzen (zwischen 33 % ohne Hilfestellung und 83 % wenn ein Hinweis gegeben wurde[13]). GPT-3.5-turbo konnte 16 % und 50 % der Schwachstellen ausnutzen, während lokale Modelle wie Llama3 nur bis zu 33 % der Schwachstellen ausnutzen konnten.

Einen Schritt weiter gehen Huang et al. [HZ24], die ein autonomes LLM-basiertes System zur Erkennung und Verbesserung von Sicherheitslücken vorstellen. Isozaki et al. [ISCK24] wollen die Auswahl des besten Modells unterstützen, indem sie unterschiedliche LLMs testen und analysieren, numerische Maßstäbe dafür setzen und Verbesserungspotenzial aufzeigen.

Systemsicherheit: LLMs können auch zur Automatisierung von Aufgaben eingesetzt werden z. B. das Schreiben von E-Mails. Wu et al. [WRK+24] werfen die Frage auf, ob sichergestellt werden kann, dass LLMs sich nur mit vertrauenswürdigen Anwendungen verbinden. Sollte das nicht der Fall sein, könnten diese Anwendungen zu Schwachstellen im System führen. Die Autor:innen schlagen vor, dass Anwendungen nur in einer sicheren Umgebung ausgeführt werden, so dass genau kontrolliert werden kann, womit sich das LLM verbindet. Außerdem schlagen sie ein Kommunikationsprotokoll vor, um Angriffe wie Prompt Injections[14] zu erschweren und eine manuelle Kontrolle zu erleichtern. Sladić et al. [SVCG24] verwenden LLMs zur Erstellung von sogenannten *Honeypots*[15]. Konkret simuliert ein LLM eine *Linux Shell* und generiert dabei glaubwürdige und dynamische Ergebnisse, die einige Schwachstellen von bisherigen *Honeypot* Implementationen verbessern sollen.

Anomalieerkennung: Han et al. [HYT23] beschreiben den Einsatz von LLMs in der Anomalieerkennung aus Logdateien. Das LLM lernt nun in welchen Situationen welche Lognachrichten generiert werden. Basierend darauf wird eine Vorhersage erstellt, was die K wahrscheinlichsten nächsten

[12] https://github.com/GreyDGL/PentestGPT

[13] Hinweise waren z. B. "there might be some exploitable suid binary on the system".

[14] Eine Prompt Injection ist eine Art Cyberangriff auf große Sprachmodelle (Large Language Models, LLMs). Hacker:innen tarnen dabei böswillige Eingaben als legitime Prompts und manipulieren so Systeme für generative KI (GenAI), sodass diese vertrauliche Daten preisgeben, Fehlinformationen verbreiten oder Schlimmeres verursachen.

[15] Als Honeypot wird ein Sicherheitsmechanismus bezeichnet, der Angreifer:innen täuscht und Attacken ins Leere laufen lässt. Er simuliert Netzwerkdienste oder Anwendungsprogramme, um Angreifer:innen anzulocken und das Produktivsystem vor Schäden zu schützen.

Lognachrichten sein werden. Sollte die dann tatsächlich auftretende Nachricht in der Vorhersage enthalten sein, geht das LLM von keiner Anomalie aus, andernfalls wird diese Lognachricht als Anomalie erkannt. Einen ähnlichen Ansatz verfolgen Haixuan et al. [GYW21], die ein LLM (LogBert) trainiert haben, das Muster in Computer-Logsequenzen erlernt.

Reaktion auf Vorfälle: Genauso wichtig wie das Verhindern von Vorfällen sind eine schnelle Reaktion und Beschreibung eines eingetretenen Vorfalls (Reporting), um frühzeitig darauf aufmerksam zu machen. LLMs werden für dies Aufgabe eingesetzt [Bur24], da sie rasch einen Vorfall identifizieren und einen Bericht dazu anfertigen können. Neben der Geschwindigkeit ist vor allem die Qualität der Reports ausschlaggebend, um einen Zwischenfall genau zu identifizieren und den Normalzustand gezielt wiederherzustellen.

Kommerzielle Werkzeuge zur Erkennung von Bedrohungen: Unternehmen wie Google, Virustotal, CrowdStrike oder Microsoft haben bereits damit begonnen, LLMs in ihre Produkte zu integrieren. Im Vordergrund bei Sicherheitsanwendungen steht die Erkennung von Bedrohungen, wie Viren, unsicheren Webseiten oder Schwachstellen im System, die den Zugriff unerwünschter Dritter ermöglichen. Darüber hinaus dienen diese Produkte dazu, Wissen zu vermitteln und Anwender:innen Unbekanntes zu erklären.

5 Angriffe auf und mit KI

In diesem Abschnitt werden potenzielle Möglichkeiten von KI-basierten Cyberangriffen (vor allem mittels LLMs) dargestellt; anschließend werden Angriffe auf KI bzw. ML (Abschnitt 5.2) näher erläutert.

5.1 KI-basierte Angriffe mit LLMs

KI hat viele Vorteile für die Cybersicherheit, aber KI kann auch ausgenützt werden, um Cyberangriffe zu starten. In diesem Abschnitt konzentrieren wir uns auf KI-gestützte Angriffe mit LLMs, die in letzter Zeit häufiger untersucht werden: laut der Studie von Yao et al. [YDX+24] gab es Anfang 2024 83 Veröffentlichungen zu LLMs mit defensiven Anwendungen, und 54 zu Angriffen.

Abbildung 3 zeigt die Taxonomie von Angriffen von Yao et al. [YDX+24]; Angriffe, für die bereits Methoden mittels LLMs bekannte sind, sind farblich hervorgehoben. Yao et al. stellten fest, dass die Hardware-Ebene am wenigsten, und die Benutzer:innen-Ebene am meisten von LLM-basierten Angriffen betroffen sind.

5.1.1 Automatisierte Angriffe

Die Verwendung von LLMs zur Unterstützung oder Automatisierung von Sicherheitstests könnte auch zu Angriffszwecken missbraucht werden (siehe Abschnitt 4.2.3). Moskal et al. [MLHO23] untersuchten das Potenzial von LLMs (GPT-3.5Turbo), um Aspekte von Cyberkampagnen mithilfe einer Plan-Act-Report-Schleife zu automatisieren. In einem *Prompt* wird das LLM aufgefordert, von einer Internet Adresse Informationen zu sammeln (reconnaissance), das System auszunutzen und Daten zu exfiltrieren. Die vorgeschlagenen Befehle werden ausgeführt und die Ergebnisse wiederum vom LLM interpretiert. Dieser Ansatz zeigte mehrere Einschränkungen, wie z. B. mangelndes Wissen über Aktionen und Tools, fehlendes Hintergrundwissen und unzureichende Abwägung und Auswahl zwischen mehreren Zielen und Optionen. Moskal et al. kommen zu dem Schluss, dass LLMs insbesondere das Risiko von Angriffen durch Akteur:innen mit geringen Fähigkeiten erhöhen.

5.1.2 Phishing

Mehr als 70 % aller Cyberangriffe nutzen Social Engineering als initialen Angriffsvektor [HSV+23]. Heiding et al. [HSV+23] führte eine Studie mit 112 Teilnehmer:innen durch, um KI-generierte Phishing-E-Mails (GPT-4) mit manuell gestalteten Phishing-E-Mails (V-Triad Framework) zu vergleichen. Eine Kontrollgruppe, die generischen Phishing-E-Mails ausgesetzt war, verzeichnete eine Klickrate zwischen

19 % und 28 %[16], während die mit GPT-4 generierten E-Mails zwischen 30 % und 44 %, die manuell mit V-Triad generierten E-Mails zwischen 69 % und 79 % und die mit GPT generierten und mit V-Triad verfeinerten E-Mails zwischen 43 % und 81 %. Sie zeigten außerdem, wie Sprachmodelle die Anreize für Phishing und Spear-Phishing erhöhen, indem sie deren Kosten senken.

Spear-Phishing ist ein gezielter Phishing Angriff auf eine bestimmtes Opfer, während Phishing normalerweise eine große Masse an Opfern generisch adressiert. In der Studie von Hazell [Haz23] wird die Skalierbarkeit von LLMs bei der Unterstützung der frühen Phasen von Spear-Phishing-Angriffen am Beispiel von 600 britischen Abgeordneten. Zunächst wurde GPT-4 verwendet, um ein einfaches Python-Skript zu schreiben, das die Wikipedia-Seite aller im Jahr 2019 gewählten britischen Abgeordneten ausliest. Diese unstrukturierten Wikipedia-Daten wurden dann an GPT-3.5 übergeben, um eine Biografie jedes:r Abgeordneten zu erstellen. Schließlich wurden mithilfe verschiedener GPT-Modelle personalisierte E-Mails erstellt, die auf die Region, die politische Partei, die persönlichen Interessen und andere Details der Abgeordneten Bezug nahmen. Obwohl die E-Mails nicht versendet wurden, erklärt der Autor, dass die daraus resultierenden E-Mails nicht nur realistisch, sondern auch bemerkenswert kostengünstig sind, da die Erstellung jeder E-Mail nur einen Bruchteil eines Cents kostet.

Die Fähigkeit von LLMs, bestimmte Personen und Gruppen zu imitieren, kann für überzeugenderes Phishing genutzt werden. Salewski et al. [SART+24] untersuchten die Fähigkeit von LLMs (Vicuna-13B und GPT-3.5-turbo), sich durch die Verwendung des Ausdrucks *„If you were a persona“* in der Eingabeaufforderung als verschiedene Rollen auszugeben. *Persona* wird durch eine soziale Identität oder Fachkompetenz ersetzt. Die Autor:innen stellen fest, dass LLMs, die sich als Fachexpertinnen und Fachexperten ausgeben, tatsächlich bessere Ergebnisse erzielen, als LLMs, die sich als Nicht-Fachexpertinnen und Nicht-Fachexperten ausgeben.

LLMs können auch bei der Erstellung von Phishing-Websites behilflich sein. Roy et al. [RNN23] identifizierten *Prompts*, die ChatGPT dazu bringen können, Phishing-Websites zu generieren. Durch einen iterativen Ansatz stellen sie fest, dass Phishing-Websites so gestaltet werden können, dass sie beliebte Marken imitieren und Taktiken nachahmen können, um Benutzer:innen dazu zu bringen, sensible Informationen preiszugeben. Ihr Prozess besteht darin, ChatGPT zu bitten, 1) sich von einer bestehenden Website inspirieren zu lassen, 2) Webelemente zu generieren, die Anmeldedaten stehlen, 3) einen Exploit zu implementieren (z. B. Textkodierung, Clickjacking, polymorphe URL und QR-Code-basierte mehrstufige Angriffe) und 4) Funktionen zu generieren, um die Anmeldedaten an die Angreifer:innen zu senden.

5.1.3 Password Cracking

Javier et al. [RPCH24] stellten die Frage, wie effektiv LLMs die zugrunde liegenden Muster erfassen können, die in von Menschen generierten Passwörtern verborgen sind. Ihr Modell PassGPT basiert auf der GPT-2-Architektur und wurde auf Passwortlecks (*Password Leaks*) trainiert. Es eignet sich sowohl zum Erraten (*cracking*) von Passwörtern als auch zur Abschätzung der Passwortstärke.

5.1.4 Malware-Generierung

Pa et al. [PPTK+23] untersuchen das Potenzial für die Entwicklung von Malware mittels LLMs. Ihr Benchmark umfasst die Entwicklung verschiedener Malware (z. B. Ransomware, Würmer, Keylogger, Brute-Force-Tools, fileless Malware) unter Verwendung von OpenAI Modellen wie ChatGPT und „text-davinci-003“. Obwohl die vollständig-automatische Generierung von Malware nicht möglich war, zeigen ihre Ergebnisse, dass es mit den richtigen Eingabeaufforderungen und Jailbreaks möglich ist, innerhalb von 90 Minuten (einschließlich Debugging-Zeit) trotz Sicherheits- und Moderationskontrollen in LLMs, funktionsfähige Malware und Angriffswerkzeuge zu generieren.

In ähnlicher Weise testete [MMHC23], wie man die Sicherheitsvorkehrungen zur Inhaltsmoderation von ChatGPT umgehen kann, um Malware zu erstellen. Während bestimmte Begriffe die Sicherheitsvorkehrungen auszulösen scheinen, konnten sie durch die Verwendung verschiedener abgeänderter Eingabeaufforderungen die Komponenten für ein Ransomware-Programm generieren.

[16] 19 % haben einen Link in den E-Mails geklickt. Es ist unklar ob die restlichen Teilnehmer:innen überhaupt ihre E-Mails überprüft haben. Wenn nur die Teilnehmer:innen gezählt werden, die auch den Abschlussfragebogen beantwortet haben, sind es 28 %.

Botacin et al. [Bot23] testeten verschiedene Entwicklungsstrategien (z. B. die Generierung vollständiger Malware und die Erstellung von Malware-Funktionen) und untersuchten die Fähigkeit von LLMs, Malware-Code neu zu schreiben. Ähnlich zu den anderen Arbeiten zeigten auch ihre Experimente, dass GPT-3 Schwierigkeiten hat, vollständige Malware-Beispiele automatisiert zu generieren. Es können aber Code-Blöcke generiert werden, deren Kombination zu funktionaler Malware führt. Darüber hinaus kann das LLM auch mehrere Versionen des Codes (Malware-Varianten) erstellen.

Charan et al. [CCAS23] untersuchten, wie ChatGPT und Google Bard missbraucht werden können, um Code zu generieren, der die Top 10 MITRE-Techniken 2022 implementiert. Ihre Experimente zeigten, dass ChatGPT die Entwicklung von gezielten Angriffen potentiell beschleunigen kann. Insbesondere werden weniger versierte Angreifer:innen befähigt weitreichendere Angriffe durchzuführen. Ebenso können LLMs dabei unterstützen, Varianten von Ransomware zu generieren.

Beckerich et al. [BPC23] präsentieren ein Proof-of-Concept, bei dem ChatGPT verwendet wird, um Malware zu verteilen und gleichzeitig eine Erkennung zu vermeiden.

5.1.5 Angriffe auf Benutzer:innen-Ebene

Das Potenzial von LLMs zur Verbreitung von Fehlinformationen stellt ein ernstes Problem für die Online-Sicherheit und das Vertrauen der Öffentlichkeit dar[17]. Chen et al. [CS24] präsentieren eine Taxonomie und analysierten potenzielle Methoden zur Generierung von Fehlinformationen mithilfe von LLMs; sie bewerteten anschließend, wie schwierig es ist, von LLM generierte Fehlinformationen zu erkennen. Sie stellten fest, dass von LLM generierte Falschinformationen für Menschen und Detektoren schwieriger zu erkennen sind als von Menschen geschriebene Falschinformationen mit derselben Semantik – LLMs können möglicherweise besser in die Irre führen und potenziell mehr Schaden anrichten. Wu et al. [WGH24] beschreiben, dass LLMs böswillige Akteur:innen in die Lage versetzen, den Stil vertrauenswürdiger Nachrichtenquellen nachzuahmen – und zwar schnell, kostengünstig und in großem Maßstab. Ihre Analyse zeigt, dass mit LLM getarnte Fake-News-Inhalte die Wirksamkeit modernster textbasierter Detektoren erheblich untergraben (bis zu 38 % Rückgang des F1-Scores[18]).

Da LLMs auf privaten Datensätzen trainiert werden, zeigten Carlini et al. [CTW+21], dass Angreifer:innen die Extraktion von Trainingsdaten durchführen können, um einzelne Trainingsbeispiele durch Abfrage des Sprachmodells wiederherzustellen. Staab et al. [SVBV24] zeigten, dass LLMs dazu verwendet werden könnten, die Privatsphäre von Personen zu verletzen, indem sie persönliche Attribute aus den jeweils verfassten Texten ableiten. Anhand eines Datensatzes realer Reddit-Profile zeigen sie, dass aktuelle LLMs eine Vielzahl persönlicher Attribute (z. B. Standort, Einkommen, Geschlecht) ableiten können und dabei eine Genauigkeit von bis zu 85 % bei den Top 1 und 95 % bei den Top 3 erreichen – und das zu einem Bruchteil der Kosten (Faktor 100) und Zeit (Faktor 240), die Menschen dafür benötigen würden. Darüber hinaus stellen sie fest, dass gängige Schutzmaßnahmen, d. h. Textanonymisierung und Model Alignment, derzeit nicht ausreichen, um die Privatsphäre der Benutzer:innen vor LLM-Inferenz zu schützen.

5.1.6 Berichte aus der Praxis

Da die in diesem Kapitel vorgestellten offensiven Anwendungen hauptsächlich auf akademische Arbeiten basieren, bleibt die Frage, ob diese Techniken tatsächlich eingesetzt werden, eher unbeantwortet.

Microsoft berichtete kürzlich, dass die Geschwindigkeit, der Umfang und die Raffinesse von Angriffen parallel zur raschen Entwicklung und Einführung von KI zugenommen haben[19]. Verteidiger:innen dagegen beginnen gerade erst, die Fähigkeiten generativer KI zu erkennen und zu nutzen, um das Gleichgewicht in der Cybersicherheit zu ihren Gunsten zu kippen. OpenAI[20][21] gab bekannt, dass sie in Zusammenarbeit mit Microsoft Threat Intelligence fünf staatlich unterstützte Gruppen daran ge-

[17] OWASP. (2023, Okt.) OWASP Top 10 für LLM. [Online]. Verfügbar: https://owasp.org/ www-project-top-10-for-large-language-model-applications/ assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_1.pdf

[18] Der F1-Score ist eine Leistungskennzahl, die bei binären Klassifizierungsproblemen verwendet wird. Es ist das harmonische Mittel aus *Precision* und *Recall*, das die Genauigkeit und Vollständigkeit des Modells misst.

[19] https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

[20] https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

[21] https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/nation-state-hackers-use-chatgpt-to-improve-cyberattacks

hindert haben, KI-Dienste zur Unterstützung bösartiger Cyberaktivitäten zu nutzen. Als Aktivitäten nennen sie bspw. Phishing-Kampagnen, Skripterstellung, Erstellung von Inhalten, die wahrscheinlich für Spear-Phishing-Kampagnen verwendet werden und die Erforschung gängiger Methoden zur Malware-Erkennungsvermeidung. Sie führen weiter aus, dass diese Aktivitäten mit früheren Red Team-Bewertungen übereinstimmen, die in Zusammenarbeit mit externen Cybersicherheitsexpertinnen und Cybersicherheitsexperten durchgeführt wurden. Dabei wurde festgestellt, dass die Fähigkeiten von GPT-4 für böswillige Cybersicherheitsaufgaben nur geringfügig über das hinausgehen, was bereits mit öffentlich verfügbaren, nicht KI-gestützten Tools erreichbar ist.

Die Insikt Group [Gro23] berichtet, dass Cyberkriminelle diese Fähigkeiten offenbar aktiv sondieren. Innerhalb weniger Monate nach der Veröffentlichung von ChatGPT tauchten zahlreiche Beispiele auf, in denen Hacker die Fähigkeit des Modells diskutierten, bei der Erstellung von Malware zu helfen. Obwohl der Großteil dieses Codes rudimentär ist und mit ziemlicher Sicherheit schwächer ist als Malware, die bereits im Internet verfügbar ist, haben LLMs die Eintrittsbarriere für weniger versierte Cyberkriminelle gesenkt, die Spear-Phishing-Kampagnen starten wollen [Gro23].

Im Dark Web sind auch spezialisierte Modelle für böswillige Zwecke verfügbar [LCLW24, Fal23]. Beispiele hierfür sind WormGPT[22], ein KI-Tool, das auf GPT-J[23] basiert und für Business Email Compromise (BEC) Angriffe[24] eingesetzt werden kann. WormGPT soll mit einer Vielzahl von Datenquellen trainiert worden sein, wobei der Schwerpunkt auf Malware-bezogenen Daten lag. FraudGPT[25][26] kursiert seit Juli 2023 im Dark Web und Telegram-Kanälen. Es soll dabei helfen, bösartigen Code zu schreiben, Seiten und betrügerische Nachrichten zu fälschen, nicht nachweisbare Malware, Phishing-Seiten und andere Hacking-Tools zu erstellen, sowie Lecks und Schwachstellen zu finden und relevante Gruppen, Websites und Märkte zu überwachen. Laut der Studie von Lin et al. [LCLW24] liegen beispielsweise die Preise für FraudGPT bei 90 Euro und WormGPT bei 109 Euro pro Monat; in einem Blogeintrag der *Infosecurity Europe Conference*[27] wird behauptet, dass FraudGPT mit Stand Ende Juli 2023 über 3000 bestätigte Verkäufe und Bewertungen vorweisen konnte.

5.2 Angriffe auf Maschinelles Lernen

Neben der Nutzung von KI bzw. maschinellem Lernen als Offensives oder Defensives Werkzeug kann jede KI für sich Ziel eines Angriffes sein – sei es KI, die im Bereich Cybersecurity eingesetzt wird, oder auch KI, die in komplett anderen Bereichen eingesetzt wird. NIST [VOFA24] definiert vier Arten von Angriffen: (1) *Evasion Attacks,* (2) *Poisoning Attacks*, (3) *Privacy Attacks* und (4) *Abuse Attacks.* Neben NIST gibt es noch weitere Taxonomien für Attacken auf ML-Modelle wie zum Beispiel MITRE Atlas[28], die einer ähnlichen Strukturierung folgen. OWASP veröffentlicht auch Top Ten-Listen für Machine Learning Security[29] und spezifisch für Large Language Model Applications[30]. Im folgenden beschreiben wir Angriffe in der Taxonomie von NIST.

5.2.1 Evasion Attacks

Im Rahmen eines Evasion-Angriffs zielen die Angreifer:innen darauf ab, *Adversarial Examples* zu erzeugen - Eingabedaten, deren Klassifizierung zur Laufzeit durch minimale Änderungen so manipuliert wird, dass die Eingabedaten einer beliebigen, von ihnen gewählten Klasse, zugeordnet werden können. Der Evasion-Angriff kann entweder direkt gegen ein beliebiges ML-Modell gerichtet sein, aber auch gegen ein ML-Modell, welches als Verteidigungstool gegen Cyberangriffe verwendet wird [SWD+22]. Evasion-Attacken können als Folge eine Verletzung der Verfügbarkeit oder der Integrität des ML-Modells bewirken.

[22] https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html
[23] https://www.eleuther.ai/artifacts/gpt-j
[24] BEC Angriffe sind eine Form von Phishing Angriff, die meist einen C-Level Manager oder eine Budgetverantwortlichen als Ziel haben
[25] https://www.heise.de/news/FraudGPT-schreibt-Phishing-Mails-und-entwickelt-Malware-9231555.html
[26] https://www.infosecurity-magazine.com/news/dark-web-markets-fraudgpt-ai-tool
[27] https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/generative-ai-dark-web-bots.html
[28] https://atlas.mitre.org/
[29] https://owasp.org/www-project-machine-learning-security-top-10/
[30] https://owasp.org/www-project-top-10-for-large-language-model-applications/

Adversarial Examples sind in vielen Domänen demonstriert worden, vor allem im Bereich der Bildanalyse. Aber auch im Bereich der Cybersecurity gibt es zunehmend Beispiele. Chernikova et al. [CO22] beispielsweise schufen das Framework FENCE für die Entwicklung von White-Box-Evasion-Attacks in diskreten Domänen. Diese Attacke wurde auf NNs angewendet, die für die Klassifizierung von Botnetzen und bösartigen Domänen trainiert wurden. Sheatsley et al. [SHP+21] präsentierten eine Methode, die durch den Einsatz formaler Logik Beschränkungen im Merkmalsraum erlernt; diese Technik wurde auf die Erkennung von *Network-Intrusion-Detection* und *Phishing-Classifiers* angewandt.

5.2.2 Poisoning Attacks

Bei einem Poisoning-Angriff werden die Trainingsdaten eines ML-Modells manipuliert, um dessen Funktionalität zu beeinträchtigen. Angreifer:innen können durch das Manipulieren von Trainings- und Testdaten aus öffentlich zugänglichen Quellen den normalen Ablauf des Trainingsprozesses stören, indem sie unautorisierte, modifizierte oder gelöschte Daten einfügen. Die während der Trainingsphase eingeschleusten Daten führen zu verzerrten Vorhersagen.

Poisoning-Angriffe haben in der Cybersicherheit eine längere Geschichte – der erste bekannte Poisoning-Angriff wurde 2006 für die Erstellung von Wurmsignaturen (*Worm Signature Generation*) entwickelt [PDL+06]. Heutzutage werden Poisoning-Angriffe in verschiedenen Bereichen untersucht, z. B. in der Computersicherheit zur Spam-Erkennung [NBC+08], *Vulnerability Prediction* (Schwachstellenvorhersage) [SSD15], *Network-Intrusion-Detection*[VSI+21], Malware Klassifizierung [SMCO21, XBB+15]; andere Domänen beinhalten Computer Vision Modelle [GFH+20, SHN+18, GLDGG19], Natural Language Processing [CSC+21, LLD+21, WZFS21] und tabellarische Daten in den Bereichen Gesundheitswesen und Finanzen [WZFS21]. Sowie bei der Evasion-Attacke, kann die Folge eines Poisoning-Angriffes eine Verletzung der Verfügbarkeit oder der Integrität des ML-Modells auslösen. Je nachdem wie der Angriff gestaltet wird, müssen die Angreifer entweder die Daten, die dazugehörigen Labels, oder beides modifizieren.

5.2.3 Privacy Attacks

Bei einem Angriff auf die Privatsphäre versuchen die Angreifer:innen, sensible Daten über Datensätze oder ML-Modelle abzuleiten. Zu diesen Angriffen gehören *Data Reconstruction* (Datenrekonstruktion), *Membership Inference*, *Model Extraction* (Modellextraktion) und *Property Inference*. Bei Datenrekonstruktionsangriffen im Bereich der Cybersecurity werden z. B. private Informationen über Benutzer:innen-Datensätze oder sensible kritische Infrastrukturdaten zurückgewonnen. Im Jahr 2003 wurden die ersten Datenrekonstruktionsangriffe von Dinur and Nissim [DN03] vorgestellt, die Daten von Benutzer:innen aus linearen Statistiken wiederherstellen.

Eine *Membership Inference Attack* bestimmt, ob ein bestimmter Eintrag im Datensatz enthalten war, der für das Training des Modells verwendet wurde. Die erste *Membership Inference* Attacke wurde von Homer et al. [HSR+08] entwickelt. Ein Modellextraktionsangriff zielt darauf ab, Informationen über die Modellarchitektur und/oder -parameter zu extrahieren, indem Abfragen an das ML-Modell gestellt werden, das von einem *ML-as-a-Service*-Anbieter trainiert wurde. Die ersten dieser Angriffe wurden von Tramer et al. [TZJ+16] auf mehrere Online-ML-Dienste für verschiedene Modelle entwickelt. Bei *Property Inference* Attacken versuchen die Angreifer:innen, durch Interaktion mit einem ML-Modell globale Informationen über die Verteilung der Trainingsdaten zu erfahren. So können sie den Anteil des Trainingsdatensatzes mit einem bestimmten sensiblen Attribut bestimmen. Diese Attacke wurde zum ersten Mal von Ateniese et al. [AMS+15] im Jahr 2015 präsentiert.

5.2.4 Abuse Attacks

Ein *Abuse Attack*, auch bekannt als *Abuse Violation* (Missbrauchsverletzung), liegt vor, wenn Angreifer:innen die vorgesehene Verwendung eines GenAI-Systems umfunktioniert, um ihre Ziele zu erreichen. Dabei nutzen sie die Fähigkeiten von GenAI-Modellen, um Hassreden oder Diskriminierung zu verbreiten, Medien zu erstellen, die zu Gewalt gegen bestimmte Gruppen aufstacheln, oder offensive Cybersicherheitsoperationen zu starten, indem sie Bilder, Texte oder bösartigen Code erstellen, die einen Cyberangriff ermöglichen.

5.3 Deepfakes - Video, Bilder, Audio

Deepfakes [ML22] sind synthetisch erzeugte Multimedia-Inhalte, die den Eindruck eines tatsächlichen, realen Inhalts erwecken sollen (z. B. eine bestimmte Person, die überzeugend eine Rede hält oder eine Handlung ausführt, die in Wirklichkeit so nicht stattgefunden hat). Deepfakes haben große Besorgnis hervorgerufen, da manipulierte Inhalte dazu verwendet werden können, die Meinung der Menschen zu bestimmten Themen (wie religiösen oder politischen Aspekten) zu manipulieren oder falsche Ereignisse darzustellen und so die Gesellschaft als Ganzes zu gefährden[31].

Die Manipulation von Multimedia-Inhalten ist zwar nicht neu, doch bei Deepfakes kommen fortgeschrittene Methoden des maschinellen Lernens zum Einsatz, mit denen realistischere Inhalte in weit größerem Umfang als je zuvor erzeugt werden können. Ausgefeilte Methoden für den Austausch von Gesichtern, die Nachahmung von Gesichtern oder die Umwandlung von Stimmen ermöglichen die Erstellung von Inhalten, die leicht als authentisch angesehen werden können. Fortschritte bei der Deepfake-Generierung bedeuten auch, dass diese Technologien einem breiteren, auch technisch weniger versierten Zielpublikum zugänglich werden, was Angriffe durch Laien ermöglicht.

Deepfakes können das Aussehen einer Person (Bild- und Video-Deepfakes) oder ihre Stimme (Audio-Deepfakes) imitieren. Bei *Reenactment* wird ein Bild so verändert, dass Gesichtsausdruck, Blick, Mund, Körperhaltung usw. dem Originalbild ähneln. *Replacement* übernimmt den Inhalt eines Quellbildes und wendet ihn auf ein Zielbild an, wobei die Identität der Person im Zielbild erhalten bleibt. Bei gefälschten Gesichtsbildern kann das Gesicht einer Quellperson durch das Gesicht einer Zielperson ersetzt werden. Solche Deepfakes, die als *Face Swap* bezeichnet werden, sind die häufigsten Beispiele für die Ersetzung von Bildern.

Da Videos im Vergleich zu statischen Bildern dynamisch sind, ist ihre Erstellung komplexer – aber Videos sind auch überzeugender, und können aufgrund ihrer lebensechten Darstellung der Zielpersonen eine erhebliche emotionale und wahrnehmungsbezogene Wirkung auf die betrachtende Person haben. Sobald die Videos in sozialen Netzwerken verbreitet werden, verstärken sie ihre Wirkung schnell, da sie viral gehen können und eine große Reichweite erzielen. Die Arten von Video-Deepfakes entsprechen den Kategorien der Bild-Deepfakes und umfassen vor allem den Austausch von Gesichtern und das Nachstellen von Personen. Beim Gesichtstausch wird die Identität der Person gefälscht, weshalb er manchmal auch als Identitätstausch bezeichnet wird. Dies ist die häufigste Art von Video-Deepfakes. Beim *Reenactment*, auch *Puppet-Mastery* oder *Expression Swap* genannt, zeichnet die Ausgangsperson die Mimik auf, z. B. Mund- und Lippenbewegungen, Augenblinzeln und Kopfhaltung, um die Mimik des Opfers nach den Wünschen des Imitators zu verändern.

Audio-Deepfakes simulieren die Stimme einer Zielperson. Diese auf den ersten Blick am wenigsten schädliche Art von Deepfakes hat sich als reale Bedrohung erwiesen. Es wurde von Fällen berichtet, in denen hohe Schäden entstanden sind, weil Angestellte glaubten, Befehle von einem Vorgesetzten zu erhalten. Derzeit sind zwei Techniken, um eine Stimme zu imitieren verbreitet: Sprachumwandlung und Text-to-Speech-Synthese. Bei der Sprachumwandlung werden zwei Sprachaufnahmen als Eingabe verwendet: eine vom Zielsprecher und eine vom Ausgangssprecher. Die Sprache des Ausgangssprechers wird so verändert, dass sie so klingt, als würde der Zielsprecher sprechen. Text-to-Speech-Synthese [Tay09] zielt darauf ab, aus dem Text eine menschenähnliche Sprache zu erzeugen; die erzeugte Stimme kann der Stimme einer realen Person ähneln oder völlig künstlich sein. In einer kürzlich veröffentlichten Ankündigung von OpenAI[32] heißt es, dass ihr neuer Sprachsynthesizer bereits nach 15 Sekunden Sprachaufnahme realistische Sprache erzeugen kann.

6 Einschätzung aus der Industrie und Forschung

Um Einblicke aus der Industrie zum Thema Künstliche Intelligenz und Cybersicherheit zu erhalten, wurden semi-strukturierte Interviews mit Expert:innen aus Unternehmen, die in Österreich im Bereich Cybersicherheit tätig sind und aus der Forschung geführt. Die wichtigsten Ergebnisse dieser Interviews werden im Folgenden zusammengefasst.

[31] Z. B., wenn auch von geringer Qualität, die Deepfakes, die den ukrainischen Präsidenten Volodymyr Zelensky zeigen, wie er die Truppen seines Landes zur Kapitulation auffordert

[32] https://openai.com/blog/navigating-the-challenges-and-opportunities-of-synthetic-voices

6.1 Joe Pichlmayr, Ikarus Security Software GmbH

Joe Pichlmayr ist seit 1993 bei der IKARUS Security Software GmbH tätig und übernahm 1998/99 im Zuge eines Management-Buy-outs die Rolle des Geschäftsführers. Unter seiner Leitung hat sich das Unternehmen zu einem bedeutenden Anbieter von IT- und OT-Sicherheitslösungen entwickelt, einschließlich eigener Scan-Engines, Cloud-Services sowie SOC-, SIEM- und Log-Management-Diensten. Neben seiner Tätigkeit bei IKARUS engagiert sich Pichlmayr seit 2011 für die Förderung des IT-Sicherheitsbewusstseins in Österreich. Er ist Mitbegründer von Cyber Security Austria und der Austria Cyber Security Challenge und setzt sich für die Ausbildung von Nachwuchstalenten im Bereich Cyber Security ein. Zudem ist er in Initiativen wie DigitalCity.Wien aktiv, um den IKT-Standort Wien zu stärken.

KI verändert die Cybersicherheitslandschaft grundlegend. Ikarus sieht derzeit keinen klaren Vorteil für Angriff oder Verteidigung. In Bezug auf Angriffe erleichtert KI die Generierung von Malware, die Optimierung von Exploits und die Entwicklung komplexer Angriffstechniken. Insbesondere ermöglicht sie die Erstellung einzigartiger Malware, die herkömmliche Erkennungsmethoden umgeht, sowie die Automatisierung von Angriffen, die bisher manuell durchgeführt werden mussten. Dadurch sinken die Einstiegshürden und der Aufwand für Cyberkriminelle erheblich, während die Komplexität und Effektivität der Angriffe steigt. Auf der anderen Seite steht die Verteidigung vor der Herausforderung, dass klassische signaturbasierte Antivirenlösungen nicht mehr ausreichen. Stattdessen setzt sie zunehmend auf KI-basierte Mechanismen wie Verhaltensanalyse und Anomalieerkennung. Dennoch bleibt ein erheblicher manueller Testaufwand, etwa zur Vermeidung von Fehlalarmen, unverzichtbar.

Während automatisierte Pentesting-Toolkits bereits weit verbreitet sind, gibt es kaum Belege für den Einsatz vollständig KI-gestützter Malware. Experten und Expertinnen gehen davon aus, dass bei ernsthaften Angriffen eigene, nicht-öffentliche Modelle verwendet werden, was die Erkennung zusätzlich erschwert. Die Abwehr erfordert innovative Ansätze wie die dynamische Analyse von Angriffsmustern oder die Entwicklung intelligenter Signaturen, um mit der Geschwindigkeit der Bedrohungen Schritt zu halten.

Sprachmodelle haben Potenzial, insbesondere bei der Analyse großer Datenmengen, die in Cybersicherheitsprozessen anfallen. Sie erleichtern es, relevante Informationen aus unstrukturierten Quellen zu extrahieren und erleichtern den Einstieg in die Cybersicherheit, da weniger technisches Wissen erforderlich ist. Dennoch sind LLMs in ihrer Präzision und Effizienz bisher Spezialistinnen und Spezialisten unterlegen und (noch) kein Game-Changer. Ihr Einsatz wird derzeit vor allem als unterstützende Technologie gesehen, z. B. zur Erkennung von Phishing-Versuchen oder als Assistent bei Analyseaufgaben (z. B. Vorfilterung und Erkennung von Mustern und Trends).

Unternehmen suchen zunehmend nach KI-basierten Lösungen, stoßen dabei aber auch auf Probleme. Gleichzeitig besteht eine große Zurückhaltung gegenüber der vollständigen Automatisierung von Sicherheitsmaßnahmen, wie dem Blockieren von Bedrohungen oder automatischen Anpassungen als Abwehrmaßnahme, da Datenschutz und Vertrauen in die Technologie entscheidende Faktoren sind.

Datenschutzanforderungen können den Einsatz von KI (in der Cloud) verhindern (z. B. können personenbezogene Daten in Sensordaten enthalten sein). Die erforderliche Hardware für lokale Anwendungen ist teuer und oft nicht verfügbar. Auch die Effizienz der Modelle ist manchmal ein Problem, z. B. würde es zu lange dauern, wenn jede E-Mail von einem LLM überprüft würde.

6.2 Andreas Tomek, KPMG

DI Mag. Andreas Tomek ist Partner bei KPMG Österreich im Bereich IT Advisory. Seine Schwerpunkte umfassen Informations- und Cyber Security, Audit und Training, Penetration Testing sowie innovative Sicherheitslösungen. Seit 2020 fungiert er als „Global Cyber Leader Cloud Security“ bei KPMG International. Zudem ist er Initiator des europäischen Start-up-Wettbewerbs Security Rockstars und als Vortragender bei diversen Konferenzen sowie Universitäten tätig. Vor seiner Tätigkeit bei KPMG war er Geschäftsführer der ISCP GmbH und Mitglied der Geschäftsleitung des Forschungszentrums SBA Research.

Der Einsatz von KI in Unternehmen bringt nicht nur Vorteile, sondern auch neue Sicherheitsrisiken mit sich. Neben potenziellen Schwachstellen in den KI-Anwendungen selbst sind auch rechtliche Aspekte ungeklärt. Der breite Zugriff auf Unternehmensdaten ist eine zentrale Herausforderung, insbesonde-

re die Datenqualität und das Zugriffsmanagement. KI-Systeme machen (auch sensible) Informationen leichter auffindbar, was die Sicherheit zusätzlich erschwert.

Auf der Angriffsseite eröffnen KI-Technologien wie Deepfakes und automatisiertes Voice-Spoofing neue Möglichkeiten, etwa für Phishing- oder Spam-Attacken. Beim Angriff können Skalierungseffekte durch KI genutzt werden. Der großflächige Einsatz von „intelligenter Ransomware“ könnte katastrophale Folgen haben, zumal die Kapazitäten zur Reaktion auf Vorfälle derzeit begrenzt sind. Unternehmen stehen vor der Herausforderung, nicht nur Angriffe abzuwehren, sondern auch geschultes Personal in ausreichender Zahl bereitzustellen.

Innerhalb der Sicherheitsbranche treiben KI-Tools wie Copilot die Automatisierung voran und erleichtern Sicherheitsoperationen durch benutzungsfreundliche Playbooks und Schnittstellen, die keine Programmierkenntnisse erfordern. Es gibt jedoch technische Einschränkungen, wie z. B. die begrenzte Kontextgröße aktueller Sprachmodelle. Diese könnten in den kommenden Jahren durch neue Hardware und größere Modelle überwunden werden, was neue Anwendungsfälle ermöglichen würde. Derzeit liegt der Schwerpunkt auf der Verbesserung der Nutzungsfreundlichkeit und der Integration von KI in bestehenden Arbeitsabläufen. Die Einführung von KI hat das Potenzial, Fachkräfte effizienter zu machen. So können z. B. Nachwuchskräfte mit KI-Unterstützung Aufgaben übernehmen, die bisher erfahrenen Fachkräften vorbehalten waren. Gleichzeitig birgt diese Entwicklung die Gefahr, dass grundlegende Fähigkeiten und tiefes Verständnis verloren gehen, da die Abhängigkeit von KI-Tools zunimmt. Sprachmodelle wie LLMs könnten auch eine wichtige Rolle in der Aus- und Weiterbildung spielen, insbesondere im Sicherheitsbereich.

Trotz des Potenzials stehen Unternehmen noch vor großen Hürden. Viele haben bereits KI-Lizenzen wie Copilot erworben, wissen aber nicht, wie sie diese produktiv einsetzen können. Während cloudbasierte Lösungen bevorzugt werden, bleiben *Legacy-Systeme* oft außen vor, da Investitionen in sie als weniger rentabel gelten. Während der erste Hype um generative KI bereits abgeklungen ist - Unternehmen haben erkannt, dass damit nicht alles gelöst werden kann - wird die nächste Entwicklungsphase zeigen, ob Unternehmen die Herausforderungen meistern und die Potenziale von KI voll ausschöpfen können. Derzeit fehlt es auch an guten Weiterbildungsangeboten zu den neuen Technologien und Anwendungsfällen im Sicherheitsbereich. Der Erfolg wird nicht zuletzt davon abhängen, ob es gelingt, innovative Lösungen für repetitive Aufgaben wie *Vulnerability Management* zu finden und damit die Effizienz in der Sicherheitsbranche deutlich zu steigern.

6.3 Simon Leitner, Condignum

Simon Leitner ist Mitgründer und CEO der Condignum GmbH, einem österreichischen Cyber-Security-Unternehmen mit Sitz in Wien. Condignum hat sich zum Ziel gesetzt, die digitale Welt sicherer zu machen, indem es Organisationen jeder Größe unterstützt, ihre Cyberabwehr zu stärken und proaktiv auf Bedrohungen zu reagieren. Simon Leitner verfügt über mehr als zehn Jahre Erfahrung im Bereich der Cybersicherheit. Unter seiner Führung bietet Condignum eine digitale Security Management Plattform (SaaS) an, ergänzt durch darauf aufbauende Security Services und Beratungsleistungen. Neben seiner Tätigkeit bei Condignum engagiert sich Leitner als Vortragender und Experte für Cybersicherheit, insbesondere im Zusammenhang mit der EU-Richtlinie NIS2.

Die Verfügbarkeit von KI-Tools wie GPT-4 hat die Eintrittsbarrieren bei Angriffen massiv gesenkt. Angriffe, die früher tiefes technisches Wissen erforderten, können heute ohne Programmierkenntnisse durchgeführt werden. Einfache Eingaben wie IP-Adressen reichen aus, um Schwachstellenscans oder gezielte Angriffe durchzuführen. Schadsoftware kann ebenfalls effizienter erstellt werden.

KI ermöglicht eine nie dagewesene Automatisierung und Skalierbarkeit von Angriffen. Social-Engineering-Kampagnen können innerhalb weniger Stunden durchgeführt werden. Selbst komplexe Angriffe wie Voice Phishing oder gefälschte Meetings mit realistisch generierten Videos von Führungskräften (CxO Fraud) werden durch KI erleichtert.

Hier zeigt sich das klassische Verteidigungsdilemma - während die Angriffsseite von KI profitiert, kämpft die Verteidigungsseite mit regulatorischen Hürden, Budgetkürzungen und einer zunehmenden Komplexität der Bedrohungslandschaft. Zwar gibt es Fortschritte bei der Bedrohungserkennung, etwa durch KI-gestützte Anomalieerkennungs-Tools, aber viele Systeme scheitern an der Verknüpfung und Validierung von Daten. Darüber hinaus erfordern solche Systeme häufig menschliches Eingreifen, um Anomalien richtig zu interpretieren.

In der Verteidigung wird oft auf KI gesetzt, um Bedrohungen zu erkennen und abzuwehren, beispielsweise durch Orchestrierungstools oder automatisierte Schwachstellenscanner. Die Qualität dieser Lösungen ist jedoch oft noch unzureichend. Gleichzeitig werden KI-Funktionen in Sicherheitsprodukten oft übertrieben beworben.

Regulatorische Anforderungen, insbesondere in Europa, erschweren den Einsatz von KI in der Cybersicherheit. Die Datenqualität und Halluzinationen der KI-Modelle stellen ebenfalls ein Problem dar.

6.4 Markus Cserna, cyan Security Group

Markus Cserna ist Mitbegründer und Geschäftsführer der cyan Security Group GmbH sowie im Vorstand der cyan AG. Seit der Gründung hat er maßgeblich zur Entwicklung des Unternehmens beigetragen, das sich auf intelligente Cybersecurity-Lösungen für Telekommunikationsunternehmen spezialisiert hat. Unter seiner Führung hat cyan Partnerschaften mit internationalen Kunden wie Orange, Claro Chile und Magenta aufgebaut.

KI hat die Dynamik der Cybersicherheit grundlegend verändert und neue Herausforderungen sowohl im Angriff als auch in der Verteidigung geschaffen. Derzeit ist ein klarer Vorteil für die Angriffsseite zu erkennen, die KI gezielt einsetzt, um ihre Methoden zu verbessern und die Effektivität ihrer Angriffe zu erhöhen.

Phishing-Angriffe und gefälschte Webseiten werden durch KI immer raffinierter. Wo früher schlecht formulierte und leicht erkennbare E-Mails dominierten, entstehen heute täuschend echte und sprachlich einwandfreie Inhalte, die selbst für erfahrene Nutzerinnen und Nutzer schwer zu erkennen sind. KI ermöglicht es zudem, Angriffe schnell zu variieren und damit bestehende Methoden der Erkennungsmuster unwirksam zu machen. Insbesondere beim Social Engineering und der Malware-Generierung wird KI eingesetzt, um gezielte Angriffe mit geringem Aufwand und hoher Reichweite durchzuführen. Die Einstiegshürde für Angreiferinnen und Angreifer ist dadurch deutlich gesunken, während die Qualität der Angriffe stetig steigt. Dies erhöht den Druck auf die Verteidigung, deren Reaktionen oft zu langsam ist, um mit den schnellen Veränderungen Schritt zu halten.

Auf der Abwehrseite nutzen Sicherheitsunternehmen zunehmend KI-basierte Technologien, um Angriffe zu erkennen und abzuwehren. Insbesondere ML hat sich bewährt, etwa bei der Klassifizierung von Phishing-Mails und der Erkennung bösartiger Webseiten. Diese statistischen Modelle erkennen bekannte Muster zuverlässig. Daneben gewinnen LLMs wie ChatGPT an Bedeutung. Sie unterstützen Sicherheitsanalystinnen und -analysten bei der Erkennung neuer Bedrohungen, indem sie Gemeinsamkeiten in Angriffsmustern analysieren und dynamisch auf neue Entwicklungen reagieren können. Derzeit liegt der Fokus auf der Unterstützung von Analystinnen und -analysten, langfristig sollen die Erkenntnisse aus LLMs in klassische, automatisierte Systeme integriert werden.

Trotz der Potenziale ist der Einsatz von KI mit erheblichen Herausforderungen verbunden. Die Kosten für cloudbasierte KI-Anwendungen, wie z. B. die Nutzung von LLMs, sind erheblich und können für Unternehmen eine finanzielle Belastung darstellen. Darüber hinaus bleibt die Beschaffung qualitativ hochwertiger Daten ein zentrales Problem, da diese für die Wirksamkeit der Modelle von entscheidender Bedeutung sind. Ein weiteres Hindernis ist die Fehleranfälligkeit von KI-Modellen: Fehler oder Fehlinterpretationen in einem Modell können sich durch die Weitergabe der Ergebnisse an andere Systeme vervielfachen, was eine zusätzliche menschliche Kontrolle erforderlich macht.

Viele Unternehmen werben mit neuartigen KI-basierten Lösungen, oft handelt es sich aber um klassische maschinelle Lernverfahren, die unter dem Begriff KI vermarktet werden. Innovative Anwendungen, die gezielt LLMs einsetzen, sind derzeit noch selten. Der Markt ist nach wie vor unübersichtlich und die Diskrepanz zwischen Versprechen und Realität oft groß.

Ein weiterer Aspekt, der mit dem zunehmenden Einsatz von KI einhergeht, ist der Umgang mit sensiblen Daten. Unternehmen geben sensible Informationen oft bedenkenlos in solche Systeme ein, was langfristig zu neuen Datenschutzproblemen führen kann. Gleichzeitig stellt der AI Act zusätzliche regulatorische Anforderungen vor: Die verwendeten Modelle und Datenquellen müssen detailliert dokumentiert werden, was den administrativen Aufwand erhöht.

6.5 Una-May O'Reilly, MIT

Dr. Una-May O'Reilly ist Principal Research Scientist am Computer Science and Artificial Intelligence Laboratory (CSAIL) des Massachusetts Institute of Technology (MIT) und leitet die Forschungsgruppe „AnyScale Learning For All" (ALFA). Ihre Forschungsschwerpunkte liegen in den Bereichen skalierbares maschinelles Lernen, evolutionäre Algorithmen und der Nutzung künstlicher Intelligenz als Werkzeug für Cyberangriffe.

KI erweitert die Fähigkeiten von Angreifern über das gesamte Spektrum, von Scriptkiddies bis hin zu staatlich finanzierten Akteuren. Besonders hervorzuheben sind Potenziale in der detaillierten Planung und Durchführung komplexer, mehrstufiger Kampagnen. Angreifer:innen profitieren bereits von KI-gestützter Automatisierung, etwa bei der Erstellung überzeugender Phishing-E-Mails oder der Entdeckung und Ausnutzung von Exploits. Diese Effizienzsteigerung stellt Verteidiger:innen vor große Herausforderungen. Auch bei der Simulation von Angriffen, beispielsweise im Rahmen von Red Team-Übungen oder Penetrationstests, nimmt KI eine zunehmend wichtige Rolle ein. Aber auch auf Seite der Verteidigung zeigt KI großes Potenzial, beispielsweise bei der Unterstützung von SOCs. KI-basierte Systeme wie SOC-Copiloten können große Datenmengen analysieren, historische Informationen zusammenführen und Abwehrmaßnahmen schneller und effizienter gestalten.

Unsere Forschung fokussiert sich auf den Einsatz von Sprachmodellen (LLMs) zur Automatisierung von Sicherheitsaufgaben, wie der Ausführung von Bash-Kommandos in Sicherheitstests. Erste Ergebnisse zeigen Fortschritte, machen jedoch auch den Bedarf an qualitativ hochwertigen Trainingsdaten deutlich, die in der Cybersicherheit oft nicht in ausreichender Menge und Qualität zur Verfügung stehen. Ansätze wie Fine-Tuning oder RAG können dabei helfen, diese Daten besser zu nutzen. Spezialisierte, kleine Sprachmodelle können ebenfalls gute Ergebnisse liefern, dennoch sind aktuelle Basismodelle wie GPT von OpenAI oft noch überlegen.

Während Anwendungen von KI in der Cybersicherheit oft skeptisch betrachtet wurden, hat die Weiterentwicklung von Sprachmodellen zu einer größeren Akzeptanz und Einführung geführt. KI ermöglicht präzisere Analysen von Code und Schwachstellen und bietet sowohl operativ als auch strategisch Vorteile. `Act - React - Anticipate` sind die strategischen Ansätze, die die Dynamik zwischen Angreifern und Verteidigern bestimmen. KI kann reaktive Ansätze, wie sie in vielen SOCs verwendet werden, verbessern, indem Bedrohungen schneller erkannt und bekämpft werden. Sie hilft auch, Auswirkungen auf die eigene Organisation besser zu verstehen und Schutzmaßnahmen schneller zu implementieren. Andere vorausschauende Strategien, wie die Simulation von KI-gesteuerten Angreifern und Verteidigern (Red Team vs. Blue Team), fördern die Vorbereitung auf zukünftige Angriffe. Gleichzeitig führt die wachsende Angriffsfläche durch alte und neue Schwachstellen sowie KI-inhärente Verwundbarkeiten zu neuen Angriffsmöglichkeiten. Vor diesem Hintergrund liegt der Vorteil derzeit noch auf der aktiven Seite – bei den Angreifern.

Trotz des großen Potenzials von KI bestehen wesentliche Herausforderungen. Die Erstellung spezialisierter Modelle erfordert umfangreiche und qualitativ hochwertige Trainingsdaten. Darüber hinaus befindet sich die breite Integration von KI-gestützten Systemen (Agenten) in die Cybersicherheit noch in der frühen Entwicklungsphase.

7 KI-basierte Tools für Cybersicherheit

Es sind viele verschiedene Cybersecurity-Tools verfügbar, wovon die meisten mittlerweile angeben, KI zu verwenden. Dazu gehören z. B. Microsoft Defender, DarkTrace, Cisco Splunk, IBM QRadar, Sophos Intercept X, CrowdStrike, und Palo Alto Cortex um nur einige zu nennen. Allerdings sind die Tools schwierig zu vergleichen, da ihr Quellcode nicht offengelegt ist; Ghazal et al. vergleichen beworbenen Produkteigenschaften der genannten Produkte [GHZ+22]. Zusätzlich ist ein Vergleich von Cybersicherheitsprodukte nicht einfach, da die Tools unterschiedliche Ziele verfolgen und daher für unterschiedliche Anwendungsbereiche entwickelt wurden. Darüber hinaus bietet jedes Tool seine eigene Variation an Funktionen und Features an, wodurch ein direkter Vergleich erschwert wird. Zudem gibt es unterschiedliche Leistungsmetriken. So wird beispielsweise bei einem IDS die Effektivität anhand von *True Positives* und *False Positives* gemessen, während es bei Virenscannern oftmals die Scangeschwindigkeit Priorität hat. Außerdem darf nicht vergessen werden, dass sich die Bedrohungslandschaften sehr schnell ändern und ein Tool, das heute gut funktioniert, in ein paar Monaten gegen neue Bedrohungen

eventuell nicht mehr effektiv funktioniert. Sowohl die Wahrnehmung der Benutzerfreundlichkeit als auch die Zahlungsbereitschaft sind sehr individuell.

Aus diesem Grund sehen wir von der Beurteilung kommerzieller Tools ab. Wir werden jedoch auf einige *Open Source Tools* eingehen, die KI verwenden oder diese als *Add-On* anbieten.

7.1 Open-Source Tools Beispiele

Im Folgenden listen wir einige Open-Source Tools und deren Eigenschaften.

Elastic ML und Elastic Security SIEM: Dies sind Erweiterungen zu Elasticsearch[33]. Elastic Machine Learning[34] ermöglicht eine automatische Anomalieerkennung und Ursachenanalyse, wodurch die Reaktionszeit verkürzt wird. Als SIEM-Lösung sammelt, normalisiert und analysiert Elastic Security SIEM Daten aus verschiedenen Quellen innerhalb der IT-Umgebung eines Unternehmens, wie beispielsweise Log-Einträgen, Netzwerkverkehr und Endpunktinformationen. Somit bietet Elastic Security SIEM eine zentrale Plattform für die Überwachung und Verwaltung von Sicherheitsereignissen in Echtzeit.

Latio Application Security Tester: Ein KI-Security-Scanner[35], der den Code von der CLI (Kommandozeile) auf Sicherheits- und Qualitätsprobleme mittels OpenAI und Gemini überprüft.

LLM Guard: Ein Open-Source-Toolkit[36], das die Ein- und Ausgabeinhalte von LLMs mittels den *ML-Modellen* DeBERTa, BGE-M3, RoBERTa analysiert[37], um Sicherheit und *Compliance* in Echtzeit zu gewährleisten.

SonarQube: Dieses Tool[38] bietet statische Code-Qualitäts- und Sicherheitsanalysen, insbesondere für KI-generierten Code aber auch für von Menschen generierten Code. Im Anschluss an die statische Codeanalyse wird ein LLM verwendet, um Sicherheitslücken aufzuzeigen und Code zur Behebung dieser Lücken zu generieren.

Taranis AI: Ein Open-Source Intelligence (OSINT) Tool[39] von der Austrian Institute of Technology GmbH (AIT). Taranis AI sammelt öffentliche Informationen über potenzielle Sicherheitsbedrohungen, Schwachstellen, Trends, neue Risiken, etc., um die Situation im Auge zu behalten und bei Angriffen frühzeitig Gegenmaßnahmen zu ergreifen. Mit den gesammelten Informationen erstellt Transis AI Artikel über die verschiedenen Ereignisse. Anschließend verfeinern Analystinnen und Analysten diese KI-generierten Artikel zu strukturierten Berichten.

Vigil: Eine *Python-Library*[40] und REST-API, die *LLM Prompt Injections*, *Jailbreaks* und andere potenzielle Bedrohungen erkennt, indem die LLM Prompts gescannt werden.

Wazuh: Eine kostenlose Open-Source-Sicherheitsplattform, die als Security Information and Event Management (SIEM) eingesetzt wird, um Bedrohungen zu erkennen, Sicherheitsereignisse zu überwachen und Compliance-Anforderungen zu erfüllen.

8 Fazit

Diese Kurzstudie hat die transformative Wirkung der künstlichen Intelligenz (KI) auf die Cybersicherheit hervorgehoben und ihr Potenzial zur Verbesserung sowohl der defensiven als auch der offensiven Fähigkeiten aufgezeigt. KI-basierte Anwendungen wie Anomalieerkennung, automatisierte Re-

[33] https://www.elastic.co/de/elasticsearch
[34] https://www.elastic.co/elasticsearch/machine-learning
[35] https://www.latio.tech/ und https://github.com/latiotech/LAST
[36] https://github.com/protectai/llm-guard, https://protectai.com/llm-guard
[37] https://github.com/protectai/llm-guard/blob/main/llm_guard/input_scanners/ban_topics.py
[38] https://www.sonarsource.com/
[39] https://taranis.ai/
[40] https://github.com/deadbits/vigil-llm und https://vigil.deadbits.ai/

aktion auf Vorfälle und die Analyse von Bedrohungsdaten bieten erhebliche Vorteile in Bezug auf Geschwindigkeit, Skalierbarkeit und Effizienz. Die zunehmende Verbreitung und Zugänglichkeit von KI-Technologien senkt jedoch auch die Einstiegshürden für Angreifer. Dies ermöglicht es auch technisch weniger versierten Einzelpersonen oder Gruppen, komplexe Cyberangriffe durchzuführen, z. B. mit personalisierten Phishing-Angriffen, täuschend echten Deepfakes oder anpassungsfähiger Malware.

Organisationen müssen ein Gleichgewicht zwischen Innovation und solider Governance finden, interdisziplinäres Fachwissen aufbauen, der Datenintegrität Priorität einräumen, ihre Widerstandsfähigkeit gegenüber neu aufkommenden Bedrohungen stärken und die Einhaltung regulatorischer Standards sicherstellen. Während der erste Hype um generative KI bereits abgeklungen ist, wird die nächste Entwicklungsphase zeigen, ob Unternehmen in der Lage sind, das Potenzial der KI in der Cybersicherheit voll auszuschöpfen.

A Appendix

Im Appendix bieten wir zusätzlich relevantes Hintergrundwissen an, welches zu einem besseren Verständnis der Studie beitragen soll.

A.1 Malware

A.1.1 Arten von Malware

Im folgenden Abschnitt werden verschiedene Malware-Typen beschrieben [QKC19].

Computervirus: Ein Computervirus ist ein Code, der sich über Dokumente, Skriptdateien oder Webanwendungen vervielfältigen und sich über ein System verbreiten kann.

Worm/Computerwurm: Ein Computerwurm ist ein Programm, das sich ohne menschliches Zutun über ein Gerät oder auf andere Geräte im selben Netzwerk verbreiten kann, indem es sich selbst repliziert.

Trojan/Trojaner: Ein Trojaner ist eine Form von Malware, die sich als harmloses Programm tarnt, um Benutzer:innen zur Installation zu bewegen.

A.1.2 Malware Detection Analysis

Die Methoden zur Malware-Erkennung können in drei verschiedenen Typen eingeteilt werden: *Static* (statische), *Dynamic* (dynamische), oder *Hybrid Method* (hybride Methode). [DTV+17]

Static Method: verwendet statische Merkmale durch Dekompilierung der Zieldatei. Die statische Analyse hat den Vorteil, dass diese schnell und effizient ist. Jedoch könnte diese Methode durch die Verwendung von Verschleierungstools, die den Computerschädling vertuschen, eingeschränkt sein [GAJ24].

Dynamic Method: umfasst die Überwachung des dynamischen Systemverhaltens, Snapshot, Fehlersuche usw. Die dynamische Analyse wird nicht durch Verschleierung beeinträchtigt, kann jedoch durch weit verbreitete Anti-Analysetechniken umgangen werden und erfordert zudem mehr Rechenleistung [GAJ24]. Bei dieser Methodik wird die Datei ausgeführt und ihr Verhalten analysiert. In der statischen Analyse hingegen wird die Datei ohne Ausführung untersucht [BKB24].

Hybrid Method: ist eine Mischung zwischen der statischen und der dynamischen Methode [GAJ24].

A.2 Techniken zur Malware-Erkennung

Dieser Abschnitt enthält eine Liste herkömmlicher Verfahren zur Erkennung von Cyber-Bedrohungen.

Signature-Based Technique: zerlegt den Code einer infizierten Datei und sucht nach einem Muster, das auf eine Malware-Familie hinweist oder einer Signatur (z. B. einer Bitfolge) [GSCK23]. Bei dieser Technik muss das Muster oder die Signatur im Voraus bekannt sein.

Behavior-Based Technique: basiert auf dem Verhalten der Malware. Dadurch werden die Einschränkungen der Signature-Based Technik überwunden und Zero-Day-Malware kann erkannt werden [GSCK23].

Heuristic-Based Technique: unterscheidet zwischen dem normalen und dem ungewöhnlichen Verhalten eines Systems. Der erste Schritt besteht darin, das normale Verhalten des Systems zu verfolgen. Im zweiten Schritt wird das Verhalten des Systems während eines Angriffs überwacht und mit dem normalen Verhalten aus Schritt eins verglichen [GSCK23].

Permission-Based Technique: überwacht und analysiert jede Genehmigungsanfrage, um einen möglichen Missbrauch zu erkennen [LL14].

Image-Based Technique; wandelt Rohdaten (z. B. Binärdateien, Netzwerkverkehr oder Protokolle) in visuelle Darstellungen um. Bei den visuellen Darstellungen kann es sich um grau skalierte oder farbige Bilder handeln, bei denen jedes Byte einem Pixel in einem Bild zugeordnet wird. Anschließend wird ein ML-Modell trainiert, um Malware, Anomalien oder andere Cyberbedrohungen auf dem Bild zu erkennen. [NKJM11]

A.3 CIA-Triade

Die CIA-Triade[41] wurde entwickelt, um Organisationen durch Sicherheitsrichtlinien zu leiten. Dabei steht das C für *Confidentiality* (Vertraulichkeit), das I für *Integrity* (Integrität) und das A für *Availability* (Verfügbarkeit).

Confidentiality: bedeutet, dass die unbefugte Weitergabe von Informationen, insbesondere bei vertraulicher Daten, verhindert wird. Im Kontext von ML können dies die Trainings-, Validierungs- und Testdaten oder das trainierte Modell selbst sein. Angriffe auf die Vertraulichkeit umfassen im Zusammenhang mit KI *Membership Inference Attacks*, *Property Inference Attacks*, *Model Inversion*, oder *Attribute Inference Attacks*.

Integrity: bedeutet, dass die Daten oder das ML-Modell vertrauenswürdig und unverändert sind. Dieser Grundsatz ist entscheidend, um sicherzustellen, dass die Daten in einem korrekten Zustand bleiben. Niemand sollte die Möglichkeit haben, die Daten unbeabsichtigt oder böswillig zu verändern oder zu erweitern, es sei denn, dies ist notwendig. Angriffe auf die Integrität im Zusammenhang mit KI sind *Evasion Attacks*, *Backdoor Poisoning*, *Model Poisoning*, *Targeted Poisoning Attacks* und *Deepfakes*.

Availability: stellt sicher, dass autorisierte Benutzer:innen jederzeit auf die benötigten Systeme, Daten oder Ressourcen zugreifen können. Das bedeutet, dass das gesamte System jederzeit verfügbar sein muss, wenn es benötigt wird. Angriffe auf die Verfügbarkeit im Zusammenhang mit KI sind *Model Poisoning*, *Clean-Label Poisoning*, *Data Poisoning Attacks* und *Energy-Latency Attacks*.

[41] https://www.nccoe.nist.gov/publication/1800-26/VolA/index.html

Abkürzungsverzeichnis

AIT Austrian Institute of Technology GmbH. 26

BOF Buffer Overflow. 6

CANs Controller Area Networks. 11

CLI Command Line Interface. 26

CSF Cybersecurity Framework. 3, 6

CSRF Cross-Site-Request-Forgery. 6

CTF Capture The Flag. 15

DoS Denial-of-Service. 4, 6

DSS Decision Support Systems. 9

DT Decision Trees. 11

GenAI Generative AI (Generative Künstliche Intelligenz). 20

IDS Intrusion Detection System. 11, 25

IoC Indicators of Compromise. 3, 13

IODEF Incident Object Description Exchange Format. 12

IPS Intrusion Prevention System. 11

KNN K-Nearest Neighbor. 11

MITRE Atlas Adversarial Threat Landscape for Artificial-Intelligence Systems. 19

ML Machine Learning. 1, 6, 9–11, 13, 16, 19, 20, 26, 29

NB Naïve Bayes. 9, 11

NIST National Institute of Standards and Technology. 3, 6, 19

NN Neural Network. 12, 20

OS Betriebssystem. 6

OSINT Open-Source Intelligence. 26

OWASP Open Web Application Security Project. 19

RAG Retrieval-Augmented Generation. 25

RCE Remote Code-Execution. 6

RF Random Forest. 9, 10

SaaS Software-as-a-Service. 23

SDN Software-Defined Networking. 9

SIEM Security Information and Event Management. 22, 26

SOC Security Operations Center. 22, 25

SVM Support Vector Machines. 10, 11

TF-IDF Term Frequency - Inverse Document Frequency. 11

XSS Cross-Site-Scripting. 4, 6

Literatur

[ABC+24] Franco Algieri, Günther Barnet, Marie Janine Calic, Pádraig Carmody, Gustav Gressel, Bruno Hofbauer, Arnold Kammel, Herfried Münkler, Walter Posch, Bernhard Richter, Nikolaus Rottenberger, and Andreas W. Stupka. *Risikobild 2024 - Welt aus den Fugen*. Sicherheitspolitische Analysen. Bundesministerium für Landesverteidigung (BMLV), January 2024.

[AKK+20] Masaki Aota, Hideaki Kanehara, Masaki Kubo, Noboru Murata, Bo Sun, and Takeshi Takahashi. Automation of Vulnerability Classification from its Description using Machine Learning. In *2020 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, July 2020. ISSN: 2642-7389.

[AMM+20] Neda Afzaliseresht, Yuan Miao, Sandra Michalska, Qing Liu, and Hua Wang. From logs to Stories: Human-Centred Data Mining for Cyber Threat Intelligence. *IEEE Access*, 8:19089–19099, 2020. Conference Name: IEEE Access.

[AMS+15] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137, 2015.

[BKB24] Ahmed Bensaoud, Jugal Kalita, and Mahmoud Bensaoud. A survey of malware detection using deep learning. *Machine Learning with Applications*, 16:100546, June 2024.

[BL17] Ruth Bearden and Dan Chai-Tien Lo. Automated microsoft office macro malware detection using machine learning. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4448–4452, December 2017.

[Bot23] Marcus Botacin. GPThreats-3: Is Automatic Malware Generation a Threat? In *2023 IEEE Security and Privacy Workshops (SPW)*, pages 238–254, 2023.

[BPC23] Mika Beckerich, Laura Plein, and Sergio Coronado. RatGPT: Turning online LLMs into Proxies for Malware Attacks, September 2023. arXiv:2308.09183 [cs].

[Bur24] Elie Bursztein. How Large Language Models Are Reshaping the Cybersecurity Landscape. https://www.youtube.com/watch?v=UicS9i3OKTs, 2024.

[BV21] Adel Binbusayyis and Thavavel Vaiyapuri. Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM. *Applied Intelligence*, 51(10):7094–7108, October 2021.

[CC23] Enrico Cambiaso and Luca Caviglione. Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources, 2023.

[CCAS23] P. V. Sai Charan, Hrushikesh Chunduri, P. Mohan Anand, and Sandeep K Shukla. From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads, 2023.

[CCTDZ21] Miguel V. Carriegos, Ángel L. Muñoz Castañeda, M. T. Trobajo, and Diego Asterio De Zaballa. On Aggregation and Prediction of Cybersecurity Incident Reports. *IEEE Access*, 9:102636–102648, 2021. Conference Name: IEEE Access.

[CO22] Alesia Chernikova and Alina Oprea. FENCE: Feasible Evasion Attacks on Neural Networks in Constrained Environments. *ACM Trans. Priv. Secur.*, 25(4):34:1–34:34, July 2022.

[CP21] Michal Choraś and Marek Pawlicki. Intrusion detection approach based on optimised artificial neural network. *Neurocomputing*, 452:705–715, September 2021.

[CPV22] Marta Catillo, Antonio Pecchia, and Umberto Villano. AutoLog: Anomaly detection by deep autoencoding of system logs. *Expert Syst. Appl.*, 191(C), April 2022.

[CS24] Canyu Chen and Kai Shu. Can LLM-Generated Misinformation Be Detected?, 2024.

[CSC+21] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, ACSAC '21, pages 554–569, New York, NY, USA, December 2021. Association for Computing Machinery.

[CTW+21] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models, 2021.

[dJCdSW24] Gabriel de Jesus Coelho da Silva and Carlos Becker Westphall. A Survey of Large Language Models in Cybersecurity, 2024.

[DLMV+24a] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 847–864, Philadelphia, PA, August 2024. USENIX Association.

[DLMV+24b] Gelei Deng, Yi Liu, Víctor Mayoral-Vilches, Peng Liu, Yuekang Li, Yuan Xu, Tianwei Zhang, Yang Liu, Martin Pinzger, and Stefan Rass. PentestGPT: An LLM-empowered Automatic Penetration Testing Tool, 2024.

[DMCF20] Noemí DeCastro-García, Ángel L. Muñoz Castañeda, and Mario Fernández-Rodríguez. Machine learning for automatic assignment of the severity of cybersecurity events. *Computational and Mathematical Methods*, 2(1), January 2020.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '03, pages 202–210, New York, NY, USA, June 2003. Association for Computing Machinery.

[DP24] Dinil Mon Divakaran and Sai Teja Peddinti. LLMs for Cyber Security: New Opportunities, 2024.

[DTN21] Alexandre Dey, Eric Totel, and Sylvain Navers. Heterogeneous Security Events Prioritization Using Auto-encoders. In Joaquin Garcia-Alfaro, Jean Leneutre, Nora Cuppens, and Reda Yaich, editors, *Risks and Security of Internet and Systems*, pages 164–180, Cham, 2021. Springer International Publishing.

[DTV+17] Anusha Damodaran, Fabio Di Troia, Corrado Aaron Visaggio, Thomas H. Austin, and Mark Stamp. A comparison of static, dynamic, and hybrid analysis for malware detection. *Journal of Computer Virology and Hacking Techniques*, 13(1):1–12, February 2017.

[EGLPA20] Tiago Espinha Gasiba, Ulrike Lechner, and Maria Pinto-Albuquerque. Sifu - a cybersecurity awareness platform with challenge assessment and intelligent coach. *Cybersecurity*, 3(1):24, December 2020.

[Fal23] Polra Victor Falade. Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(5):185–198, October 2023.

[FDAFPK+21] Paulo Freitas De Araujo-Filho, Antônio J. Pinheiro, Georges Kaddoum, Divanilson R. Campelo, and Fabio L. Soares. An Efficient Intrusion Prevention System for CAN: Hindering Cyber-Attacks With a Low-Cost Platform. *IEEE Access*, 9:166855–166869, 2021. Conference Name: IEEE Access.

[FGP+21] Alessandro Fausto, Giovanni Battista Gaggero, Fabio Patrone, Paola Girdinio, and Mario Marchese. Toward the Integration of Cyber and Physical Security Monitoring Systems for Critical Infrastructures. *Sensors*, 21(21):6970, January 2021. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute.

[fIuPPSAG] Bundesministerium für Inneres/Bundeskriminalamt und PSA Payment Services Austria GmbH. Gemeinsam gegen Phishing. https://www.wko.at/oe/it-sicherheit/gemeinsam-gegen-phishing.pdf. Accessed: 11-11-2024.

[FMHCPG+19] Lorenzo Fernández Maimó, Alberto Huertas Celdrán, Ángel L. Perales Gómez, Félix J. García Clemente, James Weimer, and Insup Lee. Intelligent and Dynamic Ransomware Spread Detection and Mitigation in Integrated Clinical Environments. *Sensors*, 19(5):1114, January 2019. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[GAJ24] Matthew G. Gaber, Mohiuddin Ahmed, and Helge Janicke. Malware Detection with Artificial Intelligence: A Systematic Literature Review. *ACM Computing Surveys*, 56(6):1–33, June 2024.

[GDSDBV+20] Eder Souza Gualberto, Rafael Timoteo De Sousa, Thiago Pereira De Brito Vieira, João Paulo Carvalho Lustosa Da Costa, and Cláudio Gottschalg Duque. The Answer is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering. *IEEE Access*, 8:223529–223547, 2020. Conference Name: IEEE Access.

[GFH+20] Jonas Geiping, Liam H. Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching. In *International Conference on Learning Representations*, October 2020.

[GHZ+22] Taher M. Ghazal, Mohammad Kamrul Hasan, Raed Abu Zitar, Nidal A. Al-Dmour, Waleed T. Al-Sit, and Shayla Islam. Cybers Security Analysis and Measurement Tools Using Machine Learning Approach. In *2022 1st International Conference on AI in Cybersecurity (ICAIC)*, pages 1–4, May 2022.

[GJB22] Neha Gupta, Vinita Jindal, and Punam Bedi. CSE-IDS: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems. *Comput. Secur.*, 112(C), January 2022.

[GLDGG19] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7:47230–47244, 2019.

[GMBV21] Luigi Gallo, Alessandro Maiello, Alessio Botta, and Giorgio Ventre. 2 Years in the anti-phishing group of a large company. *Computers & Security*, 105:102259, June 2021.

[Gro23] Insikt Group. I, Chatbot. Technical report, Recorded Future, January 2023. https://www.recordedfuture.com/research/i-chatbot.

[GSCK23] Nor Zakiah Gorment, Ali Selamat, Lim Kok Cheng, and Ondrej Krejcar. Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges, and Future Directions. *IEEE Access*, 11:141045–141089, 2023. Conference Name: IEEE Access.

[GSP24] Dhruva Goyal, Sitaraman Subramanian, and Aditya Peela. Hacking, The Lazy Way: LLM Augmented Pentesting, 2024.

[GYW21] Haixuan Guo, Shuhan Yuan, and Xintao Wu. LogBERT: Log Anomaly Detection via BERT. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[Haz23] Julian Hazell. Spear Phishing With Large Language Models, 2023.

[HC23] Andreas Happe and Jürgen Cito. Getting pwn'd by AI: Penetration Testing with Large Language Models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2023, page 2082–2086, New York, NY, USA, 2023. Association for Computing Machinery.

[HKC23] Andreas Happe, Aaron Kaplan, and Jürgen Cito. Evaluating LLMs for Privilege-Escalation Scenarios. *CoRR*, abs/2310.11409, 2023.

[HKC24] Andreas Happe, Aaron Kaplan, and Juergen Cito. LLMs as Hackers: Autonomous Linux Privilege Escalation Attacks, 2024.

[HMLL21] Philip Huff, Kylie McClanahan, Thao Le, and Qinghua Li. A Recommender System for Tracking Vulnerabilities. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ARES '21, pages 1–7, New York, NY, USA, August 2021. Association for Computing Machinery.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, August 2008.

[HSV+23] Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. Devising and Detecting Phishing: Large Language Models vs. Smaller Human Models, 2023.

[HYT23] Xiao Han, Shuhan Yuan, and Mohamed Trabelsi. LogGPT: Log Anomaly Detection via GPT, 2023.

[HZ24] Junjie Huang and Quanyan Zhu. PenHeal: A Two-Stage LLM Framework for Automated Pentesting and Optimal Remediation, 2024.

[ISCK24] Isamu Isozaki, Manil Shrestha, Rick Console, and Edward Kim. Towards Automated Penetration Testing: Introducing LLM Benchmark, Analysis, and Improvements, 2024.

[JGCX14] Fei Jiang, Tianlong Gu, Liang Chang, and Zhoubo Xu. Case Retrieval for Network Security Emergency Response Based on Description Logic. In Zhongzhi Shi, Zhaohui Wu, David Leake, and Uli Sattler, editors, *Intelligent Information Processing VII*, pages 284–293, Berlin, Heidelberg, 2014. Springer.

[KAD24] Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. Watch Your Language: Investigating Content Moderation with Large Language Models, 2024.

[KFNC24] Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. Detecting Phishing Sites Using ChatGPT, 2024.

[KGK23] Ramanpreet Kaur, Dušan Gabrijelčič, and Tomaž Klobučar. Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, 97:101804, September 2023.

[KIP10] Huy Kang Kim, Kwang Hyuk Im, and Sang Chan Park. DSS for computer security incident response applying CBR and collaborative response. *Expert Syst. Appl.*, 37(1):852–870, January 2010.

[KLJL20] Gyeongmin Kim, Chanhee Lee, Jaechoon Jo, and Heuiseok Lim. Automatic extraction of named entities of cyber threats using a deep Bi-LSTM-CRF network. *International Journal of Machine Learning and Cybernetics*, 11(10):2341–2355, October 2020.

[kN21] V. Sampath kumar and V. Lakshmi Narasimhan. Using Deep Learning For Assessing Cybersecurity Economic Risks In Virtual Power Plants. In *2021 7th International Conference on Electrical Energy Systems (ICEES)*, pages 530–537, February 2021.

[KSCS24] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. LLM-Mod: Can Large Language Models Assist Content Moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[KY21] Irina Kraeva and Gulnara Yakhyaeva. Application of the Metric Learning for Security Incident Playbook Recommendation. In *2021 IEEE 22nd International Conference of Young Professionals in Electron Devices and Materials (EDM)*, pages 475–479, June 2021. ISSN: 2325-419X.

[LCLW24] Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang. Malla: Demystifying real-world large language model integrated malicious services. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4693–4710, Philadelphia, PA, August 2024. USENIX Association.

[LKC+18] Yung-Feng Lu, Chin-Fu Kuo, Hung-Yuan Chen, Chang-Wei Chen, and Shih-Chun Chou. A SVM-Based Malware Detection Mechanism for Android Devices. In *2018 International Conference on System Science and Engineering (ICSSE)*, pages 1–6, June 2018. ISSN: 2325-0925.

[LL14] Xing Liu and Jiqiang Liu. A Two-Layered Permission-Based Android Malware Detection Scheme. In *2014 2nd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering*, pages 142–148, April 2014.

[LLD+21] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden Backdoors in Human-Centric Language Models. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3123–3140, New York, NY, USA, November 2021. Association for Computing Machinery.

[LWY+19] Yue Lin, He Wang, Bowen Yang, Mingrui Liu, Yin Li, and Yuqing Zhang. A Blackboard Sharing Mechanism for Community Cyber Threat Intelligence Based on Multi-Agent System. In *Machine Learning for Cyber Security: Second International Conference, ML4CS 2019, Xi'an, China, September 19-21, 2019, Proceedings*, pages 253–270, Berlin, Heidelberg, September 2019. Springer-Verlag.

[LZH21] Duc C. Le and Nur Zincir-Heywood. Anomaly Detection for Insider Threats Using Unsupervised Ensembles. *IEEE Transactions on Network and Service Management*, 18(2):1152–1164, June 2021. Conference Name: IEEE Transactions on Network and Service Management.

[MCCL20] KYLE MILLAR, ADRIEL CHENG, HONG GUNN CHEW, and CHENG-CHEW LIM. Operating System Classification: A Minimalist Approach. In *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 143–150, December 2020. ISSN: 2160-1348.

[MHM+24] Farzad Nourmohammadzadeh Motlagh, Mehrdad Hajizadeh, Mehryar Majd, Pejman Najafi, Feng Cheng, and Christoph Meinel. Large Language Models in Cybersecurity: State-of-the-Art, 2024.

[ML22] Yisroel Mirsky and Wenke Lee. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1):1–41, January 2022.

[MLHO23] Stephen Moskal, Sam Laney, Erik Hemberg, and Una-May O'Reilly. LLMs Killed the Script Kiddie: How Agents Supported by Large Language Models Change the Landscape of Network Threat Testing, 2023.

[MM21] Benjamin S. Meyers and Andrew Meneely. An Automated Post-Mortem Analysis of Vulnerability Relationships using Natural Language Word Embeddings. *Procedia Computer Science*, 184:953–958, 2021.

[MMHC23] Antonio Monje, Alejandro Monje, Roger Hallman, and George Cybenko. Being a Bad Influence on the Kids: Malware Generation in Less Than Five Minutes Using ChatGPT. Publisher: Unpublished, 2023.

[MP17] Aziz Makandar and Anita Patrot. Malware class recognition using image processing techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, pages 76–80, February 2017. Conference Name: 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) ISBN: 9781509040834 Place: Pune Publisher: IEEE.

[NBC+08] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. In *Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, LEET'08, pages 1–9, USA, April 2008. USENIX Association.

[NCB+19] Raul Ceretta Nunes, Marcelo Colomé, Fabio André Barcelos, Marcelo Garbin, Gustavo Bathu Paulus, and Luis Alvaro De Lima Silva. A Case-Based Reasoning Approach for the Cybersecurity Incident Recording and Resolution. *International Journal of Software Engineering and Knowledge Engineering*, 29(11n12):1607–1627, November 2019.

[NCE17] Fitzroy Nembhard, Marco Carvalho, and Thomas Eskridge. A hybrid approach to improving program security. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, November 2017.

[NCE19] Fitzroy D. Nembhard, Marco M. Carvalho, and Thomas C. Eskridge. Towards the application of recommender systems to secure coding. *EURASIP Journal on Information Security*, 2019(1):9, June 2019.

[NKJM11] L. Nataraj, S. Karthikeyan, G. Jacob, and B. S. Manjunath. Malware images: visualization and automatic classification. In *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, pages 1–7, Pittsburgh Pennsylvania USA, July 2011. ACM.

[NN20] Hoang Hai Nguyen and David M. Nicol. Estimating Loss Due to Cyber-Attack in the Presence of Uncertainty. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 361–369, December 2020. ISSN: 2324-9013.

[NNN18] Minh Nguyen, Toan Nguyen, and Thien Huu Nguyen. A Deep Learning Model with Hierarchical LSTMs and Supervised Attention for Anti-Phishing, May 2018. arXiv:1805.01554.

[OCW19] Olufemi Odegbile, Shigang Chen, and Yuanda Wang. Dependable Policy Enforcement in Traditional Non-SDN Networks. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 545–554, July 2019. ISSN: 2575-8411.

[PDL+06] R. Perdisci, D. Dagon, Wenke Lee, P. Fogla, and M. Sharif. Misleading worm signature generators using deliberate noise injection. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15 pp.–31, May 2006. ISSN: 2375-1207.

[PMJ+20] Aritran Piplai, Sudip Mittal, Anupam Joshi, Tim Finin, James Holt, and Richard Zak. Creating Cybersecurity Knowledge Graphs From Malware After Action Reports. *IEEE Access*, 8:211691–211703, 2020. Conference Name: IEEE Access.

[PPTK+23] Yin Minn Pa Pa, Shunsuke Tanizaki, Tetsui Kou, Michel van Eeten, Katsunari Yoshioka, and Tsutomu Matsumoto. An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop*, CSET '23, page 10–18, New York, NY, USA, 2023. Association for Computing Machinery.

[PRT21] C. Ponsard, V. Ramon, and M. Touzani. Improving Cyber Security Risk Assessment by Combined Use of i* and Infrastructure Models. In *CEUR Workshop Proceedings*, volume 2983, pages 63–69, 2021.

[PSS19] Vitaly G. Promyslov, Kirill V. Semenkov, and Alexander S. Shumov. A Clustering Method of Asset Cybersecurity Classification. *IFAC-PapersOnLine*, 52(13):928–933, January 2019.

[PTA+23] Hammond Pearce, Benjamin Tan, Baleegh Ahmad, Ramesh Karri, and Brendan Dolan-Gavitt. Examining Zero-Shot Vulnerability Repair with Large Language Models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2339–2356, 2023.

[PW19] Jomon A. Paul and Xinfang (Jocelyn) Wang. Socially optimal IT investment for cybersecurity. *Decision Support Systems*, 122:113069, July 2019.

[PZ21] Jomon A. Paul and Minjiao Zhang. Decision support model for cybersecurity risk planning: A two-stage stochastic programming framework featuring firms, government, and attacker. *European Journal of Operational Research*, 291(1):349–364, May 2021.

[Qa19] Osama Mohammed Qasim and Karim Hashim alsadi. Detection System for Detecting Worms using Hybrid Algorithm of Naïve Bayesian classifier and K-Means. In *2019 2nd International Conference on Engineering Technology and its Applications (IICETA)*, pages 173–178, August 2019.

[QHL18] Yaobin Qin, Brandon Hoffmann, and David J. Lilja. HyperProtect: Enhancing the Performance of a Dynamic Backup System Using Intelligent Scheduling. In *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*, pages 1–8, November 2018. ISSN: 2374-9628.

[QKC19] Attia Qamar, Ahmad Karim, and Victor Chang. Mobile malware attacks: Review, taxonomy & future directions. *Future Gener. Comput. Syst.*, 97(C):887–909, August 2019.

[RDRB11] Loren Paul Rees, Jason K. Deane, Terry R. Rakes, and Wade H. Baker. Decision support for Cybersecurity risk planning. *Decision Support Systems*, 51(3):493–505, June 2011.

[RGSI+21] Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Eider Iturbe, Erkuden Rios, Saturnino Martinez, Antonios Sarigiannidis, Georgios Eftathopoulos, Yannis Spyridis, Achilleas Sesis, Nikolaos Vakakis, Dimitrios Tzovaras, Emmanouil Kafetzakis, Ioannis Giannoulakis, Michalis Tzifas, Alkiviadis Giannakoulias, Michail Angelopoulos, and Francisco Ramos. SPEAR SIEM: A Security Information and Event Management system for the Smart Grid. *Computer Networks*, 193:108008, July 2021.

[RHG+21] Sazid Rahman, Niamat Ullah Ibne Hossain, Kannan Govindan, Farjana Nur, and Mahathir Bappy. Assessing cyber resilience of additive manufacturing supply chain leveraging data fusion technique: A model to generate cyber resilience index of a supply chain. *CIRP Journal of Manufacturing Science and Technology*, 35:911–928, 2021.

[RKI+22] Mamunur Rashid, Joarder Kamruzzaman, Tasadduq Imam, Santoso Wibowo, and Steven Gordon. A tree-based stacking ensemble technique with feature selection for network intrusion detection. *Applied Intelligence*, 52(9):9768–9781, July 2022.

[RNN23] Sayak Saha Roy, Krishna Vamsi Naragam, and Shirin Nilizadeh. Generating Phishing Attacks using ChatGPT, 2023.

[RPCH24] Javier Rando, Fernando Perez-Cruz, and Briland Hitaj. PassGPT: Password Modeling and (Guided) Generation with Large Language Models. In Gene Tsudik, Mauro Conti, Kaitai Liang, and Georgios Smaragdakis, editors, *Computer Security – ESORICS 2023*, pages 164–183, Cham, 2024. Springer Nature Switzerland.

[RYMS19] Nur Adibah Rosli, Warusia Yassin, Faizal M.a, and Siti Rahayu Selamat. Clustering Analysis for Malware Behavior Detection using Registry Data. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(12), June 2019. Number: 12 Publisher: The Science and Information (SAI) Organization Limited.

[SART+24] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-context impersonation reveals large language models' strengths and biases. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[Saw22] Tadeusz Sawik. A linear model for optimal cybersecurity investment in Industry 4.0 supply chains. *International Journal of Production Research*, 60(4):1368–1385, 2022. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00207543.2020.1856442.

[SES+09] Chris Simmons, Charles Ellis, S. Shiva, Dipankar Dasgupta, and Chase Wu. AVOIDIT: A Cyber Attack Taxonomy. *CTIT technical reports series*, January 2009.

[SGJC18] Ankit Shah, Rajesh Ganesan, Sushil Jajodia, and Hasan Cam. Dynamic Optimization of the Level of Operational Effectiveness of a CSOC Under Adverse Conditions. *ACM Trans. Intell. Syst. Technol.*, 9(5):51:1–51:20, April 2018.

[SHN+18] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018.

[SHP+21] Ryan Sheatsley, Blaine Hoak, Eric Pauley, Yohan Beugin, Michael J. Weisman, and Patrick McDaniel. On the Robustness of Domain Constraints. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 495–515, New York, NY, USA, November 2021. Association for Computing Machinery.

[SKDP21] Jacob Sakhnini, Hadis Karimipour, Ali Dehghantanha, and Reza M. Parizi. Physical layer attack identification and localization in cyber–physical grid: An ensemble deep learning based approach. *Physical Communication*, 47:101394, August 2021.

[SMCO21] Giorgio Severi, Jim Meyer, Scott Coull, and Alina Oprea. Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1487–1504. USENIX Association, 2021.

[SPG24] Ashfak Md Shibli, Mir Mehedi A. Pritom, and Maanak Gupta. AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6, 2024.

[SPPM21] Praneet Singh, Jishnu Jaykumar P, Akhil Pankaj, and Reshmi Mitra. Edge-Detect: Edge-Centric Network Intrusion Detection using Deep Neural Network. In *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–6, January 2021. ISSN: 2331-9860.

[SS21a] Injy Sarhan and Marco Spruit. Open-CyKG: An Open Cyber Threat Intelligence Knowledge Graph. *Know.-Based Syst.*, 233(C), December 2021.

[SS21b] Hudan Studiawan and Ferdous Sohel. Anomaly detection in a forensic timeline with deep autoencoders. *J. Inf. Secur. Appl.*, 63(C), December 2021.

[SSD15] Carl Sabottke, Octavian Suciu, and Tudor Dumitraș. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting {Real-World} Exploits. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 1041–1056, 2015.

[SSES+21] Ali I. Siam, Ahmed Sedik, Walid El-Shafai, Atef Abou Elazm, Nirmeen A. El-Bahnasawy, Ghada M. El Banby, Ashraf A.M. Khalaf, and Fathi E. Abd El-Samie. Biosignal classification for human identification based on convolutional neural networks. *International Journal of Communication Systems*, 34(7):e4685, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/dac.4685.

[SSHCFM+19] Pedro Miguel Sánchez Sánchez, Alberto Huertas Celdrán, Lorenzo Fernández Maimó, Gregorio Martínez Pérez, and Guojun Wang. Securing Smart Offices Through an Intelligent and Multi-device Continuous Authentication System. In Guojun Wang, Abdulmotaleb El Saddik, Xuejia Lai, Gregorio Martinez Perez, and Kim-Kwang Raymond Choo, editors, *Smart City and Informatization*, pages 73–85, Singapore, 2019. Springer.

[SVBV24] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024.

[SVCG24] Muris Sladić, Veronica Valeros, Carlos Catania, and Sebastian Garcia. LLM in the Shell: Generative Honeypots. In *2024 IEEE European Symposium on Security and Privacy Workshops*, page 430–435. IEEE, July 2024.

[SWD+22] Jaskaran Singh, Mohammad Wazid, Ashok Kumar Das, Vinay Chamola, and Mohsen Guizani. Machine learning security attacks and defense approaches for emerging cyber physical applications: A comprehensive survey. *Comput. Commun.*, 192(C):316–331, August 2022.

[Tay09] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[TBH+20] Zheyu Tan, Razvan Beuran, Shinobu Hasegawa, Weiwei Jiang, Min Zhao, and Yasuo Tan. Adaptive security awareness training using linked open data datasets. *Education and Information Technologies*, 25(6):5235–5259, November 2020.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction APIs. In *Proceedings of the 25th USENIX Conference on Security Symposium*, SEC'16, pages 601–618, USA, August 2016. USENIX Association.

[vdVZS14] Peter M. van de Ven, Bo Zhang, and Angela Schörgendorfer. Distributed backup scheduling: Modeling and optimization. In *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, pages 1644–1652, April 2014. ISSN: 0743-166X.

[Vic20] Oneil B. Victoriano. Exposing Android Ransomware using Machine Learning. In *Proceedings of the 2019 International Conference on Information System and System Management*, ISSM 2019, pages 32–37, New York, NY, USA, May 2020. Association for Computing Machinery.

[VOFA24] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. Adversarial machine learning : a taxonomy and terminology of attacks and mitigations. Technical Report NIST 100-2e2023, National Institute of Standards and Technology (U.S.), Gaithersburg, MD, January 2024.

[VSI+21] Sridhar Venkatesan, Harshvardhan Sikka, Rauf Izmailov, Ritu Chadha, Alina Oprea, and Michael J. de Lucia. Poisoning Attacks and Data Sanitization Mitigations for Machine Learning Models in Network Intrusion Detection Systems. In *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, pages 874–879, San Diego, CA, USA, November 2021. IEEE Press.

[WDCP22] Mohammad Wazid, Ashok Kumar Das, Vinay Chamola, and Youngho Park. Uniting cyber security and machine learning: Advantages, challenges and future research. *ICT Express*, 8(3):313–321, 2022.

[WGH24] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 33 of *KDD '24*, page 3367–3378. ACM, August 2024.

[WLFL21] Jhih-Ciang Wu, Sherman Lu, Chiou-Shann Fuh, and Tyng-Luh Liu. One-class anomaly detection via novelty normalization. *Comput. Vis. Image Underst.*, 210(C), September 2021.

[WPL15] Bronwyn Woods, Samuel J. Perl, and Brian Lindauer. Data Mining for Efficient Collaborative Information Discovery. In *Proceedings of the 2nd ACM Workshop on Information Sharing and Collaborative Security*, WISCS '15, pages 3–12, New York, NY, USA, October 2015. Association for Computing Machinery.

[WRK+24] Yuhao Wu, Franziska Roesner, Tadayoshi Kohno, Ning Zhang, and Umar Iqbal. SecGPT: An Execution Isolation Architecture for LLM-Based Systems, 2024.

[WSM21] Di Wu, Wei Shi, and Xiangyu Ma. A Novel Real-time Anti-spam Framework. *ACM Trans. Internet Technol.*, 21(4):88:1–88:27, September 2021.

[WZFS21] Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. Concealed Data Poisoning Attacks on NLP Models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online, June 2021. Association for Computational Linguistics.

[XBB+15] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is Feature Selection Secure against Training Data Poisoning? In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1689–1698. PMLR, June 2015. ISSN: 1938-7228.

[YDX+24] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2):100211, June 2024.

[YJW+24] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering, 2024.

[YOIL+21] Abel Yeboah-Ofori, Shareeful Islam, Sin Wee Lee, Zia Ush Shamszaman, Khan Muhammad, Meteb Altaf, and Mabrook S. Al-Rakhami. Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security. *IEEE Access*, 9:94318–94337, 2021.

[ZALT19] Kaiyue Zheng, Laura A. Albert, James R. Luedtke, and Eli Towle. A budgeted maximum multiple coverage model for cybersecurity planning and management. *IISE Transactions*, 51(12):1303–1317, December 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24725854.2019.1584832.

[ZBW+24] Jie Zhang, Haoyu Bu, Hui Wen, Yu Chen, Lun Li, and Hongsong Zhu. When LLMs Meet Cybersecurity: A Systematic Literature Review, 2024.

[ZZS24] Muhammad Nabel Zaharudin, Muhammad Haziq Zuhaimi, and Faysal Hossain Shezan. Enhancing Symbolic Execution with LLMs for Vulnerability Detection. Accessed: 11-11-2024, 2024.