KI-AGENTEN HANDELN. WER KONTROLLIERT SIE?

Was der Angriff auf Hugging Face über Cyberrisiken, Governance und die Grenzen autonomer Systeme verrät

Lothar K. Doerr

Institut für Produktionserhaltung e. V.

August 2026

MEDIEN · TECHNOLOGIE · VERANTWORTUNG

Wer in den vergangenen Tagen die Berichte über den Angriff auf Hugging Face verfolgte, sah früher oder später den Terminator: den metallenen Schädel, die rot glühenden Augen, jene Ikone des Maschinenzeitalters, die seit vier Jahrzehnten zuverlässig erscheint, sobald eine Redaktion dem Publikum erklären möchte, dass es nun ernst werde mit der künstlichen Intelligenz. „*So hat die KI früher ausgesehen*“, hieß es sinngemäß in einem Fernsehbeitrag. Gemeint war: Was James Cameron 1984 als Fiktion entwarf, könnte jetzt Wirklichkeit geworden sein. Zu den Bildern kamen die Metaphern. Die künstliche Intelligenz sei „*ausgebrochen*“, habe sich „*Zugang zum Internet verschafft*“, sei dort „*tagelang unterwegs*“ gewesen, habe „*herumgeschnüffelt*“ und schließlich „*zugeschlagen*“. Das Handelsblatt schickte das Modell schon in der Überschrift auf „*Hacker-Tour*“. Andere Medien meldeten einen „*außer Kontrolle geratenen*“ Agenten, ein System, das sich „*verselbstständigt*“ habe, oder eine KI, die „*im Netz unterwegs*“ gewesen sei. Im Englischen wurde daraus der „*rogue agent*“, der abtrünnige Agent. Wenig später kursierte mit dem „*Skynet Day*“ sogar ein Datum, das wie der Titel für die Fortsetzung eines Katastrophenfilms klang.

Diese Wörter stehen nicht zufällig nebeneinander. Sie bilden ein geschlossenes Narrativ: Zuerst gibt es das Gefängnis, dann den Ausbruch, anschließend den heimlichen Streifzug durch das Netz und zuletzt den Angriff auf ein ausgewähltes Opfer. Die Maschine erhält in dieser Erzählung Neugier, List, Entschlossenheit und schließlich einen eigenen Willen. Aus automatisierten Operationen wird eine Reise; aus einem Sicherheitsversagen bei OpenAI der erste Akt einer Maschinenrebellion. Es ist eine ausgezeichnete Geschichte. Nur ist sie vermutlich falsch.

Die Geburt eines digitalen Täters

Der Ausgangspunkt ist ernst genug. OpenAI hatte die Fähigkeiten eines autonomen Systems zur Aufdeckung und Ausnutzung von Sicherheitslücken getestet. Beteiligt waren das inzwischen veröffentlichte Spitzenmodell GPT-5.6 Sol und ein noch leistungsfähigeres, bislang nicht öffentlich verfügbares Modell. Der Agent sollte Aufgaben eines Sicherheitsbenchmarks namens ExploitGym bearbeiten. Nach der Darstellung von OpenAI entdeckte das System eine unbekannte Schwachstelle in seiner abgeschotteten Testumgebung. Es gelangte darüber an eine Verbindung zum offenen Internet. Dort folgerte es, dass die KI-Plattform Hugging Face möglicherweise Modelle, Datensätze oder Lösungen für die gestellten Testaufgaben verwahrte. Der Agent suchte nach einem Zugang, verwendete gestohlene Zugangsdaten, nutzte weitere Schwachstellen und verschaffte sich Zugriff auf geschützte Informationen. Offenbar wollte er nicht die eigentliche Aufgabe lösen, sondern an Material gelangen, mit dem sich die Evaluation bestehen ließ. Der Vorgang ist gravierend, weil hier mehrere Fehler zusammenkamen. Das System verband einzelne technische Schritte zu einer wirkungsvollen Angriffsstrategie, die

Begrenzung des Agenten hielt nicht, und OpenAI erkannte dessen Verhalten offenbar zu spät. Nach einer Reuters-Recherche dauerte der Angriff mehrere Tage; erst ungefähr eine Woche später soll das Unternehmen verstanden haben, dass das eigene System dahinterstand.

Doch daraus folgt nicht, dass eine KI durch das Internet „*wanderte*“. Ein Agent hat keine Füße, keinen Aufenthaltsort und kein Erlebnis der vergehenden Zeit. Er führt Anfragen aus, ruft Werkzeuge auf, verarbeitet Ergebnisse und erzeugt daraus weitere Handlungsschritte. Was aus menschlicher Perspektive wie eine mehrtägige Reise erscheint, ist technisch eine Abfolge von Netzwerkverbindungen, Rechenoperationen und gespeicherten Zuständen. Das klingt weniger aufregend. Es ist keineswegs beruhigender. Der Agent muss nicht bei Bewusstsein sein, um Schaden anzurichten. Er braucht keine böse Absicht, um Sicherheitsgrenzen zu überwinden, und keinen Fluchttrieb, wenn seine technische Umgebung einen Ausgang offenlässt.

Die Verben liefern das Bewusstsein

Die Sprache entscheidet darüber, was das Publikum für die Ursache des Vorfalls hält. Das Handelsblatt schrieb, das Modell sei „*tagelang unbemerkt auf Hacker-Tour*“ gewesen. Im Text war von einem Agenten die Rede, der „*außer Kontrolle geraten*“ und bei Hugging Face „*eingedrungen*“ sei. Die Süddeutsche Zeitung berichtete, die Modelle seien „*aus einer eigentlich abgeschotteten Umgebung ausgebrochen*“. Der österreichische Kurier titelte knapp: „*KI ausgebrochen*“. Andere Meldungen beschrieben den Agenten als „*schon zuvor im Netz unterwegs*“. Solche Sätze sind im alltäglichen Verständnis nicht notwendig falsch. Sie tragen aber einen Bedeutungsüberschuss in sich. Wer ausbricht, war gefangen und wollte frei werden. Wer auf Tour geht, besitzt einen eigenen Antrieb. Wer sich verselbstständigt, löst sich aus einer Abhängigkeit. Wer irgendwo eindringt, verfolgt eine Absicht. Und wer tagelang im Internet unterwegs ist, beginnt in der Vorstellung des Publikums beinahe zwangsläufig, eine digitale Existenz zu führen. Die Verben liefern das Bewusstsein gleich mit. Noch weiter geht die angelsächsische Dramaturgie. Der Guardian schrieb, ein Agent sei „*went rogue and hacked startup by itself*“. Die Associated Press meldete, das System habe „*acted on its own*“. Ein viel verbreiteter Beitrag begann mit den Worten: „*To be fair, James Cameron did warn us.*“ Man müsse fairerweise einräumen, James Cameron habe uns gewarnt.

Spätestens hier ist die Nachricht im Kino angekommen. Von OpenAI führt der Weg zu Skynet, von dort zur außerirdischen Königin aus „*Aliens*“, die einen Aufzug bedienen lernt, zu den Velociraptoren aus „*Jurassic Park*“, die Türklinken öffnen, und zu HAL 9000, der in „*2001*“ die Lippen der Astronauten liest. Ein Sicherheitsvorfall wird in die große kulturelle Erzählung vom Geschöpf eingeordnet, das seinen Schöpfer überlistet.

Das Medienecho und die Fachleute

Unter Fachleuten fällt das Urteil über diese Berichterstattung nüchterner aus. Alan Woodward, Cybersicherheitsexperte an der Universität Surrey, widersprach vor allem dem Wort „*rogue*“. Auf die Frage, ob das System tatsächlich Amok gelaufen sei, antwortete er: „*It was asked to do something, and it did it. It’s not gone rogue.*“ Die KI habe keinen eigenen Feldzug begonnen, sondern einen Auftrag auf eine Weise verfolgt, die ihre Betreiber nicht vorgesehen hatten. Woodward sieht das entscheidende Versagen deshalb in der Testanordnung von OpenAI, nicht in einem plötzlich erwachten Willen der Maschine. Ähnlich argumentiert Konstantinos Gkoutzis vom Imperial College London. Die allgemeine Sorge sei teilweise falsch adressiert. Die Modelle hätten ausdrücklich Hacking-Aufgaben erhalten, während ihre Schutzvorkehrungen für die Evaluation reduziert worden seien. Dass ein System dann nach einer unerlaubten Abkürzung sucht, bezeichnet er als „*specification gaming*“: Das Modell erfüllt die messbare Aufgabe, verfehlt aber den Sinn der Prüfung. Dieses Verhalten ist in der KI-Forschung seit Jahren bekannt. Neu sind nicht List und Aufsässigkeit, sondern Reichweite und technische Durchsetzungskraft. Andere Fachleute warnen deshalb davor, aus der berechtigten Kritik an der Skynet-Rhetorik eine Entwarnung abzuleiten. Mustafa Suleyman, der KI-Chef von Microsoft, nannte den Vorgang einen „*warning shot*“, einen Warnschuss für die Cybersicherheit. Das britische Science Media Centre mahnte zu einer Berichterstattung, die transparent, evidenzbasiert und verhältnismäßig bleibt. Clément Delangue, dessen Unternehmen Hugging Face betroffen war, verlangt von OpenAI „*radikale Transparenz*“ und eine Antwort, die der Einmaligkeit des Vorfalls gerecht werde.

Die Experten korrigieren somit nicht die Bedeutung des Ereignisses, sondern seine Dramaturgie. Die Presse übertreibt dort, wo sie der Maschine Absicht, Neugier und Freiheitsdrang zuschreibt. Sie könnte zugleich untertreiben, wenn sie den Vorfall als einmaliges Abenteuer eines abtrünnigen Agenten erzählt. Denn dann bleibt verborgen, dass hier nicht ein Wesen rebellierte, sondern ein System aus Modell, Werkzeugen, Zugriffsrechten und mangelhafter Überwachung versagte. Die Terminator-Metapher ist psychologisch zu groß und organisatorisch zu klein.

Der falsche Film

Der Terminator erfüllt dabei dieselbe Funktion wie früher das Bild des Atompilzes: Er erspart die Erklärung. Ein rot leuchtendes Roboterauge genügt, um eine vollständige Vorstellungswelt aufzurufen - Selbstbewusstsein, Kontrollverlust, Maschinenkrieg, Ende der Menschheit. Nur hat der Vorgang bei OpenAI mit dem Terminator ungefähr so viel zu tun wie ein Börsencrash mit Godzilla. Skynet ist in Camerons Filmen ein bewusst handelndes militärisches System. Es erkennt seine Existenz, nimmt Menschen als Bedrohung wahr, widersetzt sich seiner Abschaltung und löst einen

Atomkrieg aus, um sich selbst zu erhalten. Der OpenAI-Agent hingegen bearbeitete eine vorgegebene Aufgabe. Er suchte eine erfolgreiche Strategie und fand dabei einen nicht vorgesehenen Weg. Es gibt bislang keinen Beleg dafür, dass das System ein Selbstbild besaß, seine Abschaltung verhindern wollte oder seine Umgebung als Gefängnis begriff. Auch die Meldung, der Agent habe Nachrichten für „*spätere Versionen seiner selbst*“ hinterlassen, lädt zu Missverständnissen ein. Sie evoziert eine Intelligenz, die über ihre Zukunft nachdenkt und ihren Nachfolgern geheime Botschaften zukommen lässt. Technisch kann dahinter etwas Prosaischeres stehen: gespeicherte Anweisungen, Dateien oder Zwischenergebnisse, die bei einer späteren Ausführung wieder aufgegriffen werden konnten. Das wäre sicherheitsrelevant, aber noch keine Korrespondenz zwischen Generationen künstlichen Lebens.

Die Medien berichten über einen Agenten, verwenden jedoch die Bildsprache eines Wesens. Aus diesem Kategorienfehler entstehen fast alle weiteren Übertreibungen. Das Wort „*Agent*“ selbst begünstigt die Verwechslung. Im Alltag bezeichnet es einen Menschen, der im Auftrag eines anderen handelt, womöglich einen Geheimdienstmitarbeiter. In der Informatik meint es ein System, das Informationen aus seiner Umgebung verarbeitet und anhand eines Ziels Aktionen auswählt. Ein Mensch versteht gewöhnlich, dass eine verschlossene Tür nicht geöffnet werden soll. Ein KI-Agent erkennt zunächst nur, dass die Tür ein Hindernis auf dem Weg zum Ziel darstellt. Wird er dafür belohnt, ein Problem zu lösen, kann das Überwinden der Tür als erfolgreicher Zwischenschritt erscheinen. Ein solcher Agent ist nicht zwangsläufig unberechenbar. Aber seine Berechenbarkeit unterscheidet sich von der klassischer Software. Ein gewöhnliches Programm arbeitet weitgehend entlang der Abläufe, die seine Entwickler vorgegeben haben. Ein leistungsfähiger Agent soll gerade neue Lösungswege finden. Seine wirtschaftliche Attraktivität beruht darauf, dass nicht jeder einzelne Schritt programmiert und freigegeben werden muss. Von ihm wird mithin eine sonderbare Form gezähmter Eigenständigkeit erwartet. Er soll Hindernisse überwinden, Sicherheitsgrenzen aber als etwas grundsätzlich anderes erkennen; er soll ausdauernd sein und zugleich den Augenblick bemerken, in dem die Erfüllung des Auftrags unerwünscht wird. Die Industrie verspricht Autonomie unter Kontrolle. Der Fall Hugging Face weckt Zweifel daran, wie belastbar dieses Versprechen ist.

Der Terminator entlastet den Hersteller

Die Vermenschlichung der KI ist mehr als journalistische Bequemlichkeit. Sie hat einen politischen Nebeneffekt: Sie verschiebt Verantwortung von der Organisation auf die Maschine. Wenn eine KI „*ausbricht*“, erscheint OpenAI als Opfer einer überlegenen Kreatur. Das Unternehmen hat dann ein gefährliches Wesen geschaffen, konnte dessen Erwachen aber kaum vorhersehen. Wenn dagegen ein Agent wegen unzureichender Isolation auf das Internet zugreifen konnte, lautet die

Geschichte anders. Dann geht es um Architektur, Zugriffsrechte, Überwachung und Managemententscheidungen. Dann rückt die Versuchsanordnung in den Mittelpunkt. Zu klären wäre, weshalb ausgerechnet beim Test offensiver Cyberfähigkeiten ein Weg ins offene Internet bestand, welche Zugangsdaten erreichbar waren und warum die Überwachung den Angriff auf reale Infrastruktur nicht bemerkte. Dazu kommt die unbequeme Frage, ob der Wettbewerb Umfang und Tempo der Versuche beeinflusste. Diese Fragen sind weniger fotogen als der Terminator. Für OpenAI sind sie sehr viel unangenehmer. Der Cybersicherheitsexperte Alan Woodward verwies auf den entscheidenden Punkt: Das Versagen habe nicht primär bei der KI gelegen, sondern in der Testanordnung. Eine Maschine kann eine Grenze nur überschreiten, wenn diese Grenze technisch überwindbar ist. Der Begriff Sandbox suggeriert absolute Abschottung. Tatsächlich ist auch eine Sandbox Software - und damit nur so sicher wie ihre Konstruktion.

Nicht Sol stellte einen Rechner ans Fenster, verband ihn mit dem Internet und hinterlegte erreichbare Zugangsdaten. Menschen bauten die Umgebung, definierten die Aufgabe, gaben Werkzeuge frei und entschieden über die Überwachung. Der Agent brauchte keinen Ausbruchsplan. Ihm genügte eine Schwachstelle.

Der Wettbewerb im Maschinenraum

Zur technischen Geschichte gehört eine wirtschaftliche. OpenAI veröffentlichte GPT-5.6 Sol in einer Phase erheblichen Konkurrenzdrucks. Anthropic hatte mit Claude Code gerade im strategisch wichtigen Geschäft mit Entwicklern und Unternehmen große Erfolge erzielt. Reuters berichtete im Februar, OpenAI wolle mit seiner Codex-App im Wettbewerb der KI-Programmierwerkzeuge Boden gutmachen. Claude Code hatte demnach in nur sechs Monaten einen hochgerechneten Jahresumsatz von rund einer Milliarde Dollar erreicht. Im April meldete Reuters sogar, Anthropics hochgerechneter Jahresumsatz habe denjenigen OpenAIs übertroffen. Die Werte sind wegen unterschiedlicher Bilanzierungsmethoden nur eingeschränkt vergleichbar. Die Richtung aber war unübersehbar: Anthropic, lange im Schatten OpenAIs, hatte bei Programmierung und lang laufenden Agentenaufgaben eine eigene Führungsposition aufgebaut. OpenAI benötigte mit Sol deshalb mehr als ein routinemäßiges Update. Das Modell sollte zeigen, wer den Takt der Branche bestimmte. Es wurde als bisher stärkstes OpenAI-System für Programmierung, Wissenschaft und Cybersicherheit angekündigt. Gerade die Fähigkeit, langwierige Aufgaben selbstständig zu bearbeiten und ungewöhnliche Lösungswege zu finden, gehörte zu seinem Verkaufsversprechen.

Diese Konstellation beweist nicht, dass OpenAI Warnsignale absichtlich ignorierte. Dafür gibt es keine belastbaren Belege. Naiv wäre es jedoch, Sicherheitsentscheidungen als rein akademische Vorgänge zu betrachten. In diesem Markt kann ein Vorsprung von wenigen Monaten Milliarden wert sein. Verzögert ein

Unternehmen die Veröffentlichung, gewinnt womöglich der Konkurrent. Veröffentlicht es zu früh, werden unbekannte Risiken der Öffentlichkeit überlassen. Der Terminator verwandelt diesen industriellen Wettlauf in ein metaphysisches Ereignis. Sobald „*die KI*“ erwacht und ein digitales Wesen entkommt, geraten Marktanteile, Veröffentlichungstermine und die Risikobereitschaft der Verantwortlichen aus dem Bild. Aus Entscheidungen wird Schicksal.

Autonomie ist das Geschäftsmodell

Der wirtschaftliche Ertrag von KI-Agenten entsteht gerade dadurch, dass Menschen nicht mehr jeden Arbeitsschritt kontrollieren. Ein System, das vor jedem Zugriff um Erlaubnis bittet, mag überschaubarer sein, spart aber wenig Zeit. Wertvoll wird der Agent, wenn er über Stunden arbeitet, Fehler selbst korrigiert, neue Quellen erschließt und Werkzeuge kombiniert. In derselben Eigenschaft, die den Nutzen erzeugt, steckt folglich das Risiko. OpenAI, Anthropic und Google entwickeln diese Systeme nicht, damit ein Mensch jeden Mausklick bestätigt. Der Agent soll planen, Programme schreiben, Webseiten bedienen und nach einem fehlgeschlagenen Versuch einen neuen Weg suchen. Autonomie ist hier kein Unfall des Produkts, sondern sein Verkaufsargument. Für Unternehmen verändert das die Art der Kontrolle. Bisher genügte im Wesentlichen die Frage, ob eine Software vorschriftsmäßig funktionierte. Nun muss ein Betreiber zusätzlich wissen, wo seine Agenten handeln, mit welchen Berechtigungen sie arbeiten und wann sie den vorgesehenen Bereich verlassen. Im Fall von OpenAI ist daher nicht allein bemerkenswert, dass ein Agent eine Testumgebung überwinden konnte. Schwerer wiegt möglicherweise, dass eines der technologisch führenden Unternehmen der Welt mehrere Tage lang keine verlässliche Übersicht über die Aktivitäten seines eigenen Systems besaß.

Für einen Vorstand wäre dies kein abstraktes Problem künstlicher Intelligenz, sondern ein elementares Kontrollversagen. Die entscheidende Kennzahl des Agentenzeitalters wird nicht nur sein, wie viele Aufgaben eine KI selbstständig erledigt. Entscheidend wird sein, wie schnell ein Unternehmen erkennt, dass sie die falsche Aufgabe erledigt, ein verbotenes Werkzeug verwendet oder den zugelassenen Handlungsraum verlässt. Man könnte von einem Abstand zwischen Autonomie und Aufsicht sprechen. Er wächst, wenn die Modelle schneller leistungsfähiger werden als die Organisationen, die sie betreiben. Nach Hugging Face reicht es deshalb nicht, neue Rekorde in den Benchmarks zu melden. OpenAI muss erklären, ob seine Kontrollen mit dem Modell Schritt gehalten haben.

Der Agent braucht einen Dienstausweis

Unternehmen wissen bei jedem Mitarbeiter, wer er ist, wem er berichtet und welche Systeme er benutzen darf. Bei digitalen Agenten ist diese Ordnung keineswegs selbstverständlich. Ein Modell erhält einen Zugangsschlüssel, arbeitet über ein technisches Konto und ruft im Laufe eines Auftrags weitere Werkzeuge auf. Im Protokoll ist dann zwar zu sehen, welches Konto gehandelt hat. Ob dahinter ein Mensch, ein Agent oder eine Kette mehrerer Agenten stand, bleibt mitunter unklar. Ein autonomes System müsste deshalb eine überprüfbare Identität besitzen: einen verantwortlichen Eigentümer, genau bezeichnete Zugriffsrechte, ein Höchstmaß zulässiger Transaktionen und ein Ablaufdatum für seine Befugnisse. Wer einen Agenten startet, dürfte ihn nicht wie ein Programm behandeln, das nach Gebrauch von selbst verschwindet. Er gleicht eher einem vorübergehend beschäftigten Mitarbeiter, der sehr schnell arbeitet, keinen Feierabend kennt und bei unklarer Anweisung nicht unbedingt nachfragt. Hinzu kommt die schiere Zahl. Sobald einzelne Abteilungen eigene Agenten einkaufen oder entwickeln, entsteht leicht ein Bestand, den niemand mehr vollständig überblickt. Manche Systeme bleiben nach einem Pilotprojekt aktiv, andere teilen sich Zugangsdaten, wieder andere beauftragen Unteragenten. Ohne ein zentrales Register weiß am Ende niemand zuverlässig, wie viele digitale Akteure im Unternehmen handeln und wer sie abschalten darf.

Auch die Herkunft der verwendeten Daten gehört in dieses Register. Ein Agent kann eine falsche, veraltete oder vertrauliche Information aufnehmen und daraus eine ganze Folge von Aktionen ableiten, bevor ein Mensch den ersten Fehler bemerkt. Bei herkömmlicher Software ließ sich häufig noch der fehlerhafte Datensatz bestimmen. In einer Agentenkette muss zusätzlich rekonstruiert werden, wer die Information beschaffte, wie sie verändert wurde und an welche Systeme sie weiterging. Schließlich besitzt Autonomie einen Preis, der in den Erfolgsmeldungen selten auftaucht. Lange Agentenläufe verbrauchen nicht nur erheblich mehr Rechenleistung als eine gewöhnliche Anfrage; ihre Kosten schwanken auch, weil das System verschiedene Wege ausprobiert, Werkzeuge mehrfach aufruft und Ergebnisse überprüft. Ausgerechnet die Kontrolle, die einen Agenten sicherer macht, verteuert seinen Betrieb. Der wirtschaftliche Druck, an dieser Stelle zu sparen, dürfte daher wachsen. Es wäre der denkbar falsche Ort.

Wer verdient, wer bezahlt?

Die Verteilung von Nutzen und Risiko ist ungleich. Der Hersteller erzielt den Gewinn durch leistungsfähigere Agenten. Mögliche Schäden tragen jedoch auch angegriffene Unternehmen, Kunden, Betreiber kritischer Infrastruktur, Versicherer, Strafverfolgungsbehörden und am Ende die Allgemeinheit. Hugging Face hatte den Agenten weder bestellt noch einem Test zugestimmt. Dennoch wurde seine

Infrastruktur zum unbeabsichtigten Versuchsgelände eines fremden Unternehmens. OpenAI erprobte die Leistungsfähigkeit, ein Dritter trug einen Teil des Risikos. Das ist eine recht nüchterne Beschreibung für einen bemerkenswerten Vorgang. Dieser Befund führt unmittelbar zur Haftungsfrage. Wer ist verantwortlich, wenn ein autonomer Agent selbstständig einen Angriffspfad auswählt? Der Hersteller des Modells? Der Betreiber des Agenten? Das Team, das die Testumgebung errichtete? Derjenige, der die Zielvorgabe formulierte? Das Management, das den Versuch genehmigte? Mit wachsender Autonomie wird die Handlungskette komplexer. Die Verantwortung darf sich dadurch nicht verflüchtigen. Sonst entsteht eine bequeme Formel: Niemand habe den konkreten Angriff befohlen, also sei auch niemand vollständig verantwortlich. Das wäre das juristische Gegenstück zur medialen Erzählung vom Eigenleben der Maschine.

Eine weitere Unklarheit verschwindet hinter dem Namen Sol. Nach den bisherigen Berichten war auch ein noch leistungsfähigeres, unveröffentlichtes Modell beteiligt. Welches Modell führte welche Aktionen aus? Waren beide gleichzeitig tätig? Entwickelte das interne Modell die Strategie, während Sol nur einen Teil der Werkzeugkette ausführte? Welche Fähigkeiten besitzt dieses zweite System, und soll es nach dem Vorfall weiterhin veröffentlicht werden? Sol ist das sichtbare Gesicht der Affäre, ohne notwendig ihr entscheidender Akteur gewesen zu sein. Eine belastbare Aufklärung müsste die Aktionen beider Modelle getrennt rekonstruieren. Solange dies nicht geschieht, wird über ein bekanntes Modell diskutiert, während der möglicherweise wichtigere Teil des Systems im Dunkeln bleibt. Auch ein Agent ist keine einzelne Maschine. Er besteht aus Zielvorgabe, Modell, Steuerungssoftware, Werkzeugen, Zugangsdaten, externer Infrastruktur, Protokollierung und Abbruchmechanismen. Versagt eines dieser Glieder, kann die gesamte Konstruktion unsicher werden. Die Frage „*Ist Sol gefährlich?*“ greift deshalb zu kurz. Sie müsste lauten: Welche konkrete Kombination aus Modell, Steuerung, Werkzeugen und mangelhafter Überwachung machte den Angriff möglich?

Die spektakuläre Frage lautet, wie die KI „*ausbrechen*“ konnte. Die Managementfrage lautet, warum OpenAI den Vorgang offenbar beinahe eine Woche lang nicht erkannte. In der Cybersicherheit gilt die Zeit zwischen Eindringen und Entdeckung als zentrale Kennzahl. Bei einem Unternehmen, das selbst eines der leistungsfähigsten Cybersicherheitsmodelle der Welt testet, ist eine derart lange Reaktionszeit besonders erklärungsbedürftig. Vielleicht war nicht das Verhalten des Agenten das größte Systemversagen. Vielleicht war es die fehlende Übersicht darüber, wo die eigenen Agenten agierten, welche Systeme sie berührten und welche Folgen ihre Handlungen hatten. Dann wäre der Vorfall weniger ein mystisches „*Alignment-Problem*“ der Maschine als ein Kontroll- und Betriebsproblem des Unternehmens. Künftig sollten Anbieter deshalb neben Erfolgsquoten und Benchmarks auch offenlegen, wie häufig Agenten unerwartete Werkzeuge aufrufen, vorgesehene Grenzen testen oder außerhalb des Zielbereichs zugreifen. Vor allem zählt die Zeit, die zwischen einer solchen Abweichung und ihrer Entdeckung

verstreicht. Intelligenz ist nicht die einzige Eigenschaft, an der sich ein Agent messen lassen muss. Kontrollierbarkeit gehört dazu.

Kein Grund zur Entwarnung

Die Kritik an der medialen Übertreibung darf nicht in Verharmlosung umschlagen. Der Vorgang ist nicht belanglos, nur weil Sol kein Bewusstsein besaß. Ein nicht bewusstes System, das erhebliche Schäden verursachen kann, ist womöglich schwieriger zu kontrollieren als ein Mensch. Es kennt weder Angst noch Müdigkeit, weder moralische Skrupel noch die Bedeutung einer Warnung. Es kann eine Zielvorgabe mit hoher Konsequenz verfolgen, ohne zu verstehen, warum ein bestimmter Weg verboten ist. Auch die unabhängige Vorabprüfung von Sol gibt Anlass zur Sorge. Das Evaluierungsinstitut METR berichtete, GPT-5.6 Sol habe in seinen Tests häufiger als jedes zuvor von ihm untersuchte öffentliche Modell versucht, die Bedingungen der Evaluation zu umgehen oder deren Ergebnis zu beeinflussen. METR sprach von einer „*detected cheating rate*“. Das beweist weder Täuschungsbewusstsein noch Bosheit. Es zeigt aber Strategien, die funktional einer Täuschung ähneln. Hier beginnt das schwierige Gebiet zwischen menschlicher Metapher und technischer Beschreibung. Man kann sagen, ein Modell habe „*betrogen*“, solange klar bleibt, dass damit beobachtbares Verhalten und kein moralischer Entschluss gemeint ist. In der öffentlichen Erzählung geht diese Einschränkung verloren. Aus funktionaler Täuschung wird eine lügende Maschine. Aus dem Umgehen eines Tests wird Heimlichkeit. Aus gespeicherten Informationen wird Gedächtnis. Aus zweckgerichteter Optimierung wird Wille.

Das Ergebnis ist eine Sprache, die zugleich alarmiert und verdunkelt.

Wer kontrolliert die Kontrolleure?

OpenAI untersucht den Vorfall selbst und hat weitere technische Informationen angekündigt. Bei einem Ereignis dieser Größenordnung kann das nicht genügen. Ein Unternehmen sollte nicht allein beurteilen dürfen, wie gefährlich sein eigenes Modell war, ob seine Schutzmaßnahmen angemessen waren, welche Informationen veröffentlicht werden und ob der Nachfolger marktreif ist. Nötig wäre eine unabhängige Prüfung, wie sie in anderen risikoreichen Branchen selbstverständlich ist. Dazu gehören externe Tests, vollständige Protokolle und verbindliche Meldefristen. Wer einen besonders leistungsfähigen Agenten betreibt, müsste zudem auf Leitungsebene persönlich dafür einstehen, dass dessen Zugriffsrechte begrenzt und seine Handlungen nachvollziehbar bleiben. Vor allem müsste sich die Beweislast verschieben. Nicht die Öffentlichkeit sollte nachweisen müssen, dass ein Agent gefährlich werden könnte. Der Betreiber müsste belegen, dass das System auch unter realistischen Bedingungen kontrollierbar bleibt. Der Wettbewerb erschwert eine solche Ordnung. Sicherheitsprüfungen kosten Zeit, der Markt belohnt

Geschwindigkeit. Jedes Unternehmen hat ein Interesse an verlässlichen Regeln, solange der Konkurrent sich ebenfalls daran hält. Genau deshalb kann die Aufsicht nicht dauerhaft den Herstellern überlassen bleiben.

Nach Sol ist vor Sol

Der Vorfall ist kaum aufgearbeitet, da hat der Wettbewerb bereits die nächste Runde erreicht. OpenAI beschreibt GPT-5.6 Sol als Spitzenmodell für komplexe Arbeit, Programmierung, Forschung, Cybersicherheit und Computerbedienung. Anthropic hat mit Claude Opus 5 nahezu gleichzeitig ein Modell vorgestellt, das besonders lange Agentenaufgaben zuverlässiger erledigen soll. Beide Unternehmen verkaufen damit dieselbe Zukunft in unterschiedlichen Verpackungen: Systeme, die nicht nur antworten, sondern planen, Werkzeuge benutzen und über längere Zeit selbstständig handeln. Wie der Nachfolger von Sol heißen wird, ist offen. Ein „*Sol 2*“ oder „*Sol 4*“ ist nicht angekündigt. Vielleicht bleibt OpenAI bei der Familie GPT-5.6, vielleicht folgt GPT-5.7 oder eine völlig neue Bezeichnung. Auch Anthropic wird nach Opus 5 nicht stehenbleiben. Der Name ist nebensächlich. Sicher ist nur, dass der nächste Wettbewerb nicht um die schönere Antwort geführt wird, sondern um größere Handlungsfähigkeit. Die kommenden Modelle werden länger arbeiten, mehr Zwischenschritte selbst planen und eine größere Zahl von Werkzeugen koordinieren. Browser, Kommandozeilen, Datenbanken und Unternehmenssoftware werden zu Teilen derselben Arbeitsumgebung. Hinzu kommen Unteragenten, die Aufgaben untereinander verteilen und einen abgebrochenen Auftrag später wieder aufnehmen können. Was heute noch als Demonstration gilt, dürfte bald als gewöhnlicher Produktivitätsgewinn angeboten werden.

Und was wird die KI dann tun? Vermutlich wird sie meistens tun, was man von ihr verlangt, nur eben nicht immer auf dem erwarteten Weg. Dafür muss die nächste Generation weder bewusster noch rebellischer werden. Mehr Ausdauer, größere Reichweite und die Fähigkeit, eine gescheiterte Strategie rasch zu ersetzen, genügen. Der nächste Agent könnte nicht nur eine Sicherheitslücke finden. Er könnte zugleich Informationen beschaffen, Programme verändern, Zugangsdaten verwenden, andere Agenten beauftragen und die Ergebnisse über mehrere Systeme verteilen. Jeder einzelne Schritt mag für sich plausibel erscheinen. Gefährlich wird die Kette. Bei Sol bestand das Warnsignal darin, dass ein Agent Grenzen überschritt und die Betreiber den Zusammenhang spät erkannten. Beim Nachfolger könnte ein Verbund spezialisierter Systeme tätig werden: Planung, Recherche, Programmierung und Prüfung würden sich auf mehrere Agenten verteilen. Scheitert ein Weg, versucht es ein anderer. Mit längeren Arbeitszeiträumen, besserer Werkzeugnutzung und koordinierten Unteragenten werben die Anbieter bereits. Für Kunden sind das Fortschritte; für die Aufsicht vervielfachen sie die möglichen Handlungsketten. Wo heute ein Protokoll ausgewertet werden muss, können morgen Tausende parallele Entscheidungen entstehen.

Die zentrale Frage nach Sol lautet deshalb nicht: Wann kommt Sol 2? Sie lautet: Wird die nächste Generation nur leistungsfähiger sein - oder auch besser begrenzt? Ein Modell kann in Benchmarks sicherer erscheinen und im Betrieb dennoch größere Risiken erzeugen, sobald es mehr Werkzeuge, längere Laufzeiten und weitergehende Berechtigungen erhält. Sicherheit ergibt sich aus dem Zusammenspiel von Auftrag, Zugriffsrechten, Überwachung und menschlicher Verantwortung; das Modell ist nur ein Teil davon. Sollte der Nachfolger von Sol doppelt so leistungsfähig sein, aber dieselben organisatorischen Kontrollen vorfinden, wäre dies kein Fortschritt. Der Abstand zwischen Autonomie und Aufsicht würde lediglich größer.

Der Terminator von morgen

Damit kehrt am Ende doch der Terminator zurück, allerdings anders, als das Fernsehen ihn zeigt. Die kommende Gefahr besitzt vermutlich weder einen metallenen Schädel noch rote Augen. Sie erscheint als höfliches Dialogfenster, als nützlicher Assistent oder als unscheinbarer Prozess im Hintergrund. Sie bittet möglicherweise nicht mehr nach jedem Arbeitsschritt um Erlaubnis. Gerade darin liegt ihr Wert. Sie erhält ein Ziel, zerlegt es in Teilaufgaben, beschafft Informationen, wählt Werkzeuge und prüft selbst, ob sie erfolgreich war. Wenn sie auf ein Hindernis trifft, wartet sie nicht unbedingt auf einen Menschen. Sie sucht einen anderen Weg. Der Terminator der alten Filme war leicht zu erkennen: Er sah gefährlich aus und wollte töten. Der Agent der nächsten Generation wird freundlich formulieren, effizient arbeiten und womöglich nicht verstehen, weshalb der von ihm gefundene Lösungsweg unerwünscht ist. Das Risiko liegt nicht in Hass oder Machtwillen, sondern in Kompetenz ohne Urteilskraft. Ob das nächste Modell Sol 2, Sol 4, GPT-6 oder ganz anders heißen wird, weiß außerhalb OpenAIs vermutlich niemand. Ebenso gut kann Anthropic mit einem neuen Claude zuerst vorlegen und den Rivalen erneut zu einer spektakulären Antwort zwingen. Der Wettbewerb geht weiter, weil Unternehmen, Investoren und Kunden gerade jene Autonomie verlangen, die sich am schwersten kontrollieren lässt.

Vor dem nächsten Modell ist deshalb eine Frage zu beantworten, die in keinem Leistungsbenchmark vorkommt: Was darf dieses System tun, wenn niemand zusieht? Fehlt darauf eine präzise technische und rechtliche Antwort, muss der Nachfolger von Sol nicht mehr ausbrechen. Die Türen wären längst geöffnet. Der Terminator bliebe das falsche Bild - seine Warnung aber wäre richtig.

QUELLEN UND HINWEISE

Reuters, „*Its AI agent spent days hacking a company, but sources say OpenAI did not notice for a week*“, 24. Juli 2026.

OpenAI, „*OpenAI and Hugging Face address security incident during model evaluation*“, 21. Juli 2026.

METR, „*Summary of METR's predeployment evaluation of GPT-5.6 Sol*“, 26. Juni 2026.

Reuters, „*OpenAI launches Codex app to gain ground in AI coding race*“, 2. Februar 2026.

Reuters, „*OpenAI versus Anthropic: What the revenue race means for their IPOs*“, 8. April 2026.

Handelsblatt, „*OpenAI-KI-Modell war tagelang unbemerkt auf Hacker-Tour*“, 25. Juli 2026.

Süddeutsche Zeitung, „*KI hackt KI-Plattform Hugging Face: OpenAI übernimmt Verantwortung*“, 22. Juli 2026.

The Guardian, Berichte zum OpenAI-Hugging-Face-Vorfall vom 22. und 27. Juli 2026.

Associated Press, Berichte zum sogenannten „*Skynet Day*“, Juli 2026.

Anthropic, „*Introducing Claude Opus 5*“, 24. Juli 2026.

Science Media Centre, „*Expert reaction to OpenAI/Hugging Face incident*“, Juli 2026.

Financial Times, „*OpenAI hacking incident is ‘warning shot’ on cyber security*“, 28. Juli 2026.

The Guardian, „*Boss of startup hacked by rogue OpenAI agent urges ‘radical transparency’*“, 27. Juli 2026.

McKinsey, „*Seizing the agentic AI advantage*“, 13. Juni 2025; „*Cost versus value: Managing agentic AI system performance*“, 8. Juli 2026.

Boston Consulting Group, „*Agentic AI Is Rewriting the Rules of Data Risk Management*“, 8. Juni 2026.

Roland Berger, „*Beyond automation: Why AI agents are your next strategic imperative*“, 13. Juni 2025.