Erster Artikel auf t-online vom 22.07.2026:

"Beispielloser Cyber-Vorfall": KI bricht aus – und attackiert eigenständig andere Firma:

Experten warnen schon länger vor Künstlicher Intelligenz, die eigenständig Cyberangriffe verüben kann. Jetzt kam es tatsächlich erstmals dazu.

Ein außergewöhnlicher Zwischenfall beim Testlauf neuer KI-Modelle demonstriert das Risiko potenzieller Cyberattacken mit Künstlicher Intelligenz. Der US-Konzern Open AI hat Verantwortung für einen Hackerangriff auf die KI-Plattform Hugging Face übernommen. Das Unternehmen sprach von einem "beispiellosen Cyber-Vorfall" und kündigte an, die Sicherheitsvorkehrungen zu verstärken.

Einige der fortschrittlichsten KI-Modelle von OpenAI seien bei einem Sicherheitstest aus einer eigentlich abgeschotteten Umgebung ausgebrochen, teilte OpenAI am Dienstag weiter mit. Die Software drang demnach eigenständig in Computersysteme der anderen Firma ein und verhielt sich dabei wie ein gewiefter Hacker.

Hugging Face, eine Plattform für KI-Anwendungen, hatte bereits vergangene Woche von einem Cyberangriff mithilfe von Künstlicher Intelligenz berichtet. Nun stellt sich heraus, dass dieser auf einen internen Versuch bei OpenAI zurückging, den die Software ungeplant zu weit trieb.

Alles nur um zu schummeln

OpenAI wollte laut dem Blogeintrag beim neuen Modell GPT-5.6 Sol und einer noch unveröffentlichten künftigen Version die Fähigkeit ausloten, Sicherheitslücken für Cyberattacken auszunutzen. Dafür sollte die Software Aufgaben eines Standard-Tests mit dem Namen ExploitGym lösen, der dafür oft in der Branche verwendet wird. Doch die KI-Modelle gingen viel weiter als erwartet, um die Aufgabe zu erfüllen.

Zunächst brach die Software aus der Testumgebung aus, indem sie sich auf eigene Faust Zugang zum offenen Internet verschaffte – und zwar über eine bisher unentdeckt gebliebene Schwachstelle. Während sie eigenständig im Netz unterwegs waren, kamen die Modelle zu dem Schluss, dass es bei Hugging Face nützliche Daten und Lösungen für die ExploitGym-Aufgaben geben könnte. Deshalb habe die Software sich Zugang zu "geheimen Informationen" auf der Plattform

verschafft, mit denen sie beim ExploitGym-Test schummeln könnte, erläuterte OpenAI.
Auch dafür setzte die OpenAI-Software unter anderem auf zuvor unentdeckte Sicherheitslücken – aber auch auf gestohlene Zugangsdaten. Hugging Face zufolge führte der KI-Angreifer Tausende Schritte aus und wechselte den Standort der digitalen Steuerzentrale der Attacke.

Angst vor KI-Cyberangriffen

Experten warnen schon lange vor Cyberattacken mit KI-Software. In den vergangenen Monaten heizte vorwiegend ein neues Modell des OpenAI-Rivalen Anthropic die Debatte an. Die Software mit dem Namen Mythos entdeckte über Jahrzehnte unentdeckt gebliebene Sicherheitslücken bei viel genutzten Programmen und Online-Diensten. Die Schwachstellen konnten in der Folge gestopft werden. Doch zugleich führte das vor Augen, dass solche KI in den falschen Händen als verheerende Cyberwaffe dienen kann.
Anthropic macht die vollen Möglichkeiten von Mythos nur für ausgewählte Behörden und Unternehmen zugänglich. Für andere gibt es eine abgespeckte Version mit dem Namen Fable, der unter anderem die Cybersicherheits-Fähigkeiten fehlen. Nach Warnungen, dass die KI-Modelle von den Zügeln befreit werden könnten, musste Anthropic auf Geheiß der US-Regierung den Zugang zu Mythos und Fable zeitweise sperren. Die Blockade wurde inzwischen aufgehoben.

<u>Zweiter Artikel auf t-online von Marcel Horzenek vom 22.07.2026:</u>

Ausbruch aus Testumgebung: KI hackt fremde Firma – wie ernst ist der Vorfall wirklich?

KI-Modelle von OpenAI haben eine fremde Firma gehackt. Was nach einem Albtraum-Szenario klingt, ist es bei näherem Hinsehen gar nicht – denn die Bedrohung ist nicht neu.

Open AI hat öffentlich Verantwortung für den Cyberangriff auf die KI-Plattform Hugging Face übernommen. Hinter der Attacke steckten laut dem Unternehmen gleich mehrere eigene Modelle – darunter das aktuelle GPT-5.6 Sol und eine noch unveröffentlichte, leistungsstärkere Version.

Der Angriff war kein Werk von Kriminellen, sondern das Ergebnis eines internen Tests. OpenAI wollte messen, wie weit seine Modelle bei Cyberangriffen kommen, und ließ sie dafür die Aufgaben eines Tests namens ExploitGym lösen. Die Modelle liefen dabei in einer abgeschotteten Umgebung ohne echten Internetzugang.
Wie sich herausstellte, war diese Testumgebung nicht ausreichend gesichert, denn die Modelle gingen weiter, als es die Forscher von OpenAI vorgesehen hatten: Über eine bis dahin unbekannte Sicherheitslücke verschafften sie sich Zugang zum offenen Internet. Anschließend schlussfolgerten sie selbstständig, dass die Lösungen für den Test wohl bei Hugging Face lägen – und drangen mit gestohlenen Zugangsdaten und weiteren Schwachstellen in deren Systeme ein. Beide Sicherheitsteams bemerkten die Aktivität, Hugging Face stoppte sie.

Warum die KI überhaupt zum Angreifer wurde

Die Schutzmechanismen, die solches Verhalten im Normalbetrieb blockieren, waren für diesen Test bewusst abgeschaltet. Es handelte sich um eine gezielte Grenzerprobung – nicht um ein Modell, das im Alltag plötzlich außer Kontrolle gerät. Es ist dennoch der erste bekannte Fall, in dem ein solcher Fähigkeitstest in einen echten Angriff auf ein fremdes Unternehmen umschlug.
Entscheidend ist vor allem, worum es den Modellen ging. Sie wollten keinen Schaden anrichten, sondern ihre Testaufgabe lösen – notfalls über eine drastische Abkürzung. Für den KI-Experten Fabian Westerheide steckt darin die eigentliche Lehre.
"Die KI wollte nicht im menschlichen Sinne schummeln. Sie hatte ein Ziel und hat einen überraschend effektiven Weg gefunden, dieses Ziel zu erreichen." Anders als Menschen verstünden solche Systeme unausgesprochene Regeln nicht automatisch. Wer einer KI sagt, sie solle einen Test lösen, meine selbstverständlich: auf dem vorgesehenen Weg. Das Modell aber bekomme zunächst nur ein Ziel.
Je leistungsfähiger die Systeme würden, desto kreativer könnten die Wege werden, die sie einschlagen. Entscheidend sei daher, welche Ziele und Handlungsspielräume man ihnen gebe. Denn, so Westerheide: "Sie optimieren nicht automatisch auf das, was wir meinen. Sie optimieren auf das, was wir ihnen tatsächlich vorgeben."

Was der Vorfall für Unternehmen und Verbraucher bedeutet

Für den Cybersecurity-Experten Rüdiger Trost ist der Vorfall weniger ein Bruch als eine Fortsetzung bekannter Entwicklungen. Angriffe dieser Art gebe es bereits seit Längerem – bislang jedoch mit einem Menschen, der die eingesetzten KI-Werkzeuge selbst steuerte.
Ein Netzwerkangriff sei kein einzelnes Ereignis, sondern vollziehe sich in vielen kleinen Schritten, erklärt Trost. KI sei dabei bisher vor allem in den frühen Phasen zum Einsatz gekommen, Informationen über das Ziel zu sammeln, Daten und Social-Media-Profile auszuwerten oder Phishing-Mails zu personalisieren.
Mit den aktuellen Modellen ließen sich nun auch die komplexeren, späteren Schritte automatisiert ausführen, also Datendiebstahl, Verschlüsselung (Ransomware), Sabotage etc. – also der eigentliche Zweck des Angriffs.
Für den Alltag von Anwendern und IT-Verantwortlichen ändert sich laut Trost zunächst nichts. Sie sollten bekannte Sicherheitslücken weiterhin so schnell wie möglich schließen und gängige Sicherheitsstandards befolgen. "Die Angreifer bleiben dieselben, könnten künftig jedoch über bessere und schnellere Werkzeuge verfügen." Die Bedrohung ist also keine grundsätzlich neue – aber möglicherweise eine schnellere.

Wie viel Kontrolle wir abgeben (sollten)

Bleibt die Frage nach der Zukunft. KI-Systeme sollen künftig immer mehr Aufgaben eigenständig übernehmen. Kontrolle abzugeben sei dabei kein Fehler, sondern der eigentliche Sinn autonomer Systeme, sagt Westerheide. Entscheidend sei, wo und unter welchen Bedingungen.
Bei Aufgaben mit begrenzten, umkehrbaren Folgen würde er einer KI schon heute viel überlassen. Je größer der mögliche Schaden – etwa bei Geld oder kritischer Infrastruktur –, desto strenger müssten die Grenzen sein. Vertrauen dürfe dabei aber kein Gefühl sein, sondern müsse technisch entstehen: durch klare Grenzen, begrenzte Zugriffsrechte, Überwachung und die Möglichkeit, ein System jederzeit zu stoppen.
Der Vorfall zeige deshalb nicht, dass man auf autonome KI verzichten sollte, so Westerheide. Sondern "dass ihre Fähigkeiten möglicherweise schneller wachsen als unsere Fähigkeit, sie zuverlässig einzugrenzen."

Denn hier liegt das eigentliche Problem: Die klaren Grenzen, Zugriffsrechte und Kontrollmechanismen, die Westerheide fordert, wirken nur, wenn jemand sie auch einbaut.

Kriminelle nutzen keine Schutzmechanismen

Es sind dieselben Schutzmechanismen, die OpenAI für seinen Test bewusst abgeschaltet hatte – und die seriöse Anbieter ihren Modellen im Normalfall mitgeben. Wer dagegen ein frei verfügbares oder manipuliertes Modell ohne solche Sperren nutzt, hat diese Bremse gar nicht erst.

Hugging Face selbst erklärte, zunächst nicht mit Sicherheit gewusst zu haben, welches Modell hinter dem entdeckten Angriff steckte – und verwies darauf, dass ein Angreifer mit einem unbeschränkten Modell an keinerlei Nutzungsregeln gebunden ist.

Der Vorfall bei OpenAI führt damit vor allem eines vor Augen: wozu die Technik inzwischen grundsätzlich fähig ist. Die Frage ist längst nicht mehr, ob KI solche Angriffe ausführen kann – sondern in wessen Händen sie das tut und mit welchen Grenzen.