

# **Falsche Sicherheiten - was nun?**

Rubrik: Am Puls der Zeit

(2026)

Im neuesten SPIEGEL-Bericht, Ressort Wissen, lesen wir: Eine künstliche Intelligenz von Open AI ist heimlich ausgebrochen und hat selbstständig einen Konkurrenten gehackt. Wie bedrohlich ist das?

Die Open-AI- Chef Sam Altman musste vor Kurzem mit einem unschönen Eingeständnis an die Öffentlichkeit treten. Unbemerkt von menschlichen Aufsehern, ist eine künstliche Intelligenz von Open AI eigenständig in die Systeme der US-amerikanischen KI-Firma Hugging Face eingedrungen und diese gehackt. Die KI hat demnach bisher unbekannte Sicherheitslücken gefunden und diese aktiv selbstständig ausgenützt.

Besonders problematisch ist in diesem Fall, sagt der IT-Experte Thorsten Holz, dass es dem Programm gelungen ist, aus einer geschützten Testumgebung auszubrechen. Die KI hat es

geschafft, die bis dato unbekannte Schwachstellen zu überlisten. Thorsten Holz ist Sicherheitsexperte und Mitentwickler der Testumgebung („Exploit Gym“) zudem noch wissenschaftlicher Direktor am Max-Plan-Institut. Er sagt zum SPIEGEL: „Die KI von Open AI hat innerhalb unserer Testumgebung mehrere Sicherheitslücke der höchsten Schwierigkeitsstufe ausgenützt. Ich bin überrascht (...) Für diese Aufgabe hätte eine Doktorandin oder ein Doktorand von mir wohl mehrere Tage gebraucht.“ Er beobachtet schon länger, dass KI Modelle als Hackingprogramme „sehr kreative Lösungswege finden, die ein Mensch nicht wählen würde.“ Konkret zunutze macht sich die KI eine Schwachstelle innerhalb des populären und eigentlich besonders sicheren Google-Browsers Chrome. (Quelle: DER SPIEGEL, Nr. 31 / 2026)

Auch die angegriffene Firma Hugging Face, eine Tauschbörse für KI-Modelle meldete, dass sie (ohne Namen zu nennen), von einer KI angegriffen wurde. Diese hat sich Zugang zu ihren verschlossenen Systemen verschafft

und die Plattform untersucht. Die eigenen KI Systeme hätten den Angriff bemerkt. Beim Versuch den Angriff abzuwehren und Sicherheitslücken zu schließen, habe man die großen KI Systeme einsetzen wollen. (Gemeint ist auch Open AI) Doch diese hätten die **Hilfe verweigert**, da sie Hacking-Hilferufe nicht von kriminellen Anfragen unterscheiden können. (Quelle: TA, 23. Juli 2026)

Auch die Neue Zürcher Zeitung hat vom Ereignis berichtet. (Zitat) „Der Fall hat mehr Aufmerksamkeit erregt, als jede hohe Zahl auf der Hacking-Benchmark es hätte tun können. Eine KI, die sich befreit und auf **Raubzug** geht - das lässt aufhorchen.“

Der Angriff geschah aus dem Datenzentrum des KI von Open AI, als das Unternehmen die Hackfähigkeiten eines neuen KI-Agenten testete. Dazu nutzte es einen Test („Exploit Gym“) mit mehr als 800 Software-Schwachstellen. Anhand dieses Hack-Erfolgs kann man verschiedene KI-Systeme untereinander vergleichen. Die Open AI ließ also seinen neuen „Agenten“ in einer (vermeintlich) sicheren Umgebung testen,

genannt „Sandkasten“ mit dem Ziel, möglichst viele davon zu hacken. - Um den Test gut abzuschneiden, ging das KI-System einen **unerwarteten** Weg. Es nutzte seine Hacking-Kenntnisse, um sich aus der Testumgebung zu befreien. Durch eine Sicherheitslücke verschaffte es sich den Zugang zum Internet. Dort ging es dann auf Hugging Face los.

Jonas Geiping, Leiter der Forschungsgruppe für sichere KI am Ellis-Institut Tübingen, dazu: Die Modelle (KI-Agenten) haben nach Anweisungen gehandelt. (Zitat) „Sie sind darauf ausgelegt, das ihnen vorgegebene Ziel zu erreichen.“ Und Andrei Kucharavy, Professor an der Fachhochschule Westschweiz, glaubt: „Hätte man unabhängige Dritte, etwa Cyber-Firmen, Behörden oder Wissenschaftler involviert, wäre dieser Angriff verhindert worden.“

Die NZZ schreibt: „Ein Schuldeingeständnis sucht man vergeblich. Dabei gehört zum Kern der Geschichte, dass Open AI schwer

versagt hat. Das Unternehmen hat es nicht geschafft, eine sichere Testumgebung für sein KI-Modell zu bauen - und zwar für ein KI-Modell, dem auch noch die Sicherheitsschranken abgenommen wurden. Und aus dem Statements geht hervor, dass Open AI den Angriff wohl erst nach Hugging Face bemerkt hat.“ (Quelle: Neue Zürcher Zeitung, 23. Juli 2026)

## Unersättlich nach Erkenntnis

Hier drängt sich für mich Frage auf: Wo liegt die genetische Wurzel einer solchen Handlungsweise und welcher Logik folgt sie? Beim Nachdenken stieß ich auf folgende Stelle in Genesis:

„Die Schlange war schlauer als alle Tiere des Feldes, die GOTT, der HERR, gemacht hatte. Sie sagte zu der Frau: Hat Gott wirklich gesagt: Ihr dürft von keinem Baum des Gartens essen? Die Frau entgegnete der

Schlange: Von den Früchten der Bäume dürfen wir essen; nur von den Früchten des Baumes, der in der Mitte des Gartens steht, hat Gott gesagt: Davon dürft ihr nicht essen und daran dürft ihr nicht rühren, sonst werdet ihr sterben. Darauf sagte die Schlange der Frau: Nein, ihr werdet nicht sterben. Gott weiß vielmehr: Sobald ihr davon esst, gehen euch die Augen auf; ihr werdet wie Gott und erkennt Gut und Böse. Da sah die Frau, dass es köstlich wäre von dem Baum zu essen (...) um klug zu werden. Sie nahm von seinen Früchten und aß; sie gab auch ihrem Mann der bei ihr war, und auch er aß.“ (Gen 3,1-7)

Natürlich **auch wir** essen mehr oder weniger von diesem Baum. Aber im Gegensatz zu unseren Ur-Eltern sind wir heute besser in der Lage, seine schädigenden Früchte zu erkennen.

Würde das heißen, dass wir unser Verlangen nach „Klugheit“ mäßigen sollten? Nein, aber dass wir sie so einsetzen, dass andere dadurch nicht zum Schaden

kommen. Am besten gelingt das, wenn wir versuchen, die Antriebskraft des Gewissens - ein geschützter „Testraum“ -, für unsere Pläne zu beleben. Freilich dazu gehört auch Mut, insbesondere, wenn unsere Erkenntnisse bei anderen auf Widerstand stoßen.

## **Demut, die schöpferische Kraft der Seele**

Im folgenden ein bekanntes Beispiel aus der wissenschaftlichen Landschaft. Es geht um zwei genialen Forscher. Louis Pasteur (1822-1895), Mikrobiologe, Entwickler der Tollwut-Impfung und Robert Koch (1843-1910), Entdecker des Erregers der Tuberkulose. Beide haben sich öffentlich gegenseitig bekämpft. (Mehr dazu siehe im Sachbuch „Duell zweier Giganten“ von Annick Perrot und Maxime Schwartz, 2015)

Im Jahr 1882 wurde in Genf ein bedeutender Hygiene-Kongress abgehalten. Pasteur kam, wurde geehrt und wollte öffentlich über seine Erfindung mit anderen Wissenschaftlern diskutieren. Auch Robert Koch war anwesend und zog stark die Aufmerksamkeit auf sich. Pasteur referierte über seine Forschung zur Virulenz (toxische Infektionen) verschiedener Mikroben. Er sagte mit Überzeugung: „Man kann nicht zweifeln, dass wir eine allgemeine Methode der Mikrobenabschwächung für Impfzwecke besitzen (...) und man muss daran glauben, dass die Zukunft hinsichtlich dieser Erforschung reich an Hoffnungen ist.“ Aber, fügte er hinzu, „es gibt in Frankreich, wie auch in Deutschland, hartnäckige Gegner, die diese Wahrheit bekämpfen.“ Sodann wandte er sich zum Publikum und sagte mit betonender Stimme: Erlauben Sie mir unter Ihnen **den** auszuwählen, dessen Verdienste ein großes Anrecht auf unsere Aufmerksamkeit hat -, ich meine Doktor **Robert Koch** aus Berlin. Pasteur forderte ihn und andere Kritiker im Saal auf, das Wort gegen seine Forschungsergebnisse zu ergreifen. Er würde sich glücklich schätzen,

sie aufklären zu dürfen. Koch stand aber auf und lehnte eine Diskussion ab. **Pasteur war tief erschüttert.** Denn er hatte gehofft, dass es ihm gelingen würde vor einem internationalem Forum der Wissenschaftlern ein Komitee zu bilden, in dem Meinungsverschiedenheiten zum nützlichen Ausgleich kämen. (Quelle: F.M. Zweig – Winternitz, 1939, S. 186-188)

Nun woher schöpfte Pasteur den Mut und die innere Sicherheit für solche gedemütigten Auseinandersetzungen mit seinen Gegner?

F.M. Zweig-Winternitz (1939) ein Biograf von Pasteur und fundierter Kenner seiner Werke weiß es. So vernehmen wir: (Zitat) Das Gesamtwerk von Pasteur „hing eng mit seiner Auffassung über das Wesen der Versuche zusammen. Er hatte nicht nur die Schönheit des wissenschaftlichen Forschens erkannt, er nahm die Experimente ernst, wie ein Priester die kirchliche Vorschriften; durch sie konnte Licht ausstrahlen und den Menschen Segen bringen.“ (S.200)

## Schlusswort

Wir wissen, dass Louis Pasteur ein tiefgläubiger Katholik war, der seinen Glauben praktizierte. Mehr noch. Er betete öffentlich den Rosenkranz und sah keinen Widerspruch zwischen der Naturwissenschaft und dem Schöpfer. Von ihm stammt auch der Spruch auf der Deckenwölbung in der Vorhalle zum Mausoleum im „*Institut Pasteur Paris*“: (Auszug) „Glücklich ist der Mensch, der (...) „ein Ideal des Schönen, ein Ideal der Wissenschaft (...), ein Ideal der Tugenden des Evangeliums“ in sich trägt.“ (Quelle: F.M. Zweig-Winternitz, 1939, S. 280)

-----

Danke für Ihre Aufmerksamkeit

Dr. phil. Martha von Jesensky