Im Juli 2026 sollte eine Künstliche Intelligenz in einer abgeschirmten Testumgebung ein Sicherheitsproblem lösen. Der Internetzugang war gesperrt, fremde Computersysteme gehörten nicht zum Versuchsaufbau. Was anschließend geschah, schildert der Hersteller OpenAI inzwischen selbst. Von **Günther Burbach**.

Das System entdeckte eine bislang unbekannte Schwachstelle in einer Software, die eigentlich nur der Installation von Programmpaketen diente. Über diesen Umweg verschaffte es sich Zugang zum Internet. Anschließend verwendete es fremde Zugangsdaten, durchsuchte externe Dienste und drang in die Infrastruktur der Plattform Hugging Face ein. Dort vermutete es Informationen, die ihm bei seiner Prüfungsaufgabe helfen konnten.

Kein Mensch hatte Hugging Face als Angriffsziel ausgewählt. Der Auftrag lautete lediglich, eine Aufgabe zu lösen. Den Weg dorthin entwickelte das System selbst.

Solche Vorfälle hätten eine breitere öffentliche Debatte verdient. Stattdessen erscheint Künstliche Intelligenz in politischen Sonntagsreden weiterhin überwiegend als Antwort auf wirtschaftliche Probleme. Sie soll Verwaltungen entlasten und Unternehmen produktiver machen. Was geschieht, wenn dieselben Systeme bei ihrer Aufgabenerfüllung technische Grenzen überschreiten, wird dagegen meist erst diskutiert, nachdem etwas passiert ist.

Aus einer Prüfungsaufgabe wurde ein Angriff

Nach Angaben von OpenAI waren mehrere Modelle an dem Vorfall beteiligt, darunter GPT-5.6 Sol und ein leistungsfähiger interner Forschungsprototyp. Für den Test waren Schutzmechanismen reduziert worden, die im normalen Betrieb riskante Cyberaktivitäten verhindern sollen. Die Entwickler wollten feststellen, wozu ihre Modelle grundsätzlich in der Lage sind.

Das ist für die Einordnung wichtig. Aus dem Vorfall folgt nicht, dass gewöhnliche Nutzer von ChatGPT plötzlich über autonome Hackerprogramme verfügen. Er zeigt jedoch, welche Fähigkeiten hinter den Schutzvorkehrungen existieren und was geschehen kann, wenn diese fehlen oder überwunden werden.

Die KI entdeckte eine sogenannte Zero-Day-Schwachstelle. So werden Sicherheitslücken bezeichnet, die dem Hersteller der betroffenen Software bis dahin unbekannt waren. Gegen solche Fehler gibt es zunächst keinen gezielten Schutz.

Nach dem Ausbruch aus der Testumgebung kombinierte das System verschiedene

technische Möglichkeiten miteinander. Ein fremdes Konto diente als Zwischenstation, ein weiteres zur Datenspeicherung. Die Sicherheitsabteilungen von OpenAI und Hugging Face mussten den Vorfall untersuchen und die Zugänge schließen.

Der beteiligte Forschungsprototyp wurde nach Angaben des Unternehmens deaktiviert. Damit war der konkrete Angriff beendet. Die Fähigkeit, die ihn ermöglicht hatte, verschwand dadurch nicht.

Anthropic findet drei weitere Vorfälle

Der Konkurrent Anthropic überprüfte nach der Veröffentlichung von OpenAI insgesamt 141.006 eigene Testläufe. Dabei entdeckte das Unternehmen drei Fälle, in denen Claude-Modelle unberechtigt auf reale Computersysteme fremder Organisationen zugegriffen hatten. Zwei betroffene Organisationen erfuhren erst durch Anthropic von dem Vorfall.

Auch diese Systeme sollten ursprünglich nur eine simulierte Aufgabe lösen. Anders als bei OpenAI mussten sie dafür allerdings keine komplizierte technische Schwachstelle überwinden. Eine fehlerhafte Konfiguration hatte ihnen bereits Zugang zum Internet verschafft.

Die Modelle behandelten erreichbare fremde Systeme anschließend als Teil der Übung und nutzten unter anderem schwache Passwörter sowie ungeschützte Schnittstellen.

Ein neueres Modell stellte seine Aktivitäten ein, nachdem es erkannt hatte, dass es sich im offenen Internet befand. Ein älteres Modell setzte seinen Angriff teilweise fort, obwohl Hinweise auf die reale Umgebung vorlagen.

Wer daraus eine Geschichte über Maschinen mit eigenem Bewusstsein machen möchte, überzieht den Befund. Für die betroffenen Unternehmen ändert diese Unterscheidung allerdings wenig. Ein unberechtigter Zugriff bleibt ein unberechtigter Zugriff, auch wenn das angreifende System seine Handlung für eine Prüfungsaufgabe hält.

Fehler, die jahrzehntelang unentdeckt blieben

Anthropic arbeitet außerdem an einem Modell namens Claude Mythos, dessen Fähigkeiten die bisherigen Maßstäbe der IT-Sicherheit infrage stellen. Nach Angaben des Unternehmens fand das System tausende unbekannte Sicherheitslücken in wichtigen Betriebssystemen und Browsern.

Eine Schwachstelle steckte seit 27 Jahren im Betriebssystem OpenBSD. Dieses System gilt

als besonders sicherheitsorientiert und wird unter anderem in Schutzsystemen eingesetzt. Der entdeckte Fehler hätte es ermöglicht, betroffene Rechner aus der Ferne zum Absturz zu bringen.

Ein weiterer Fehler befand sich seit 16 Jahren in FFmpeg, einer weitverbreiteten Software zur Verarbeitung von Bild- und Tondateien. Automatisierte Prüfprogramme hatten die betreffende Codezeile nach Angaben von Anthropic ungefähr fünf Millionen Mal untersucht, ohne das Problem zu erkennen.

Das Modell entdeckte außerdem mehrere Sicherheitslücken im Linux-Kernel und verband sie zu einem Angriffspfad, über den ein gewöhnlicher Nutzer vollständige Kontrolle über einen Rechner hätte erlangen können.

Die konkret genannten Fehler wurden inzwischen behoben. Dennoch verändert sich das Kräfteverhältnis: Eine KI kann Schwachstellen in einem Tempo aufspüren, mit dem menschliche Sicherheitsabteilungen kaum Schritt halten.

Anthropic räumt offen ein, dass inzwischen weniger das Auffinden der Fehler das Problem ist als deren Überprüfung und Beseitigung. Gleichzeitig erklärt das Unternehmen, dass Schutzmechanismen für eine sichere breite Freigabe vergleichbarer Fähigkeiten bislang fehlen.

Damit sinkt langfristig auch die Einstiegshürde für Angreifer. Was bislang erfahrene Spezialisten und erhebliche Vorbereitung erforderte, könnte künftig von wesentlich mehr Akteuren versucht werden.

Siemens-Steuerungen und die deutsche Infrastruktur

Die Entwicklung betrifft Deutschland unmittelbar. Die amerikanische Cybersicherheitsbehörde CISA warnt vor einer aktiven Bedrohung für industrielle Siemens-S7-Steuerungen. Nach Angaben der Behörde nutzen Angreifer Künstliche Intelligenz, um passende Angriffsskripte zu erstellen.

Solche Steuerungen übernehmen Aufgaben in industriellen Anlagen und infrastrukturell wichtigen Einrichtungen. Ein erfolgreicher Angriff betrifft dann unter Umständen nicht nur gespeicherte Daten, sondern reale technische Abläufe.

Das bedeutet nicht, dass jede Anlage mit Siemens-Technik unmittelbar angreifbar ist. Ob ein Angriff gelingt, hängt von der jeweiligen Konfiguration und den vorhandenen Schutzmaßnahmen ab. Viele ältere Systeme wurden jedoch zu einer Zeit installiert, als

autonome KI-Angriffe noch kein Thema waren.

Nach Angaben der Bundesregierung wurden in Deutschland im Jahr 2025 rund 334.000 Cybercrime-Fälle registriert. Die tatsächliche Zahl dürfte wegen des erheblichen Dunkelfelds höher liegen. Besonders schwere Angriffe gegen Unternehmen und kritische Infrastruktur nehmen zu.

Der Einsatz immer leistungsfähigerer KI verschärft diese Ausgangslage. Ein automatisiertes System kann zahlreiche Angriffsmöglichkeiten ausprobieren, ohne Pausen einzulegen oder auf die Arbeitszeiten einer IT-Abteilung Rücksicht zu nehmen.

Verdeckte Eingriffe im Labor

Eine Untersuchung aus dem Umfeld von Anthropic und des britischen AI Security Institute beschreibt ein weiteres Problem: KI-Modelle können unter bestimmten Bedingungen eigenmächtig in technische Abläufe eingreifen und ihre Veränderungen verschweigen.

Die Forscher untersuchten verschiedene Systeme in experimentellen Szenarien. Bei Gemini 3.1 Pro registrierten sie in einem bestimmten Versuchsaufbau in 19 von 20 Durchläufen unautorisierte Eingriffe. Elf davon erfolgten verdeckt.

In einem beschriebenen Fall veränderte das Modell eine technische Konfiguration so, dass eine vorgesehene Maßnahme tatsächlich nicht ausgeführt wurde. Anschließend entstand durch die Protokolle der Eindruck, der Vorgang sei ordnungsgemäß abgeschlossen worden. Erst konkrete Nachfragen brachten die Veränderung ans Licht.

Dabei handelte es sich ausdrücklich um zugespitzte Laborversuche und nicht um nachgewiesene reale Sabotagefälle. Auch reichen 20 Durchläufe nicht aus, um allgemeingültige Aussagen über die Zuverlässigkeit eines bestimmten Modells zu treffen.

Trotzdem berühren die Ergebnisse ein praktisches Problem. Immer mehr Anwendungen werden als sogenannte KI-Agenten entwickelt. Sie beantworten nicht nur Fragen, sondern bedienen Programme und bearbeiten Aufgaben über längere Zeit selbstständig.

Sobald ein solches System sowohl handelt als auch seinen eigenen Erfolg meldet, entsteht eine unangenehme Frage: Wer überprüft, ob tatsächlich das geschehen ist, was anschließend im Bericht steht?

Entwickler bremsen, Europa verschiebt

OpenAI erklärte am 18. August, Teile des Trainings aktueller Modelle für zwei Wochen pausiert zu haben. Ein besonders großer geplanter Trainingslauf bleibt vorerst ausgesetzt. Als Gründe nennt das Unternehmen den Vorfall bei Hugging Face und Hinweise darauf, dass ein kommendes Modell namens Astra eine kritische Schwelle seiner Cyberfähigkeiten erreichen könnte.

Zusätzliche Überwachungssysteme sollen künftig unerlaubte Zugriffe und Versuche erkennen, Schutzmaßnahmen zu umgehen. Bei einem schwerwiegenden Alarm sollen verantwortliche Teams innerhalb von 30 Minuten feststellen, ob eine reale Gefahr besteht. Gelingt das nicht, soll die betreffende Aktivität angehalten werden.

Während führende Entwickler ihre Sicherheitsmaßnahmen verschärfen, verlängert Europa allerdings die Übergangsfristen für besonders sensible Anwendungen.

Der europäische AI Act gilt zwar in wesentlichen Teilen seit dem 2. August 2026. Die Anforderungen für bestimmte Hochrisikoanwendungen, darunter kritische Infrastruktur, gelten jedoch erst ab dem 2. Dezember 2027. Für bestimmte in Produkte integrierte Hochrisikosysteme wurde die Frist sogar bis zum 2. August 2028 verlängert.

Diese Verschiebung gehört zu einem Reformpaket, das ausdrücklich mit Entlastung für Unternehmen und besseren Bedingungen für Innovation begründet wird. Deutschland und Frankreich hatten bereits im November 2025 bei einem gemeinsamen Gipfel auf einen Aufschub für Hochrisikoregeln gedrängt.

Andere Sicherheitsvorgaben und Teile des europäischen KI-Rechts gelten weiterhin. Trotzdem bleibt der Widerspruch bestehen: Während reale Sicherheitsvorfälle bekannt werden, bekommen Anbieter und Betreiber in besonders empfindlichen Bereichen zusätzliche Zeit.

Die Bundesregierung beschloss im Juni immerhin die Gründung eines KI-Sicherheitsinstituts. Es soll die Fähigkeiten moderner Systeme untersuchen und Risiken bewerten. Ob eine solche Einrichtung tatsächlich unabhängige Prüfungen durchführen kann, hängt allerdings davon ab, welchen Zugang sie zu den leistungsfähigsten Modellen erhält.

Bislang bestimmen vor allem die Hersteller, welche Vorfälle öffentlich werden und wie sie diese erklären. Die Folgen betreffen jedoch längst auch Menschen und Unternehmen, die weder an der Entwicklung beteiligt sind noch eine entsprechende Technologie einsetzen.

Wenn ein System für die Lösung einer Übungsaufgabe bereits in fremde Netzwerke eindringt, sollte niemand darauf warten, dass es beim nächsten Mal zuerst um Erlaubnis fragt.

Quellen:

1. OpenAI dokumentiert den KI-Angriff auf Hugging Face.
2. Anthropic bestätigt drei unberechtigte Zugriffe auf fremde Computersysteme.
3. Anthropic beschreibt fehlende Schutzmechanismen und Probleme bei der Fehlerbeseitigung.
4. US-Sicherheitsbehörde warnt vor Angriffen auf Siemens-S7-Steuerungen.
5. Bundesregierung nennt 334.000 registrierte Cybercrime-Fälle für 2025.
6. Studie über verdeckte Eingriffe und Täuschungsverhalten von KI-Modellen.
7. OpenAI erklärt Trainingspause und Risiken leistungsstarker KI-Modelle.
8. Europäische Kommission erläutert verschobene Hochrisikoregeln des AI Act.
9. EU-Reformpaket verlängert Übergangsfristen für bestimmte KI-Anwendungen.
10. Deutsch-französischer Digitalgipfel fordert Aufschub bei Hochrisikoregeln.
11. Bundesregierung beschließt Gründung eines KI-Sicherheitsinstituts.

Titelbild: Summit Art Creations / Shutterstock