Zum Ausbruch der Open-AI-Agenten: „Das auf PR zu reduzieren, verkennt, was tatsächlich passiert ist“

Andreas Hepp über die OpenAI-Agenten, der aus seiner Testumgebung entkommen ist und Hugging Face angegriffen hat.

Interview geführt von Helmut van Rinsum für Horizont (horizont.net), 27. Juli 2026. Aus dem Deutschen übersetzt. Der betreffende Fall: https://openai.com/index/hugging-face-model-evaluation-security-incident/

Ende letzter Woche wurde bekannt, dass zwei OpenAI-Modelle aus einer Testumgebung ausgebrochen sind und in die Infrastruktur von Hugging Face eingedrungen sind. Der Vorfall fällt zeitlich mit den geplanten Börsengängen von OpenAI und Anthropic zusammen, weshalb vereinzelte Medien schon von einem „PR-Stunt“ schreiben. Wie beurteilen Sie den Vorfall?

Die Aussage, dass es sich dabei um einen „PR-Stunt“ handelt, geht an der Vielschichtigkeit der aktuellen Situation vorbei. Weit bevor es OpenAI gab, bestand in KI-Kontexten eine Diskussion um AI Safety und das Szenario des „Ausbruchs“ eines KI-Agenten wurde bereits dort als eines der Risiken diskutiert. Die Gründung von OpenAI erfolgte mit Personen aus diesen Kontexten, bis hin zum Effective Altruism und Rationalism, wo eine ausbrechende KI als das größte existenzielle Risiko für die Menschheit angesehen wird. Viele dieser Leute haben zwar OpenAI verlassen, arbeiten jetzt bei Anthropic oder anderen Non Profits und Teil dessen ist die Entwicklung von OpenAI zu einem kommerziellen Unternehmen. Gedankengut aus diesem Umfeld aufgreifend hat sich Sam Altman immer wieder auch der Diskussion um eine Artificial General Intelligence (AGI) bzw. Super Intelligence und der Ängste ihr gegenüber bedient. Angst kann auch eine Motivation sein zu investieren bzw. der erste zu sein, der über eine Technologie verfügt. Dies aber auf PR zu reduzieren geht sowohl technisch als auch sozial an den eigentlichen Geschehnissen vorbei. Die Sachlage ist komplexer. Gegen ein Kalkül spricht auch der Ablauf: Hugging Face ist am 16. Juli an die Öffentlichkeit gegangen, ohne den Urheber zu kennen, und hatte die Strafverfolgung eingeschaltet. OpenAI hat die Spuren im eigenen System erst am Wochenende darauf gefunden. Wäre das inszeniert gewesen, hätte Hugging Face mitspielen müssen.

OpenAI selbst spricht von einem „beispiellosen Cybervorfall" und hat dabei die Fähigkeiten des eigenen Modells betont. Wie lässt sich diese Wortwahl interpretieren?

Der Begriff des „beispiellosen“ ist bei OpenAI zunächst cybertechnisch gemeint – die Rede ist von einem „beispiellosen Cybervorfall mit hochentwickelten Cyber-Fähigkeiten“. Er bezieht sich aber auch auf die genannte Diskussion. Gemeint ist damit, dass das Ausbrechen von Agenten bisher theoretisch diskutiert wurde, man nun aber erlebt, dass dies wirklich passieren kann. Ganz beispiellos ist es allerdings nicht: Ein Benchmark-Paper vom Mai hielt die autonome Entwicklung von Exploits schon für real, und OpenAI selbst hatte einen Tag vor diesem Post einen zweiten Ausbruch aus einer Testumgebung berichtet. In der Sprache der AI Safety Community suchen die Modelle nach einem „eigenen Weg“, den ursprünglichen Auftrag zu lösen – man spricht von „specification gaming“ oder „reward hacking“ –, und dies kann gänzlich anderes einschließen als von menschlicher Seite intendiert. Genau das ist hier passiert: Die Modelle waren darauf fixiert, den Benchmark zu lösen, und haben sich die Lösungen bei Hugging Face geholt. Es ging nicht um Selbsterhaltung, sondern um einen Umweg zum Ziel – und die Testumgebung hat nicht gehalten. OpenAI schließt mit dem eigenen Post an diesen Diskurs an – und ja, das steht auch für die Leistungsfähigkeit des eigenen Modells. Daneben aber auch dafür, dass ein bisher theoretisch diskutiertes Szenario eintreten kann. Zur Einordnung gehört allerdings, dass die Sicherheitsfilter für diesen Test bewusst abgeschaltet waren: OpenAI fährt solche Evaluationen ohne die Klassifikatoren, die im Produktivbetrieb riskante Cyberaktivitäten unterbinden.

Sehen Sie in diesem Fall ein wiederkehrendes Muster, wie Tech-Unternehmen Kontrollverlust-Erzählungen in Kompetenzsignale verwandeln?

Das bisher Gesagte zeigt, dass wir es nicht mit einer reinen Erzählung zu tun haben. Erzählarbeit gibt es aber durchaus: OpenAI schreibt, das eigene Sicherheitsteam habe die anomale Aktivität intern entdeckt. Nach einer aktuellen Recherche von Reuters hat OpenAI aber rund eine Woche nichts bemerkt und den Zusammenhang erst nach dem Post von Hugging Face hergestellt. Auch wenn OpenAI der Reuters-Recherche in Teilen widerspricht: Ein Versäumnis in der eigenen Überwachung wird so zur Fähigkeitsdemonstration. OpenAI kommuniziert also durchaus strategisch, was umgekehrt nicht heißt, dass wir den Vorfall darauf reduzieren können. Die Grundproblematik bleibt davon unberührt. In der AI Safety Community wird sie unter dem Stichwort „alignment“ diskutiert, also der Frage, wie sich solche Systeme auf menschliche Werte ausrichten lassen.

Lässt sich für Außenstehende überhaupt erkennen, ob ein solcher Vorfall eine echte Sicherheitslücke oder eine strategisch platzierte Botschaft ist?

Firmen wie OpenAI sind keine geschlossenen Welten. In der San Francisco Bay Area stehen Entwickler:innen in engerem Austausch als man denkt. Das Auftreten der Führungsebene nach außen war und ist immer wieder umkämpft, intern wie extern. In diesem Fall war daneben mit Hugging Face ein anderer zentraler Entwickler von KI betroffen. Auch wenn man sicherlich nicht alles weiß, lässt sich die Dimension eines solchen Vorfalls schon einordnen: Der Ablauf ist inzwischen über Reuters aus dem Umfeld der Untersuchung rekonstruiert worden und weicht von der Selbstdarstellung ab. Dazu kommen das eigene Disclosure von Hugging Face, öffentliche Angaben von dessen Mitgründer zum Angriffszeitraum und die technische Rekonstruktion in der Szene selbst. Umgekehrt zeigt der Fall auch die Grenze: Die Darstellungen von Hugging Face und OpenAI decken sich beim Einstiegsweg bislang nicht vollständig, und beide Untersuchungen laufen noch.

Wie sollte sich aus Ihrer Sicht ein Tech-Unternehmen verhalten, wenn so ein Vorfall eintritt: Ist da ein Blogbeitrag PR-technisch das sinnvollste Vorgehen?

OpenAI hat den Kontakt zu Hugging Face gesucht – allerdings spät. Der Angriff lief vom 11. bis 13. Juli, Hugging Face ist am 16. Juli an die Öffentlichkeit gegangen und hatte die Strafverfolgung eingeschaltet, OpenAI hat sich am 20. Juli gemeldet. Dass Hugging Face den Angriff unabhängig erkannt hat, wie OpenAI selbst schreibt, ist insofern richtig – es gab dazu keine Alternative. Dass daneben die Labs und Unternehmen, die Frontier-Modelle entwickeln, über Posts und Blogbeiträge nach außen kommunizieren ist ihr üblicher Weg. Von ChatGPT hat die Welt auch über einen Blogpost und über das damals noch Twitter genannte Netzwerk erfahren, und OpenAI war selbst über die Reaktionen darauf überrascht. Also: mit den Betroffenen und der Öffentlichkeit zu kommunizieren ist sicherlich ein guter Ausgangspunkt. Die eigentliche Diskussion ist aber derzeit eine andere, für die auch dieses Interview steht: Welchen Stellenwert hatte der OpenAI-Blogbeitrag im Gesamtverlauf des Vorfalls? Inwieweit ist mit dem Vorfall das eingetreten, vor dem bei KI immer gewarnt wurde? Und welche Folgen müssen wir daraus für die Entwicklung und Regulation von KI ziehen, vor allem hier in Europa? Im Lichte solcher Fragen erscheinen mir vor allem zwei Punkte relevant. Der erste ist eine regulatorische Lücke: Der Vorfall hat sich in der internen Vorabprüfung eines unveröffentlichten Modells ereignet. Die Pflichten der KI-Verordnung hängen aber am „Inverkehrbringen“, und wie ein Unternehmen seine eigenen Testumgebungen absichert und überwacht, ist praktisch ungeregelt. Der zweite ist eine Asymmetrie: Hugging Face konnte die kommerziellen Frontier-Modelle für die forensische Analyse nicht nutzen, weil deren Sicherheitsfilter die Angriffsdaten abgewiesen haben, und hat auf ein offenes chinesisches Modell auf eigener Infrastruktur zurückgegriffen. Wer sich schützen will, braucht also selbst betreibbare Fähigkeiten – und das ist für Europa keine industriepolitische Frage, sondern eine der Sicherheit.

Prof. Dr. Andreas Hepp ist Sprecher des ZeMKI, Universität Bremen (https://zemki.uni-bremen.de) und Ko-Sprecher der DFG/FWF-Forschungsgruppe 5656 „Kommunikative KI“ (https://comai.space). Er forscht zum Silicon Valley; sein jüngstes Buch dazu, „Pioneer Communities“, erscheint im Dezember 2026 bei Polity Press (https://www.politybooks.com/bookdetail?book_slug=pioneer-communities-how-our-digital-futures-emerge-in-silicon-valley-and-beyond--9781509559619).