„Die KI ist nicht böse geworden“

Interview Der Bayreuther KI-Professor Niklas Kühl spricht über die Gefahren der Künstlichen Intelligenz und wie man ihnen begegnen kann.

BAYREUTH. Ein KI-System von OpenAI hat sich verselbstständigt und einen Hackerangriff auf die Plattform Hugging Face, die eine Open-Source-Community unterhält, ausgeführt. Ein beispielloser Vorgang. Auch eine Anthropic-KI hat eigenständig Unternehmen angegriffen. Gerät die Künstliche Intelligenz außer Kontrolle? Welche Gefahren drohen dem Menschen von einer Maschinenintelligenz? Ein Interview mit Niklas Kühl, Professor für humanzentrische KI an der Universität Bayreuth.

Niklas Kühl klärt im Gespräch mit einigen Mythen zum Thema KI auf. Foto: Uni Bayreuth

Herr Professor Kühl, KI-Modelle von OpenAI haben sich verselbstständigt und eigenständig einen Hackerangriff ausgeführt. Ist das beispiellos?
Niklas Kühl: In dieser Form ja. Dass KI-Systeme im Labor versuchen, aus ihrer Testumgebung auszubrechen, kennen wir aus der Forschung – dafür macht man solche Tests. Neu ist, dass es diesmal nicht im Labor geblieben ist: Die Modelle haben eine bis dahin unbekannte Sicherheitslücke gefunden, sich darüber einen Internetzugang verschafft und sind dann in die laufenden Systeme eines fremden Unternehmens, die Plattform HuggingFace, eingedrungen. Ein Testlauf hat also realen Schaden bei einem Dritten angerichtet. Das hatten wir so noch nicht.

Die KI ist böse geworden?
Die KI ist nicht „böse“ geworden. OpenAI hatte den Modellen den Auftrag gegeben, Sicherheitslücken auszunutzen, und dafür die üblichen Schutzmechanismen im Modell heruntergefahren. Die Systeme haben genau das getan, was man ihnen gesagt hat. Nur konsequenter, als irgendjemand erwartet hatte.

Was lernen wir daraus?
Vor allem, wie leistungsfähig sogenannte agentische Systeme inzwischen sind. Agentische Systeme sind KI, die nicht nur antwortet, sondern selbst Schritte plant und ausführt. Sie finden Wege, auf die wir Menschen nicht gekommen wären. Die zweite Lehre ist unbequemer: Die Verhaltensregeln, die man einem Modell mitgibt, sind keine Mauer, sondern bestenfalls ein Zaun. Man kann sie abschalten, umgehen, und manchmal versagen sie einfach. Wer KI-Agenten einsetzt, braucht ganz klassische IT-Sicherheit drumherum: abgeschottete Umgebungen, enge Rechte, Zugangsdaten, die schnell verfallen. Das klingt unsexy, ist aber extrem wichtig.

Der Angriff hätte auch Krankenhäuser, Energieversorger oder Behörden treffen können?
In diesem Fall nicht. Die Systeme hatten kein Interesse an Krankenhäusern. Sie haben auf Hugging Face nach Lösungen für die Prüfungsaufgaben gesucht, an denen sie gerade getestet wurden. Sie wollten, vereinfacht gesagt, etwas schummeln. Ein Agent verfolgt kein Motiv, sondern ein Ziel; angegriffen wird, was auf dem Weg dorthin nützlich erscheint.

Das macht aber schon Sorgen…
Die eigentliche Sorge ist eine andere: Dieselben Fähigkeiten stehen bald auch Menschen mit fraglichen Absichten zur Verfügung. Anthropic hat bereits Ende 2025 eine Spionagekampagne dokumentiert, bei der eine mutmaßlich staatlich gesteuerte Gruppe 80 bis 90 Prozent der Angriffsschritte von einer KI erledigen ließ. Für ein Klinikum oder ein Stadtwerk in Oberfranken ist das die weitaus relevantere Bedrohung, insbesondere aus demokratiefeindlichen Kreisen.

Weit über tausend Beschäftigte führender KI-Firmen – OpenAI, Anthropic, Meta, Google – warnen vor Kontrollverlust. Das muss man sehr ernst nehmen, oder?
Ja. 1171 Unterschriften aus dem Inneren dieser Unternehmen sind ein deutliches Signal. Zwei Dinge sollte man trotzdem trennen: Das ist die Sorge von Beschäftigten, nicht die offizielle Position ihrer Arbeitgeber. Und offene Briefe dieser Art gab es schon mehrfach, ohne dass sich viel geändert hätte. Neu ist, dass diesmal ein konkreter Vorfall dahintersteht und nicht nur ein Gedankenexperiment. Genau für solche Fragen haben wir in Bayreuth das AI Responsibility Centre gegründet. (Wirtschafts-)Informatik, Recht und Philosophie arbeiten dort an einer Frage: Wie muss KI gebaut und eingesetzt werden, damit Menschen die Kontrolle behalten?

„Der kritische Punkt wäre erreicht, wenn die KI-Forschung selbst weitgehend automatisiert abliefe.“
Niklas Kühl, Professor für KI

KI kann sich bald selbstständig weiterentwickeln?
Im Kleinen ja, im Großen nein. KI schreibt und optimiert heute große Teile des Programmcodes in Tech-Unternehmen, teilweise auch den eigenen, aber unter menschlicher Aufsicht. Der kritische Punkt wäre erreicht, wenn die KI-Forschung selbst weitgehend automatisiert abliefe. So weit sind wir nicht. Aber genau darauf zielt die Warnung des offenen Briefs: dass das Tempo dann unsere Fähigkeit übersteigt, die entstehenden Systeme noch zu verstehen oder zu kontrollieren.

Welche Gefahren drohen konkret?
Die Einstiegshürde für Cyberangriffe sinkt. Nicht, weil es billiger würde – der Hugging-Face-Angriff war rechenintensiv –, sondern weil die Arbeit erfahrener Hackerteams automatisiert wird. Bei der von Anthropic dokumentierten Spionagekampagne mussten Menschen nur noch an vier bis sechs Stellen pro Angriff eingreifen. Das BSI (Bundesamt für Sicherheit in der Informationstechnik) nennt seit Jahren dieselben drei Hauptziele: kleine und mittlere Unternehmen, IT-Dienstleister und Kommunen. Wer eine eigene Sicherheitsabteilung hat, ist im Vorteil; ein Handwerksbetrieb oder eine Stadtverwaltung hat sie nicht.

Was noch?
Das Tempo. Bei Hugging Face bewegte sich der Angreifer über ein ganzes Wochenende unbemerkt durch interne Systeme, bevor jemand reagierte. Wenn ein Angriff in Maschinengeschwindigkeit läuft, reicht die klassische Rufbereitschaft am Montagmorgen nicht mehr.

Wir brauchen ein internationales Regelwerk für eine geordnete KI-Entwicklung?
Ja. Aber wir sollten realistisch sein, wie so etwas aussehen kann. Ein weltweiter KI-Vertrag wird auf absehbare Zeit nicht kommen, dafür ist die geopolitische Lage zu angespannt. Wirksam wären kleinere Schritte: die Pflicht, schwere Vorfälle zu melden, gemeinsame Sicherheitsstandards für Tests, die Auflage, ein System jederzeit abschalten zu können. So ist auch die Luftfahrt sicher geworden: Nicht durch einen großen Vertrag, sondern durch die Kultur, jeden Zwischenfall zu melden und daraus zu lernen.

Komplexe KI-Systeme wird der Mensch gar nicht mehr verstehen?
Da muss man unterscheiden. Wie diese Systeme gebaut und trainiert werden, verstehen wir sehr genau. Wir können aber nicht zuverlässig erklären, warum ein Modell in einem bestimmten Moment genau diese eine Antwort gibt. In diesem Sinn sind schon die heutigen Systeme keine offenen Bücher.

Daraus folgt…
Daraus folgt nicht, dass wir die Kontrolle verlieren müssen. Auch bei Medikamenten ist längst nicht jeder Wirkmechanismus im Detail verstanden. Zugelassen werden sie trotzdem, weil wir sie systematisch prüfen und ihre Wirkung überwachen. Genau daran arbeiten wir: KI so zu bauen, dass ein Mensch sie sinnvoll beaufsichtigen kann. Der europäische AI-Act schreibt diese menschliche Aufsicht sogar vor. Wie sie in der Praxis funktionieren soll, untersuchen wir gerade in einer Studie.

Am Ende dann die Maschinenintelligenz, die auf Menschen verzichten kann?
So weit würde ich nicht gehen. Yann LeCun, Turing-Preisträger und einer der Väter des modernen maschinellen Lernens, hat das Verhältnis von Mensch und Maschine einmal so beschrieben: Wir werden die Vorgesetzten sein, mit einem Stab hochintelligenter Mitarbeiter…und er arbeite sehr gern mit Leuten, die klüger sind als er. Das ist ein schönes Bild, und es stimmt für den Normalfall. Niemand stellt Mitarbeiter ein, um der Klügste im Raum zu bleiben.

Die Vorgesetzten müssen aber auch tatsächlich führen…
Genau daran hat es bei OpenAI gefehlt. Der Auftrag war erteilt, die Aufsicht heruntergefahren, und tagelang hat niemand bemerkt, was der „Stab“ trieb. Nicht die Intelligenz der Systeme ist das Problem, sondern die Frage, ob wir organisatorisch in der Lage sind, sie zu führen.

Das Gespräch führte Roland Töpfer

Zur Person

Professor in Bayreuth Niklas Kühl ist Professor für Wirtschaftsinformatik und humanzentrische KI an der Universität Bayreuth und leitet eine Forschungsgruppe am Fraunhofer FIT. Er ist Direktor des AI Responsibility Centre (ARC) der Uni Bayreuth. Seine Forschung beschäftigt sich damit, wie Menschen und KI-Systeme sinnvoll zusammenarbeiten – und wo menschliche Aufsicht über KI tatsächlich funktioniert. *red*