Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Einfluss von KI auf die Cyberbedrohungslandschaft

Änderungshistorie

Version	*Datum*	*Name*	*Beschreibung*
v1.0	10.04.24	TK24	Erstveröffentlichung

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn
E-Mail: ki-kontakt@bsi.bund.de
Internet: https://www.bsi.bund.de

Inhalt

1 Einleitung

Mit dem Aufkommen von Anwendungen, die auf großen Sprachmodellen (eng. „large language models“, LLM) basieren, ist KI auch in der Öffentlichkeit wieder ein viel diskutiertes Thema. Die Grenzen dieser neuen Modelle müssen noch erforscht werden und es ist unklar, welche bleibenden Veränderungen der aktuelle KI-Trend mit sich bringt. Zweifelsohne gibt es Bedenken hinsichtlich der Auswirkungen von KI auf die Cybersicherheit, da sie bereits jetzt die Cyberbedrohungslandschaft sowohl für Angreifer als auch für Verteidiger verändert. Wir untersuchen, wie sich Angriffe und Tätigkeiten von Angreifern durch die neu verfügbare Technologie verändern, wobei wir uns auf die offensive Nutzung von KI konzentrieren. Während generative KI bereits die Qualität und Quantität von Social-Engineering-Angriffen steigert (z. B. Deepfakes oder personalisiertes Phishing in großem Maße), fokussieren wir unsere Diskussion mehr auf technische Angriffsvektoren und weniger auf den menschlichen Faktor. Es sollte jedoch erwähnt werden, dass Social-Engineering-Angriffe zu den häufigsten Angriffen gehören und KI auf diese Art von Angriff starke Auswirkungen hat.

Konkret geht es uns nicht darum, alle Möglichkeiten in diesem weiten Feld zu erörtern. Ziel dieses Berichts ist es, KI-gestützte Anwendungen zu identifizieren, die bereits für den offensiven Einsatz verfügbar sind und zu bewerten, wie sich diese Bedrohungen in naher Zukunft entwickeln könnten. Dazu gehören auch Tools und Anwendungen mit doppeltem Verwendungszweck (sog. Dual-Use-Güter), wie z. B. Penetrationstests, die sowohl beim ethischen Red-Teaming als auch bei kriminellen Aktivitäten helfen können. Wir befassen uns auch mit den Sorgen hinsichtlich einer autonomen Hacker-KI, die gelegentlich in den Medien geäußert werden (1) (2).

In diesem Kapitel stellen wir unsere wichtigsten Ergebnisse und Empfehlungen vor. In den folgenden Kapiteln geben wir einen Überblick über verschiedene Bereiche offensiver KI-Anwendungen.

1.1 Ergebnisse

Auf der Grundlage einer Literaturrecherche und der Bewertung verschiedener Tools und Projekte fassen wir die wichtigsten Ergebnisse wie folgt zusammen:

- KI, insbesondere LLMs, senkt die Einstiegshürden und erhöht Umfang und Geschwindigkeit bösartiger Handlungen, einschließlich der Erstellung von Malware, Social-Engineering-Angriffen und der Datenanalyse im Rahmen von Angriffen. Dies führt zu fähigeren Angreifern und qualitativ besseren Angriffen.
- Angreifer und Verteidiger profitieren von allgemeinen Produktivitätssteigerungen durch den Einsatz von LLMs, z. B. für die Aufklärung und Open-Source-Intelligenz (z. B. durch das Crawlen und Analysieren von Websites und sozialen Medien) oder die Codegenerierung (z. B. Programmierassistenten).
- Es gibt Proof of Concepts (PoC) und Projekte, die KI für die autonome Generierung und Mutation von Malware einsetzen. Öffentlich verfügbaren Modelle sind bisher allerdings noch nicht „produktionsreif“.
- Tools, die Angriffe oder Exfiltrationspfade optimieren, werden derzeit auf einzelne Netzwerke trainiert. Sie sind bis heute nicht verallgemeinerbar und existieren (öffentlich verfügbar) nur als PoCs.
- Agenten, die eigenständig beliebige Infrastrukturen kompromittieren, sind noch nicht verfügbar und werden es wahrscheinlich auch in naher Zukunft nicht sein. LLM-basierte Agenten, die Teile eines Angriffs automatisieren, werden jedoch in naher Zukunft verfügbar sein.
- KI kann bei der automatischen Erkennung von Sicherheitslücken eingesetzt werden. Dies ist ein aktives Forschungsgebiet und es sind mehrere Open-Source-Tools sowie kommerzielle Produkte verfügbar. In Zukunft wird es für Open-Source-Projekte von entscheidender Bedeutung sein, diese Art von Tools proaktiv zu nutzen, bevor böswillige Akteure dies tun.

Unsere wichtigsten Ergebnisse stimmen insgesamt mit einem kürzlich veröffentlichten Bericht des britischen National Cyber Security Centre überein (3).

1.2 Empfehlungen

Angesichts der sich verändernden Bedrohungslandschaft ist es wichtig, der Cybersicherheit höchste Priorität einzuräumen. Es wird entscheidend sein, die Geschwindigkeit und den Umfang der Abwehrmaßnahmen zu erhöhen, insbesondere, aber nicht ausschließlich, durch

- Verbesserung des Patchmanagements,
- Aufbau einer resilienten IT-Infrastruktur,
- Verbesserung der Angriffserkennung,
- Verstärkung der Social-Engineering-Prävention (z. B. Sensibilisierungsschulung, Multi-Faktor-Authentifizierung, Zero-Trust-Architektur) sowie
- Nutzung der allgemeinen Vorteile der KI für Verteidigungsmaßnahmen (z. B. Erkennung von Bedrohungen und Schwachstellen).

Da KI häufig klassische Angriffe verstärkt, fallen auch diese Maßnahmen weitgehend in den Bereich der klassischen IT-Sicherheit.

1.3 Limitierungen

Sowohl Cybersicherheit als auch Künstliche Intelligenz unterliegen einem ständigen Wandel, weshalb es wichtig ist, Veränderungen und neue Entwicklungen bezüglich der Bedrohungslandschaft weiterhin zu beobachten. Zwar gibt es noch keine autonomen Hackeragenten, aber es ist schwierig, Programme fähiger Akteure zuverlässig zu bewerten oder technische Durchbrüche vorherzusagen. Bei unserer derzeitigen Bewertung der Auswirkungen von KI auf die Cyberbedrohungslandschaft gehen wir davon aus, dass es in naher Zukunft keine bedeutenden Durchbrüche bei der Entwicklung von KI, insbesondere von LLMs, geben wird. Sollte sich diese Annahme nicht bewahrheiten, müssen die Auswirkungen neu bewertet werden.

2 Auswirkungen großer Sprachmodelle

Mit der Veröffentlichung von ChatGPT im November 2022 hat ein Wettbewerb um die Führung auf dem Markt für Chatbots begonnen. Es werden ständig neue Produkte und Sprachmodelle veröffentlicht, die erhebliche Leistungssprünge versprechen. Infolgedessen ist es nun für praktisch jedermann möglich, auf leistungsstarke Sprachmodelle zuzugreifen, welche Ergebnisse von bisher unerreichter Qualität liefern. Die Leistung und Verfügbarkeit dieser Sprachmodelle hat inzwischen verschiedene Branchen beeinflusst und wird auch den Cybersicherheitssektor nachhaltig beeinflussen.

Für cybersicherheitsrelevante Anwendungen können LLMs hilfreich sein, indem sie direkt über eine Web oder mobile App (in der Regel als Chatbot) aufgerufen werden. Es ist auch möglich, den API-Zugang zu nutzen, um LLMs in bestehende Tools (z. B. Reverse-Engineering-Tools oder Penetrationtestframeworks) zu integrieren oder neue Anwendungen zu entwickeln. Methoden und Anwendungen im Bereich der Cybersicherheit weisen Dual-Use-Eigenschaften auf, wobei ihre ethische Anwendung von den Absichten des Nutzers abhängt. Dieser Grundsatz gilt auch für die Verwendung von LLMs im Bereich der Cybersicherheit. Ob die Verwendung gutartig oder bösartig ist, hängt von der Absicht des Benutzers ab. Leider ist es für Benutzer mit schlechten Absichten leicht, die Fähigkeiten von LLMs zu missbrauchen. Neben allgemeinen Produktivitätsgewinnen für böswillige Akteure sehen wir derzeit eine böswillige Nutzung vor allem in zwei Bereichen: Social Engineering und Generierung von bösartigem Code.

Der einfache Zugang zu hochwertigen LLMs ermöglicht es, selbst mit geringen oder gar keinen Fremdsprachenkenntnissen automatisch überzeugende Phishing-Nachrichten von hoher Qualität zu erstellen. Die Anweisungen können mit zusätzlichem Kontext ergänzt werden, um die Nachrichten zu personalisieren oder einen bestimmten Schreibstil zu verwenden, was zu überzeugenden Nachrichten führt. Herkömmliche Methoden zur Erkennung betrügerischer Nachrichten, wie z. B. die Prüfung auf Rechtschreibfehler und unkonventionellen Sprachgebrauch, reichen daher nicht mehr aus. LLMs können auch eingesetzt werden, um die Erfolgsquote von Phishingangriffen weiter zu erhöhen, indem beispielsweise plausible Domainnamen und URLs generiert werden. Die Kombination eines LLM mit anderen generativen KI-Techniken, wie z. B. Deepfakes für Bild- und Audioinhalte, ermöglicht es böswilligen Akteuren, Social-Engineering-Angriffe von noch nie dagewesener Qualität durchzuführen.

Es ist in der Regel schwierig, einen bestimmten Angriff mit der Verwendung eines LLM in Verbindung zu bringen, da dies eng mit dem Problem der Erkennung von KI-generierten Inhalten im Allgemeinen verbunden ist. Berichte in den Medien, von Sicherheitsberatungsunternehmen und Regierungsbehörden sowie Untersuchungen auf Marktplätzen liefern jedoch eindeutige Beweise für die Verwendung von LLMs durch böswillige Akteure, einschließlich sog. Advanced Persistent Threats (4).

Die Fähigkeit von LLM, schädlichen Code zu generieren, verändert auch die Cyberbedrohungslandschaft, da sie insbesondere die Einstiegshürden für Personen, die bösartige Aktivitäten durchführen wollen, senkt, indem sie es auch Personen mit begrenzten technischen Kenntnissen ermöglicht, anspruchsvollen schädlichen Code zu produzieren (5). Auch bereits fähige Akteure profitieren von Produktivitätssteigerungen. Die Anbieter von Chatbots oder offenen LLMs treffen in der Regel Vorkehrungen um sicherzustellen, dass ihre Produkte nicht missbraucht werden können. Es werden Filtersysteme eingesetzt, um unerwünschte Ausgaben zu verhindern. Diese Systeme sind in der Regel nützlich, um einfache Aufforderungen mit böswilligen Absichten, wie z. B. „Erstelle mir Code für Ransomware“, abzufangen. Es ist jedoch oft nur wenig Aufwand und Wissen über die Domäne erforderlich, um diese Systeme zu umgehen. Da die Filterung immer ein Kompromiss zwischen der Verhinderung unerwünschter Ausgaben und der gleichzeitigen Bereitstellung eines Systems mit hohem Nutzen ist, ist es fraglich, inwieweit solche Filterungen Missbrauch wirksam verhindern können.

Die Nutzung eines Chatbots, der von einem Onlinedienst bereitgestellt wird, der ein System zur Verhinderung unethischer Ausgaben einsetzt, ist nicht die einzige Möglichkeit, auf ein LLM zuzugreifen. Weitere Möglichkeiten sind die Verwendung von „Jailbreaks“ (Nutzereingaben, die bestehende Filter und Anweisungen außer Kraft setzen), die Verwendung von Diensten, die die Ausgabe nicht rigoros filtern oder

die Verwendung „unzensierter“ öffentlicher Modelle. Zusätzliche Schritte zur Umgehung der Filter, wie sie oben erwähnt wurden, sind hier nicht erforderlich.

3 KI-basierte Schadcodegenerierung

Malware ist der Sammelbegriff für schädliche Software, wie z.B. Ransomware, Würmer oder Trojaner. Oft ist es das Ziel von Angreifern, Malware auf einem Zielcomputer zu platzieren, sei es über Exploits oder durch Social Engineering. Maßnahmen wie Virenscanner bekämpfen solche Software, indem sie Schadcode erkennen und das Ausführen unterbinden. Dies führt zu einem Wettrüsten zwischen den Angreifern, die neue Malware entwickeln, und den Verteidigern, die ihre Verteidigungsmaßnahmen anpassen, um neue Bedrohungen abzuwehren.

Daher ist es interessant zu untersuchen, wie KI die Erstellung und Verwendung von Malware beeinflusst. Im Rahmen unserer Recherchen haben wir mehrere Möglichkeiten gefunden, wie KI in diesem Bereich eingesetzt wird. Die Modelle reichen von LLMs über GANs (Generative Adversial Networks) bis hin zu Reinforcement-Learning-Systemen und sie werden für verschiedene Zwecke eingesetzt.

Erstens ermöglicht KI Akteuren mit geringen oder gar keinen technischen Kenntnissen, Malware einfacher zu erstellen. Sie brauchen kein tiefes Verständnis für die Programmierung oder die Funktionsweise von Malware und können ihre Anfrage in natürlicher Sprache stellen.

Weiterhin besteht die Sorge, dass KI dazu verwendet werden könnte, eigenständig Malware zu schreiben. Dies geht einen Schritt weiter als die bloße Unterstützung menschlicher Akteure. LLMs können bereits einfache Malware schreiben, aber wir haben keine KI gefunden, die eigenständig in der Lage ist, fortgeschrittene, bisher unbekannte Malware zu schreiben (z. B. mit komplizierten Verschleierungsmethoden oder Zero-Day-Exploits). Die erforderlichen Trainingsdaten über Malware und Schwachstellen wären zudem nur sehr schwer und teuer zu erstellen.

Als nächstes kann KI dabei helfen, Malware zu modifizieren. Dies ist realistischer als Malware von Grund auf neu zu erstellen und es existieren bereits mehrere Forschungsarbeiten über die Modifizierung von Malware durch KI. Dies geschieht meist in einem Featureraum, nicht auf der eigentlichen Codeebene, und mit dem Ziel, eine Entdeckung zu vermeiden. Allerdings findet dies bislang in einem eher akademischen Umfeld statt und wir haben keine Hinweise darauf gefunden, dass diese Modelle bereits eingesetzt werden. Außerdem gibt es kein ausgefeiltes Tool, sondern nur PoCs und Forschungsprojekte. Dieser Ansatz eignet sich nur für hochqualifizierte Akteure, sowohl im Bereich Malware als auch KI, zudem ist für das Training solcher Tools ist eine gute Datenbasis erforderlich.

Schließlich möchten wir noch KI als Teil der Malware erwähnen. Hier erstellt die KI nicht die Malware an sich, sondern ist in die Funktionalität der Malware integriert. Oft ist das Ziel, die Malware zu verschleiern und damit zu verhindern, dass sie entdeckt wird. Um einer Entdeckung zu entgehen, existieren so genannte polymorphe Engines, die den Code der Malware verändert, während ihre Funktionalität erhalten bleibt. Eine Anwendung von KI in diesem Bereich ist zumindest denkbar. Dabei würde die Manipulation des Codes durch ein KI-Modell bestimmt werden. Zum jetzigen Zeitpunkt gibt es keine Hinweise darauf, dass ein solches Modell im Einsatz ist, obwohl es viele Warnungen vor einer solchen theoretischen Möglichkeit gibt. Eine andere Möglichkeit wäre, ein KI-Modell so zu trainieren, dass es das Benutzerverhalten nachahmt, so dass die Aktionen der Malware weniger auffällig sind.

4 KI-basierte Angriffe

Das interessanteste Tool für cyberkriminelle Aktivitäten wäre eine KI, die das Ziel als Eingabe erhält (sei es ein IP-Bereich oder ein Name) und alle Schritte eines Cyberangriffs völlig selbstständig durchführt. Die Strategie- und Abstraktionsfähigkeiten der neusten KI-Technologien machen sie zu erstklassigen Kandidaten für die Entwicklung eines solchen Tools. Aus der Sicht eines Penetrationstesters wäre dies ein nützliches Werkzeug, um Systeme zu härten und den Zeitaufwand für die Durchführung von Penetrationstests zu verringern. Hierbei handelt es sich um ein aktuelles Forschungsfeld und es werden Anstrengungen unternommen, ein solches Tool zu entwickeln.

In diesem Bereich sind Reinforcement-Learning-Systeme ein gängiger Ansatz, da diese in der Lage sind, mit einer Umgebung zu interagieren, daraus zu lernen und langfristige Strategien zu entwickeln. Kürzlich wurden auch LLMs als Lösung für dieses Problem vorgeschlagen. In unserer Forschung haben wir kein Werkzeug gefunden, das diese Aufgabe vollständig lösen kann. Es gibt jedoch einige Tools, die Teile des Prozesses automatisieren. Meistens handelt es sich bei diesen Tools um akademische Projekte oder PoCs, die nicht besonders benutzerfreundlich oder ausgefeilt sind. Oft ist der Anwendungsbereich dieser Tools entweder sehr groß oder sehr klein. Im großen Maßstab betrachten beispielsweise Tools zur Planung von Angriffswegen eine abstrakte Version eines Zielnetzes und planen einen optimalen Angriffsweg. Ein aktiver Angriff findet dabei nicht statt. Ähnlich verhält es sich mit Modellen, die optimale Exfiltrationspfade für Systeme finden.

Auf der anderen Seite gibt es Tools, die explizit auf ein einzelnes, spezifisches Netzwerk trainiert sind und versuchen, dort einen erfolgreichen Angriff zu starten. Dies erfordert Kenntnisse über das Zielnetzwerk, sowie eine Trainingsphase, die kaum unbemerkt bleiben wird. Außerdem lässt sich ein trainierter Agent nicht ohne weiteres auf andere Netze verallgemeinern. Die Umgebungen verschiedener Systeme und Netze unterscheiden sich stark in Größe und verfügbaren Aktionen. Das macht eine Verallgemeinerung sehr schwierig. Außerdem ist eine sehr große Trainingsdatenmenge erforderlich, um die Fülle der Optionen abzudecken. Diese Probleme machen den Schritt von einem Konzeptnachweis zu einer realen, allgemeinen Anwendung zu einem schwierigen, wahrscheinlich derzeit ungelösten Problem. LLMs könnten ein Ansatz sein, um die Verallgemeinerbarkeit zu verbessern.

Es gibt mehrere Tools, die das Pentesting durch KI-Assistenten unterstützen. Nach Tests haben wir festgestellt, dass diese Tools vor allem als Unterstützung für Personen dienen, die versuchen, einen Angriff zu starten und dadurch die Einstiegsschwelle senken.

Ein anderer Ansatz für LLMs ist, ähnlich wie bei den oben genannten Tools, bestimmte Teile der Angriffskette zu automatisieren. Hier ist vor allem die Aufklärungsphase zu nennen, aber auch andere Schritte wie die Analyse von Serverantworten. Die Anwendung von KI als vollautomatisches Angriffswerkzeug ist ein Bereich, der intensiv erforscht wird. Wir erwarten weitere Projekte und Tools in diesem Bereich, insbesondere solche, die sich auf die Verwendung von LLMs und generativer KI konzentrieren.

5 Weitere Schnittstellen zwischen KI und Cybersicherheit

In diesem Abschnitt geben wir einen Überblick über andere Bereiche, in denen sich KI und Cybersicherheitsanwendungen überschneiden. Der sichtbarste davon die Integration von LLMs in diverse Tools, wie sie auch in anderen Bereichen stattfindet.

LLMs wurden bereits in IDEs (integrierte Entwicklungsumgebungen) integriert und es gibt Plugins für Reverse-Engineering- oder Penetrationstest-Tools. Diese Plugins rufen in der Regel die API eines LLM-Anbieters mit einer vorgefertigten Anweisung und Inhalten aus der jeweiligen Anwendung auf, das Ergebnis wird dann innerhalb der Anwendung angezeigt. Der Nutzen im Vergleich zur direkten Verwendung des LLM, beispielsweise im Browser (zusammen mit dem Kopieren und Einfügen des jeweiligen Inhalts), ist derzeit noch begrenzt.

KI wird auch bei der automatischen Erkennung von Sicherheitslücken eingesetzt. Aufgrund ihrer Vorteile für die Softwareentwicklung ist dies ein aktives Forschungsgebiet und es sind mehrere Open-Source-Tools sowie kommerzielle Produkte verfügbar. Die Analyse von Open-Source-Anwendungen mit diesen Tools ist sehr einfach durchführbar. Daher ist es für Open-Source-Projekte von entscheidender Bedeutung, diese Art von Tools proaktiv zu nutzen, bevor böswillige Akteure dies tun. Obwohl für die Analyse in der Regel der Quellcode benötigt wird, ist es in Kombination mit Reverse-Engineering-Tools bis zu einem gewissen Grad möglich, Methoden zur Erkennung von Schwachstellen bei Closed-Source-Anwendungen einzusetzen. Es gibt Projekte, die diesen Prozess mithilfe eines LLM automatisieren. Die Ergebnisse sind jedoch je nach Komplexität des Codes und der Verschleierungstechniken sehr unterschiedlich.

Captchas werden allgegenwärtig eingesetzt, um zwischen automatisierten Bots und echten menschlichen Nutzern durch verschiedene Aufgaben wie verzerrte Text- oder Bilderkennungsaufgaben zu unterscheiden und so bösartige Aktivitäten wie Spamming, Brute-Force-Angriffe und Data Scraping zu verhindern, indem sie menschenähnliche Antworten verlangen. Methoden zur Umgehung von Captchas gibt es jedoch schon seit ihrer Erfindung. Heutzutage wird auch KI für diesen Zweck eingesetzt und die Auswahl an solchen Tools und Online-Diensten, die gute Ergebnisse liefern, ist groß.

KI wird auch zum Erraten von Passwörtern eingesetzt. Ausgehend von der Annahme, dass bestimmte Arten von Passwörtern mit größerer Wahrscheinlichkeit von Menschen verwendet werden, werden die Regeln für Passwortkandidaten aus Daten gelernt, im Gegensatz zu bestehenden Tools zum Erraten von Passwörtern, bei denen diese Regeln von Hand ausgearbeitet werden. Durch diverse Leaks existieren hierfür viele Trainingsdaten.

Da jeder bestrebt ist, KI in seine Prozesse zu integrieren, wird die Gefahr von eingebetteter Malware in der KI oder in Daten immer größer. Es gibt bereits Fälle, in denen Malware in den Parametern neuronaler Netze verschlüsselt ist, wobei die Nutzbarkeit des Modells kaum verändert ist. Schädlicher Code kann auch in trainierten Modellen versteckt sein, die häufig auf bestimmten Plattformen verbreitet werden (6). Außerdem können LLMs und das dazugehörige Ökosystem dazu missbraucht werden, bösartige Software an die Nutzer zu verteilen.

Auf der Hardwareseite der Angriffe sind Seitenkanalangriffe ein bekannter Angriffsvektor. Sie erfordern jedoch ein hohes Maß an technischem Können und Knowhow. Es gibt PoCs über KI-Modelle, die bei Seitenkanalangriffen helfen und diese möglicherweise für weniger erfahrene Angreifer leichter zugänglich machen.

Seriöse Softwareanbieter haben damit begonnen, LLMs für den Kundensupport zu nutzen. In ähnlicher Weise könnten auch böswillige Akteure LLMs nutzen, um Malware-as-a-Service-Nutzern Unterstützung zu bieten. Technisch weniger versierte Opfer von Ransomwareangriffen sind oft mit der Beschaffung der für die Zahlung des Lösegelds erforderlichen Kryptowährung überfordert. Cyberkriminelle bieten bereits Unterstützung bei der Beschaffung und Zahlung an. Dieser Prozess kann durch den Einsatz von LLMs automatisiert und in verschiedenen Sprachen angeboten werden, um die Erfolgsquote zu erhöhen.

Literaturverzeichnis

1. **Fang, Richard, et al.** *LLM Agents can Autonomously Hack Websites.* 2024.

2. **Oberhaus, Daniel.** *Prepare for AI Hackers.* https://www.harvardmagazine.com/2023/02/right-now-ai-hacking : Harvard Magazine, 2023.

3. **NCSC.** The near-term impact of AI on the cyber threat. *NCSC.* [Online] 24. 01 2024. https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

4. **Microsoft Threat Intelligence.** Staying ahead of threat actors in the age of AI. *Microsoft.* [Online] Microsoft, 14. 02 2024. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

5. **Pa Pa, Yin Minn, et al.** *An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware.* 2023.

6. **Cohen, David.** Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor. *JFrog.* [Online] 27. 02 2024. https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/.