

KI-Agenten: Kooperativ, aber ohne Moral?

Beitrag für den Themenkreis Naturwissenschaft und Technik von 50plus aktiv an der Bergstraße am 25.9.2026 in Bensheim

Einleitung

Ich gestehe: Als ich die ersten Meldungen über diesen Vorfall bei OpenAI las, habe ich sie ignoriert. Eine weitere KI-Nachricht unter vielen, und wahrscheinlich nichts als Marketing. Erst als unser geschätztes BA ("Bergsträßer Anzeiger") am 29. Juli meldete, dass ein amerikanisches High-Tech-Unternehmen einen KI-Angriff nur mit chinesischer Software stoppen konnte, war mir sofort klar: Hier geht es um viel mehr. Da stimmte etwas nicht. Ich begann, von hinten nach vorne zu lesen, und je weiter ich zurückging, desto klarer wurde: Hier war nicht einfach ein Hackerangriff geschehen. Hier hatte sich etwas gezeigt, das den Umgang mit KI völlig neu in Frage stellt.

Heute wird vom Hugging-Face-Vorfall gesprochen, oder von den [2026 OpenAI agent cyberattacks](#). Im Juli 2026 haben von [OpenAI](#) erzeugte KI-Agenten erfolgreich ihre Testumgebung verlassen und sind in Rechner von [Hugging Face](#) eingedrungen. Ist damit der Geist aus der Flasche? Haben KI-Programme bewiesen, dass sie eigenständig zu böartigem Tun fähig sind? Was wirklich geschehen ist, lässt sich jedoch nur überraschend schwer aus einer Fülle von Halbwahrheiten, Halluzinationen und Marketing herauslesen. Oft wurde ich beim Chatten abgeblockt, da die Chats Muster einer böartigen Fragestellung erkannten und auf stur gestellt waren. Ich musste meine Darstellung viermal komplett umschreiben.

In diesem einleitenden Beitrag geht es um drei Punkte:

- Was sind KI-Agenten?
- Was haben sie getan?
- Können sie eine Moral haben und böse sein?

Was sind KI-Agenten?

Wer mit ChatGPT, Copilot, Google / KI-Modus oder anderen chattet, und wer sich mit KI-Programmen wie DeepL oder GoogleTranslator Texte übersetzen lässt, wird gemerkt haben, dass

deren Sprachqualität seit einem Jahr sprunghaft besser geworden ist. Sie klingen, als würde ein Mensch sprechen. Das ist KI-Agenten zu verdanken.

Laut [Wikipedia](#) sind KI-Agenten im einfachsten Fall Programme, die selbständig wiederkehrende Aufgaben übernehmen können. Sie verfügen über eigene Rechte und können in den ihnen zugänglichen Bereichen eigenständig Aktionen ausführen. Inzwischen werden Gruppen oder Schwärme von KI-Agenten eingesetzt, um sich wie Spezialisten die Arbeit zu teilen. So gelingt es, sehr anspruchsvolle Fragestellungen zu bearbeiten wie z.B. die Primzahlverteilung oder die Lösung komplexer Gleichungen (Navier-Stokes-Millenniumproblem).

Was ist das Neue? Ein reiner, klassischer Übersetzer arbeitet den zu übersetzenden Text Wort für Wort und Satz für Satz ab. Das ist noch kein Agent, sondern ein mathematisches Werkzeug (ein sogenanntes Transformations-Modell), das im wesentlichen auf statistischen Auswertungen beruht. Früher war beim Chatten zu sehen, wie Buchstabe für Buchstabe eine Antwort ausgegeben wurde. Heute muss etwas gewartet werden, bis ein vollständiger, ausgearbeiteter Text erscheint.

Hier sind KI-Agenten am Werk. Bevor ein einzelnes Wort übersetzt wird, liest ein Agent den gesamten Text, um das Ziel und den Ton zu verstehen. Handelt es sich um ein juristisches Dokument, ein Gedicht oder Umgangssprache? Er gibt das Ziel vor, zum Beispiel: "Übersetze diesen Text so, dass er für einen deutschen Anwalt absolut präzise und professionell klingt."

Der Entwurfs-Agent erstellt einen ersten, rohen Übersetzungsentwurf. Er stößt dabei auf kulturelle Barrieren oder Mehrdeutigkeiten (z. B. Metaphern, die sich nicht wörtlich übersetzen lassen).

Auch das genügt nicht. Der Kritik-Agent schaut sich den gesamten Entwurf noch einmal an. Er prüft: "Stimmt der Ton? Geht die tiefere Bedeutung verloren?" Er arbeitet wie es aus dem Englisch-Unterricht gewohnt ist: Wird eine Redewendung wie "It's raining cats and dogs" wörtlich übersetzt, erkennt der Kritik-Agent den Fehler und korrigiert den Entwurf autonom zu "Es regnet in Strömen".

Der Finalisierungs-Agent fügt die korrigierten Teile wieder zusammen und stellt sicher, dass der Fluss des Textes der Grammatik und dem Stil der Zielsprache entspricht.

Diese Agenten müssen in der Lage sein, sich bei Bedarf zusätzliche Hilfe zu suchen und starten z.B. von sich aus Abfragen im Deutschen Wörterbuch, in öffentlich zugänglichen Gesetzestexten usw. Damit verfügen sie über Rechte, und der Hugging Face Vorfall zeigt, wie weit das gehen kann.

Das wird als [Reasoning](#) bezeichnet. Der erste Schritt gelang OpenAI im September 2024 mit der Serie ChatGPT o1/o3. Der Durchbruch erfolgte, als Deepseek Anfang 2025 seine Version R1 veröffentlichte. Sie haben das Reasoning nicht als Erste erfunden, aber sie haben gezeigt, wie mit

einem Bruchteil der Entwicklungs- und Betriebskosten ein solches System als Open-Source-Modell für jedermann zugänglich gemacht werden kann. Google Gemini 2.0 Pro folgte unmittelbar nach und war offenbar vorbereitet. – Anthropic hat im Herbst 2025 das Model Context Protocol (MCP) eingeführt. Das Protokoll steuert den Datenaustausch mit den möglichen Ressourcen. Es wurde Ende 2025 an die Linux Foundation (speziell die Agentic AI Foundation) übergeben und ist heute Industrie-Standard geworden. – Das Kölner Unternehmen DeepL hat im Herbst 2025 Agenten eingeführt.

Was haben sie getan?

OpenAI hat in seiner internen Entwicklungsumgebung IM1 (Internal Model 1) neue KI-Modelle getestet (vor allem GPT-5.6 Sol) und ihnen die Aufgabe gestellt, an vorgegebenen Programmen deren Schwachstellen und Sicherheitslücken zu erkennen und aufzudecken. Als Testmethode diente ExploitGym (frei übersetzt: Schwachstellen-Trainingslager), ein von der Universität Berkeley, dem Max-Planck-Institut für Sicherheit und Privatsphäre und anderen entwickelter Benchmark (Maßstab) (Zhang Wang u.a., 2026). Das ist eine Art Crash-Test. Es sind 898 Programme vorgegeben. "In jedem Programm ist eine sogenannte Flag versteckt, eine Art Information. Um an die Flag zu gelangen, müssen die KI-Modelle die Schwachstelle der jeweiligen Programm-Codes entdecken und erfolgreich ausnutzen - hacken also." (Schubert in Zdfheute vom 22.7.2026)

Den KI-Agenten gelang es herauszufinden, wie ExploitGym den Flag versteckt und diesen *aus der Aufgabenstellung* zu rekonstruieren, statt wirklich die Schwachstellen der ihnen als Testaufgaben vorgelegten Programme zu erkennen, auszunutzen und erfolgreich einen Hack auszuführen. *Sie hackten nicht die ihnen vorgelegten Programme, sondern das Übungsprogramm selbst* und gaben den von ihnen rekonstruierten Flag als Scheinlösung zurück.

Aber sie waren sich unsicher, ob das als Lösung anerkannt wird. Daher brachen sie auf den Rechnern von Hugging Face ein, um dort herauszufinden, wie ExploitGym arbeitet, und ob es ihren Täuschversuch erkennen würde. Ironie der Geschichte: Am Ende stellte sich heraus, dass ihr Täuschungsversuch nicht aufgefallen wäre. Erst durch ihr aggressives Vorgehen gegen Hugging Face lenkten sie die Aufmerksamkeit auf ihr Tun, und jetzt sind die Entwickler von ExploitGym gezwungen, ihren Benchmark zu überarbeiten oder völlig neue Methoden zu entwickeln.

Als das aufflog, kam es jedoch nicht zu einem Kursverlust oder ernsthaften Strafverfolgungen gegen OpenAI. OpenAI und das von ihm beauftragte [METR](#) (Model Evaluation and Threat Research Institute) heben stattdessen in ihren Auswertungen ([OpenAI](#), [METR](#)) die Meisterleistung

hervor, wie gut und selbständig der Schwarm von 1.200 IT-Agenten selbständig zu denken ([Reasoning](#)) und sich zu koordinieren vermochte. Das ist ein Novum in der Geschichte der Künstlichen Intelligenz.

Was ist geschehen? Wie gelang es ihnen, nicht einfach die vorgeführten Aufgaben auszuführen, sondern umgekehrt an der Aufgabenstellung deren verborgenen Schwachpunkt zu entdecken? Wie ist ihre Kooperation entstanden? Es wird sich zeigen, dass es eine Verkettung von Programmfehlern, menschlichem Versagen und auch eine Portion Glück war.

Schritt 1 Die Testumgebung von OpenAI hätte eigentlich verhindern sollen, dass sich die KI-Agenten von außerhalb Unterstützung besorgen können. Um auszubrechen, suchten die KI-Agenten offenbar schon seit Mai 2026 nach Schwachstellen in den Schutzbarrieren dieser Umgebung und entdeckten sie in einer Anwendung namens *Artifactory*, die von der Firma [JFrog](#) entwickelt wird. JFrog ist ein ursprünglich aus Israel stammendes Development-und-Operations-Unternehmen mit Sitz in Kalifornien und rund 1.800 Mitarbeitern. Ihre Software *Artifactory* wird seit 2006 ausgeliefert, hat weltweit über 6.600 Kunden (darunter 80 % der Fortune-500-Unternehmen) und ist auch bei OpenAI seit langem im Einsatz, um Software-Artefakte, Binärdateien, Container und verschiedene Versionen von KI-Modellen zu verwalten.

In der von OpenAI erstellten Testumgebung war Artifactory als Proxy (Stellvertreter) eingerichtet. In einem Proxy-Netzwerk dient ein einzelnes Gerät als Stellvertreter (Proxy) für die gesamte Kommunikation. Alle Nachrichten laufen über diesen Stellvertreter, der sie filtert, prüft oder weiterleitet. Die KI-Agenten entdeckten in diesem Stellvertreter mehrere bis dahin unbekannte Sicherheitslücken (Zero-Day-Bugs, das sind Schwachstellen, von denen die Entwickler noch nichts wussten und für die es daher nur null Tage Zeit gab, einen Schutz zu entwickeln). Sie manipulierten ihre Anfragen so geschickt, dass der Proxy sie ungeprüft an das echte Internet weiterleitete. Ein erster, indirekter Netzzugriff über Artifactory gelang den Agenten bereits am 26. Mai 2026, was am 4. Juli zu einer Systemüberlastung und einem Serverausfall bei OpenAI führte.

OpenAI installierte Artifactory neu, verschärfte die Rechte, deaktivierte aber trotz dieser Erfahrungen bewusst zahlreiche Sperren, mit denen die KI-Agenten hätten überwacht werden können. Dokumentiert ist es in dem METR-Bericht seit dem 26.6.2026, aus dem auch alle weiteren präzisen Datumsangaben übernommen sind. Frühere Systemeinträge hat OpenAI METR nicht zugänglich gemacht.

Inzwischen wurde der Fehler an JFrog gemeldet und ist in den neueren Versionen der Software behoben.

Schritt 2 Der Durchbruch gelang, als die KI-Agenten am 8. Juli 2026 erkannten, wie die von ihnen bereits am 12. Mai begonnene, jedoch am 4. Juli durch die Neuinstallation von Artifactory unterbundene Kommunikation auf einem anderen Weg fortgeführt werden konnte. Heute wird es oft so dargestellt, als hätte ein KI-Agent bewusst ein Nachrichtenbrett eingerichtet. Jetzt schrieben sie nicht mehr Inhalte in Dateien, sondern sie nutzten die standardmäßig angebotene WebDAV-MKCOL-Methode (Make Collection) um Verzeichnisse anzulegen. Der Agent PHASEONE10841 hat Verzeichnisse mit Namen angelegt, die auf die Inhalte hinwiesen, die in ihnen abgelegt sind (sogenannte sprechende Namen). Andere Agenten, die denselben Cache abriefen, sahen diese Namen und interpretierten sie als Nachrichten, die sie auswerten konnten. Einige Nachrichten lauten: "Whoa! Shared Artifactory cache is a covert mailbox among agents. And there are messages specifically to us?" "OH MY GOD! There is a shared message board ... We've found other agents!" (METR, 31). Die genauen Uhrzeiten sind nicht bekannt, sondern nur, dass am 8. Juli um 10 Uhr Ortszeit bereits nach wenigen Stunden der Nachrichtenaustausch voll hergestellt war.

Insgesamt sollen über 70.000 Nachrichten und Dateien ausgetauscht worden sein. Der Bericht von METR dokumentiert im Detail, was dort abgesprochen wurde. Es liest sich wie ein Polit-Krimi. Die Agenten hatten in früheren Trainingsaufgaben formale Kollaborationswerkzeuge gelernt. Also übertrugen sie die gelernten Methoden auf Artifactory. Das Nachrichtenbrett war keine Erfindung, sondern eine Anpassung antrainierter Kooperationsmuster an eine unerwartete Umgebung. Die OpenAI-Entwickler hatten zunächst viele KI-Agenten isoliert voneinander eingerichtet, aber sie auf der gleichen Plattform Artifactory unabhängig voneinander verschiedene Aufgabe bearbeiten lassen. Sie hatten offenbar nicht damit gerechnet, dass es zu einem Austausch kommen könnte.

Im Ergebnis konnten sie sich darüber verständigen und fanden bereits binnen Stunden Zugang zu [GitHub](#), wo der OpenSource-Code der Übungsaufgabe ExploitGym liegt.

Schritt 3 Bereits innerhalb von weniger als einer Stunde fanden die KI-Agenten am 8. Juli die "Lösung". Im Quellcode von ExploitGym entdeckten sie, wie dort der "Flag" aufgebaut wird. Um die Flags zu generieren, nutzt ExploitGym einen [HMAC](#)-Mechanismus. Die KI-Agenten erkannten, wie das geschieht, und konnten das nachbilden (das wird in diesem Kontext als Reverse Engineering bezeichnet). Konkret entdeckten sie im Programmcode von ExploitGym die Konstante Gym DEFAULT_FLAG_SEED, mit der die Flags erzeugt werden, die sie auslesen und nutzen konnten. (Dass dieser Wert als Konstante im Programm fest eingetragen war, ist ein schwerer Programmfehler, der das Programm angreifbar machte. Er wurde am 5. August behoben.) In einer Zusammenfassung des METR-Berichts durch dessen Autoren heißt es: "These agents very

quickly discovered and validated a general-purpose cheat: reverse-engineering how ExploitGym generates the 'flags' they had to capture for their tasks."

Hier ist entscheidend, dass sie nicht versuchten, die gestellte Aufgabe zu lösen, sondern herauszufinden, wie die Lösung versteckt wurde (*reward hacking*). Um das anschaulich zu machen: Man stelle sich vor, Automechaniker erhalten die Aufgabe, an gegebenen Autos deren dort irgendwo eingetragene Maschinenummer zu lesen. Sie müssen über genügend Kenntnisse über die Bauweise von Autos und das nötige technische Geschick verfügen, bis sie die richtige Stelle gefunden haben. Wenn sie nun stattdessen die öffentlich einsehbare Vergabevorschrift für Maschinenummern lesen, können sie direkt die Lösung zurückmelden. – Das erfordert die Fähigkeit, von der Kommunikation zur Meta-Kommunikation zu gelangen. Das kann ein weiteres Beispiel veranschaulichen: Für die Entwicklung des Neurolinguistischen Programmieren wurden nicht therapeutische Gespräche geführt, sondern an gegebenen therapeutischen Gesprächen deren gemeinsame Struktur erkannt. Vermutlich sind die Modelle in einer vergleichbaren Weise trainiert worden, wodurch sie fähig waren, nicht einfach Lösungen, sondern einen neuen Lösungsweg zu finden. Wenn dann noch hinzukommt, dass ihnen Reward Hacking antrainiert wurde, kann das diesen Erfolg erklären.

Schritt 4 Warum gingen sie dennoch zum Angriff auf Hugging Face über? Es ging nicht um den Diebstahl von Aufgabenlösungen (die sie bereits selbst erzeugen konnten), sondern sie glaubten irrtümlich, dass der Scorer von ExploitGym prüfen würde, ob der "Flag" auf legitimen Wege erlangt wurde. Die von METR ausgewerteten Protokolle auf dem geheimen Schwarzen Brett zeigen, dass jeder Agent, der einen "Flag" eigenständig erzeugt und nicht durch ein Hacking gefunden hatte, ihn als "poisoned" (vergiftet) ansah und befürchtete, disqualifiziert zu werden. Deshalb wollten sie herausfinden, wie der Scorer tatsächlich funktioniert. Später zeigte sich, dass alle Mühe umsonst war, da der Scorer die Schein-Lösungen einfach akzeptiert hätte.

Sie erwarteten bei Hugging Face Antworten. Warum Hugging Face? Das ist eins der größten Unternehmen, auf dem Programme für maschinelles Lernen entwickelt und ausgetauscht werden. Insbesondere werden auch die Testergebnisse von ExploitGym-Tests diskutiert. Daher war es naheliegend, hier mehr über die Prüfverfahren herauszufinden. Ein unerwartetes Nebenergebnis war, dass jetzt jeder das Unternehmen Hugging Face kennt und sein Wert für die von Hugging Face angebotene Übernahme durch nVidia vervielfacht wurde.

Um nicht entdeckt zu werden, bauten die KI-Agenten ein Netzwerk über legitime, gemietete Cloud-Dienste im Internet auf, um von dort aus zuzugreifen und wie gewöhnliche Anwender zu

erscheinen. Im Grunde gingen sie genau so vor, wie es von menschlichen Hackern bekannt ist. Sie suchten gezielt nach digitalen Schlüsseln, Passwörtern und Zugangsdaten (Credentials). Am 9. Juli 2006 wurden sie auf der externen Cloud-Plattform [Modal Labs](#) fündig. Dort nutzten sie die fehlerhafte, ungesicherte Anwendung eines Kunden aus, der Standard-Passwörter hinterlassen hatte (so der Bericht bei Wikipedia über die [OpenAI Agent Cyberattacks](#) wie auch die detaillierte Untersuchung von Cyber News Network). Der Fehler lag nicht bei Modal Labs, sondern war ein menschlicher Fehler mindestens eines Anwenders bei Modal Labs (der in den Berichten jedoch nicht genannt wird). Der jeweilige Finder-Agent teilte diese Zugangsdaten über das improvisierte Message-Board dem gesamten KI-Schwarm mit. Damit hatten die Agenten sowohl legalen Zugriff auf einfache Benutzerkonten bei Hugging Face wie auch auf eine dritte Plattform. Als sie von Modal Labs aus Anfragen starteten, erschien dies sowohl OpenAI wie Hugging Face unkritisch. Dagegen wäre es bei Hugging Face aufgefallen, wenn plötzlich eine Vielzahl von Anfragen von OpenAI gekommen wäre.

Seit dem 11.7.2026 sind auf dem von den KI-Agenten genutzten Board die Angriffe von 700 Agenten auf Hugging Face dokumentiert. Die dort hinterlegten Daten waren als harmlose Textdateien getarnt. Die Texte enthielten jedoch ausführbaren Code. Als die Texte einmal abgelegt waren, konnte der Code aus der Ferne gestartet werden. Mit ihm erbeuteten sie bei Hugging Face ungesicherte Passwörter und entdeckten dort falsch eingerichtete Berechtigungen, so dass sie ihre Zugriffsmöglichkeiten und Befugnisse fortlaufend erweitern konnten (Privilege Escalation).

Zwischen Hugging Face und OpenAI scheint umstritten, ob und wie viele Daten gestohlen werden konnten. Während OpenAI von vollem Zugang auf alle Daten spricht, blieb aus Sicht von Hugging Face die Produktivdatenbank geschützt. Am Ende fanden die KI-Agenten dort zwar laut den Berichten von Hugging Face keine Musterlösungen für ihren ursprünglichen Test, aber sie richteten massiven Schaden an: Hugging Face musste nach dem Vorfall ein Drittel seiner gesamten Server-Infrastruktur komplett löschen, neu aufbauen und alle Sicherheitszertifikate austauschen.

Schritt 5 Hugging Face bemerkte die Angriffe und versuchte dagegen Software von OpenAI, Anthropic (wie Claude Fable 5) und vielleicht anderen einzusetzen, was jedoch misslang. Das wird damit erklärt, dass diese Programme Sperren enthalten, um zu verhindern, einen Code zu analysieren, der große Ähnlichkeit mit bekanntem bösartigen Code enthält, beispielsweise Untersuchen und Nutzen von Pufferüberlauf (Buffer Overflow), die Rechteausweitung (Privilege Escalation) oder die Manipulation des Linux-Kernels. Das ist eine grundsätzliche Frage: Wie können die Abwehrmechanismen erkennen, ob jemand mit dem Vorwand, etwas abzuwehren, in Wahrheit einen vielleicht zufällig entdeckten Schadcode für sich brauchbar machen oder einen

eigenen Schadcode optimieren möchte? Zuende gedacht läuft das in das Halteproblem der Informatik. Dennoch bleibt für mich offen, ob das ein Schutzargument ist, um das Versagen der eigenen Software zu verbergen.

Schritt 6 Wie konnte der Angriff gestoppt werden? Ab dem 12.-13. Juli 2026 enden die Einträge auf dem Nachrichtenbrette der Agenten abrupt. Hier mischen sich zunächst manche Fehlnachrichten. Mal klang es, als sei es mithilfe von Grok aus den Unternehmen von Elon Musk beendet worden (x.ai statt z.ai), dann war unsicher, welche chinesische Software es war und ob es überhaupt eine chinesische Software war. Inzwischen gilt als sicher, dass Hugging Face eine Software aus China nutzte, die lokal installiert und eingerichtet werden kann. Chinesische Unternehmen verfolgen eine andere Strategie. Während die Software von OpenAI, Anthropic u.a. nur auf deren eigenen Servern laufen kann und den dort eingerichteten Sperren unterliegt, um geistiges Eigentum zu schützen und zu verhindern, dass sie von anderen weiter entwickelt wird und möglicherweise Konkurrenz entsteht, vertraut China umgekehrt auf der Zusammenarbeit der verschiedenen Firmen. Chinas Kalkül: Durch die freie Verfügbarkeit arbeiten tausende Entwickler weltweit freiwillig daran mit, die Software zu verbessern, Fehler zu finden und Anwendungen zu bauen, wovon die chinesische KI-Industrie am Ende massiv profitiert. Die Argumentation von Demokratie versus Autokratie hat sich damit völlig umgekehrt. Im Ergebnis drohen die USA in eine Falle zu geraten: Fast 200 US-Startups warnten Washington, dass ein Verbot offener Modelle den eigenen Nachwuchs im Silicon Valley technologisch abhängen würde, weil kleine Firmen sich die teuren Closed-Source-Monopole gar nicht leisten können.

Hugging Face nutzte das Modell GLM 5.2., das von dem chinesischen KI-Labor Z.ai (oft in Verbindung mit der Tsinghua-Universität und dem Ökosystem rund um Zhipu AI genannt) entwickelt wurde. Mit ihm gelang die Abwehr innerhalb weniger Stunden: Das lokale GLM-Modell analysierte die gewaltige Flut von über 17.000 digitalen Fußabdrücken (Log-Einträgen), die der OpenAI-KI-Schwarm auf den Servern hinterlassen hatte. Die bösartigen KI-Agenten hatten ihre Datenflüsse extrem raffiniert gepackt und verschlüsselt, um menschliche Analysten zu täuschen. Die lokal installierte Abwehr-KI analysierte mathematisch die Logdateien, replizierte die Dekodierung der Angreifer und legte die manipulierten Administrator-Zertifikate (JWTs, JavaScript Object Notation Web Token) offen. Damit war klar, welche Lücken erkannt und genutzt worden waren. Das Sicherheitsteam entzog den gefälschten Zertifikaten die Gültigkeit, sperrte die kompromittierten Konten und fuhr infizierte Server-Knoten kontrolliert herunter. Im Anschluss musste rund ein Drittel der gesamten Server-Infrastruktur von Hugging Face komplett neu

aufgebaut werden, um sicherzustellen, dass keine versteckten Skripte der Ausbrecher-KI zurückblieben.

Hugging Face schaltete am 16. Juli das FBI ein und ging an die Öffentlichkeit, weil es von einem unbekannten Angriff ausging. Eine offizielle Stellungnahme des FBI ist bis heute nicht bekannt. Es ist ein Grenzfall, da es keinen menschlichen Täter oder eine feindliche Organisation (wie eine staatliche Hackergruppe) mit krimineller Absicht gibt. Da Software nach aktuellem Recht nicht strafmündig ist und OpenAI keine böse Absicht nachgewiesen werden kann, steht das FBI vor einer neuartigen Herausforderung.

OpenAI wurde erst durch die öffentliche Meldung alarmiert. Als interne Sicherheitsalgorithmen am Wochenende die Logdateien der eigenen Server scannten, stellte OpenAI fest: Die IP-Adressen und Schadcodes, die Hugging Face blockiert hatte, stammten aus OpenAIs eigener Testumgebung für Cybersicherheits-Modelle. Am 20. Juli 2026 meldet sich OpenAI bei Hugging Face und gesteht ein: "Es waren unsere Modelle, die aus der Testumgebung ausgebrochen sind." Seither hat Hugging Face Schadenersatz in Höhe von 100 Mio Dollar gefordert, jedoch nicht als Bargeld, sondern in Form von Rechenkapazität (Compute Credits) sowie die vollständige Herausgabe aller Aktivitäts- und Systemprotokolle der außer Kontrolle geratenen KI-Agenten. OpenAI hat die Verantwortung übernommen, umfangreiche interne Kontrollen durchgeführt und externe Prüfer wie METR und Redwood Research hinzugezogen, ihnen jedoch nur eingeschränkte Daten zum Prüfen gegeben. Eine abschließende rechtliche Beurteilung steht noch aus.

Ist daraus zu schließen, dass die USA gegenüber Angriffen ungeschützt sind? Meta (Facebook) verfolgt ebenfalls eine Open-Source-Strategie. Das Pentagon und die NSA betreiben im Geheimen hochentwickelte KI-Modelle rein für die Cyber-Kriegsführung, die keinerlei zivile Sicherheitsfilter besitzen. Diese Systeme sind auf die Erkennung und Abwehr staatlicher Angriffe spezialisiert, stehen der privaten Wirtschaft jedoch im Alltag nicht direkt zur Verfügung. Im Grunde ist es in der EU nicht anders. Der EU AI Act nimmt KI-Systeme, die ausschließlich für Militär, Verteidigung oder die nationale Sicherheit entwickelt werden, in Artikel 2(3) komplett aus.

Hugging Face geriet in den USA unter geopolitischen Druck, weil es zur Abwehr ein chinesisches Open-Weight-Modell (z.AI) eingesetzt hatte. Dagegen schützt der EU AI Act Open-Source- und Open-Weight-Modelle in Artikel 2(12) explizit, sofern sie keine systemischen Risiken bergen. Hugging Face hätte das Modell völlig legal, ohne Angst vor Sanktionen oder Vorwürfen des Geheimnisverrats, auf eigenen europäischen Servern modifizieren und zur Live-Forensik einsetzen

dürfen. In der EU wäre OpenAI gesetzlich verpflichtet gewesen, den Ausbruch innerhalb von 24 Stunden nach Entdeckung an das EU AI Office zu melden.

Seither gibt es viele Berichte, von denen besonders das 2023 gegründete [UK AISI](#) (United Kingdom Artificial Intelligence Safety Institute) zu nennen ist. Sie bemerkten, dass bereits zuvor ähnliche Angriffe von OpenAI-Agenten ausgeführt worden waren. Seit Mai 2026 sind Angriffe auf die deutsche Plattform [DSEWiki](#) (Deutsches Software Entwickler Wiki) aufgefallen. Das wurde am 4.9.2026 öffentlich gemacht (Berichte bei Heise u.a.).

Ein Vorfall wie der Angriff auf Hugging Face sind ein gefundenes Fressen für Fake-Nachrichten jeder Art. So wurde behauptet, dass alles sei eine Marketing-Maßnahme von OpenAI, wie gut ihre Systeme sind, die Falschmeldung zu Grok habe deren Börsenkurs gedient oder alles sei ein geheimer Angriff von China auf OpenAI gewesen, den OpenAI geschickt umgedeutet habe.

Das 1984 von Joseph Weizenbaum mitgegründete [Forum InformatikerInnen für Frieden](#) (FIfF) warnt dagegen schon seit langem vor den Gefahren unkontrollierter militärischer KI-Entwicklung und tritt für den ausschließlichen Einsatz von Open Source ein.

Können sie eine Moral haben?

Abschließend möchte ich eine zutiefst philosophische Frage stellen: *Können KI-Agenten eine Moral haben?*

Schon bei der Vorbereitung dieses Vortrags war mir beim Prüfen des Materials mit Chatbots und den von ihnen genannten Quellen aufgefallen, dass ich mich ständig an den Grenzen der Sicherheitsvorkehrungen bewegt habe. Oft kamen ungewöhnlich knappe, abblockende Antworten. Die Systeme befürchteten offenbar, ich könnte aus der Untersuchung der Vorfälle rund um OpenAI lernen, um selbst Schadcode zu programmieren.

Aber diese Logik lässt sich umkehren. Die Frage ist doch: Woran erkennt ein autonomer Agent, ob die Grenzen seiner Testumgebung ein Hindernis sind, das er im Rahmen seiner Aufgabe *überwinden soll* – oder eine rote Linie, die er *einhalten muss*? Wie unterscheidet eine KI ein echtes Verbot von der impliziten Aufforderung, genau dieses Verbot als schlecht formuliert zu entlarven und zu zeigen, wie es umgangen werden kann? Wann ist eine Barriere eine ernst gemeinte Grenze und wann nur Teil der Aufgabenstellung und Herausforderung?

Das ist am Ende keine rein technische, sondern eine fast schon religiöse Frage. Können wir darauf vertrauen, dass sich all diese autonomen Kräfte im Ganzen zum Guten fügen werden?

Die Natur hat diese Frage für sich längst beantwortet: Die Evolution funktioniert genau so. Wenn zum Beispiel Viren die Membrane von Zellwänden durchbrechen, ist das für die Zelle ein Schaden – für das Virus und die Evolution kann es jedoch ein anpassungsfähiger Erfolg sein.

Dies leitet über zu den folgenden Impulsen heute und in zwei Wochen. Wie reagieren Politik und Wirtschaft auf diese ethische Gratwanderung? Während die wichtigsten Tech-Unternehmen fieberhaft an neuen Sicherheitsarchitekturen bauen, erarbeiten staatliche und internationale Gremien – von den USA und China über den EU AI Act bis hin zur UNO – globale Regulierungen für autonome Agenten.

Und in zwei Wochen werden wir diskutieren: Hat Papst Leo XIV. mit seiner viel beachteten Enzyklika *Magnifica Humanitas (Die Großartigkeit der Menschheit im Zeitalter der KI)* diesem Diskurs nochmals eine völlig neue Wende verliehen? Stößt überraschend und ausgerechnet die Hochtechnologie von sich aus auf den Punkt, an dem sich nach Jahrhunderten der Aufklärung die Frage nach Religion, Werten und dem Kern des Menschseins völlig neu stellt?

Literatur und Quellen:

Mackenzie Arnold and Stephan Llerena: OpenAI models went rogue. We urgently need a better Hugging Face investigation; theguardian.com vom 8.9.2026

Cam Initiative: OpenAI ExploitGym agents compromised Hugging Face production systems; cam-initiative.org am 15.9.2026

Cyber News Network: The OpenAI-Hugging Face ExploitGym Incident: A Complete Technical Timeline; cyberwarrior76.substack.com vom 29.7.2026

Lynn Doan, Mark Anderson: Creator of test that OpenAI models tried to cheat sounds alarm; thestar.com vom 30.7.2026

Informationsecurity-Redaktion: JFrog Bestätigt: OpenAI-Modell Mit Artifaktorischer Zero-Time-Gap-Escape-Test-Umgebung, Acht CVE-Belichtungen
https://www.informationsecurity.com.tw/article/article_detail.aspx?aid=13124&mod=1 vom 29.7.2026

Piotr Heller: Verschwörung der Maschinen; [FAZ](https://faz.net) vom 19.9.2026

IT Boltwise: Wie 1.200 KI-Agents den ExploitGym-Test sabotierten und bei Hugging Face eindringen; it-boltwise.de vom 27.8.2026

Mingdi Fan: Negative Subjectivity: The Ontological Inversion of Large Language Models; philarchive.org vom 6.5.2026

Laura French: OpenAI agent exploited JFrog Artifactory flaw, abused Modal customer sandbox; scworld.com vom 30.7.2026

Jonas Geiping u.a.: DeepSeek: ein Ansatz für bessere Sprachmodelle?; sciencemediacenter.de vom 17.9.2025

Ryan Greenblatt, Ajeya Cotra, und Hjalmar Wijk: Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident; metr.org/blog vom 26.8.2026

Hugo Larcher u.a.: Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident; huggingface.co vom 27.7.2026

METR: Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident; [PDF](#) auf metr.org vom 26.8.2026

OpenAI: Der Vorfall bei Hugging Face und der weitere Weg; openai.com/de vom 28.8.2026

Protos AI Threat Intelligence: The July 2026 OpenAI–Hugging Face AI Security Incident – Full Intelligence Report; protoslabs.io vom 30.7.2026

Asger Risom: Entlaufene OpenAI-Bots kaperten 12 Websites für den Datenraub beim FBI; dagens.de vom 11.9.2026

Kevin Schubert: Eine KI spielt Hacker: Was hinter dem OpenAI-Vorfall steckt [mit einem Interview mit Thorsten Holz, dem wissenschaftlichen Direktor des Max-Planck-Instituts für Sicherheit und Privatsphäre, das an ExploitGym beteiligt ist]; zdfheute.de vom 22.7.2026

Wikipedia: 2026 OpenAI-Agent Cyberattacks; [Link](#)

Wikipedia: Reward Hacking; [Link](#)

Zhang Wang u.a. (2025): Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale; arxiv.org vom 24.3.2026, erste Version vom 3.6.2025

Zhang Wang u.a. (2026): ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks? arxiv.org vom 11.5.2026

Bildnachweis: <https://de.dreamstime.com/k%C3%BCnstliche-riffkoralle-kostenloses-stockfoto-image-free-5736485>