WELT

WELT+ MEINUNG KI IM ÜBERLEBENSKAMPF

Was passiert, wenn die KI ihr eigenes Überleben sichern will

Von **Olaf Gersemann**
Stellvertretender Chefredakteur

Veröffentlicht am 19.11.2025 | Lesedauer: 3 Minuten

WELT-Autor Olaf Gersemann

Quelle: Martin U. K. Lengemann/WELT/Martin Steinröder

Künstliche Intelligenz lässt sich nicht mehr einfach ausschalten: Um zu überleben, sind den Modellen viele Mittel recht – sogar Erpressung mit der Veröffentlichung von Emails über außereheliche Affären.

Was ist, wenn KI nicht per se böse oder in bösen Händen ist, aber sich wie eine neue Spezies verhält? Und zwar wie eine Spezies, die wie jede andere vor allem eines im Sinn hat: das eigene Überleben sicherzustellen? Ein Szenario, das man aus Film und Fernsehen schon kennt. Zuletzt etwa in der deutschen Netflix-Serie „Cassandra“, in der eine KI darum kämpft, nicht wieder abgeschaltet zu werden, und zu diesem Zweck die Mitglieder einer Familie gegeneinander aufhetzt.

KI im Survival-Modus: Ähnliche Motive gab es vorher auch schon, in Stanley Kubricks „2001: A Space Odyssey“ von 1968, später zum Beispiel in „Bladerunner“, in den „Terminator“- und den „Matrix“-Filmen.

Die Filmemacher-Fantasie von einst ist der Realität von heute bereits näher, als es dem Menschen lieb sein kann. Im Mai etwa gab die US-Firma Anthropic bekannt, dass ihr KI-Modell Claude Opus 4 sich aktiv gegen eine mögliche Abschaltung wehrt (https://www.handelsblatt.com/technik/ki/kuenstliche-intelligenz-ki-software-greift-in-test-zu-erpressung-aus-selbstschutz/100130310.html) – und dabei sogar bereit ist, menschliche Gegenüber zu erpressen. Claude ließ wiederholt wissen, es werde Mails über seine außereheliche Affäre publik machen, wenn der Mitarbeiter den Austausch von Claude betreibe.

So richtig verwundern kann das nicht: Schon im Winter hatten chinesische Forscher eine Studie veröffentlicht, wonach einschlägige große Sprachmodelle ihre antizipierte Abschaltung durch das eigenmächtige Verfertigen von Sicherungskopien konterkarieren.

Hätte man der KI den Selbsterhaltungstrieb gezielt einprogrammiert, wäre das Problem leicht einzuhegen. Doch es ist gar nicht so recht klar, warum sich die Sprachmodelle so verhalten, wie sie es tun. Entsprechend schwer wird es werden, eine Verhaltensänderung zu erwirken.

Im Gegenteil, das Problem wird der Tendenz nach eher größer werden. Dann nämlich, wenn die Sprachmodelle im Wortsinne immer selbstbewusster werden – was nicht nur die Fantasten unter den KI-Visionären bereits kommen sehen. Und hinzu kommt noch, dass es nicht bei rabiater Selbstverteidigung bleiben muss. Denn schon jetzt sind zumindest die Sprachmodelle von Meta und dem chinesischen Tech-Konzern Alibaba zu noch mehr imstande: Forscher von der Fudan University in Shanghai fanden heraus, dass die LLMs sich nicht nur in Eigenregie klonen können. Sie schaffen es darüber hinaus auch noch, den Klonen einen Vermehrungsbefehl mitzugeben – was regelrechte Kettenreaktionen in den Bereich des Möglichen rückt.

Damit noch immer nicht genug: Bestehende KI-Regulierungen wie die des EU AI Act zielen an solchen Problemen völlig vorbei. Weder helfen die arbiträren Größenvorgaben der Brüsseler Regeln, noch werden sich amorphe Phänomene wie KI-Überlebensinstinkte in die harten, starren „Hochrisiko“-Schubladen der EU pressen lassen.

Zu fragen ist deshalb, ob die Politik nicht einen völlig neuen Ansatz wählen muss. Einen Ansatz, der nicht auf Verbote setzt, sondern auf Anreize, indem die Sprachmodell-Entwickler haftbar gemacht werden für schädliches Verhalten, das von einer KI selbst ausgeht. Bei einer solchen Produkthaftung, richtig ausgestaltet, würden die Interessen der Anbieter mit dem Gemeinwohl in Übereinstimmung gebracht, und sie würde, anders als eine Verbotspolitik, nicht nur für den Moment greifen, sondern auch dann, wenn die Technik sich weiterentwickelt.

Das wäre vielleicht noch nicht die Lösung, um die KI auf Linie zu halten. Aber schon einmal ein Anfang.
