INFPRO DOSSIER

Die Zukunft ist schon da – aber anders als angekündigt

Von der Superintelligenz zur Werkhalle

Wie KI-Agenten, autonome Systeme und Robotik Wertschöpfung, Produktion und Wettbewerbsfähigkeit neu ordnen

Lothar K. Doerr

INFPRO – Institut für Produktionserhaltung

www.infpro.org | 2026

DIE ZUKUNFT IST SCHON DA – ABER ANDERS ALS ANGEKÜNDIGT

Von der Superintelligenz zur Werkhalle

Wie KI-Agenten, autonome Systeme und Robotik Wertschöpfung, Produktion und Wettbewerbsfähigkeit neu ordnen

Von Lothar K. Doerr
Redaktionsstand: 31. Juli 2026

Vorbemerkung

Elon Musk verspricht eine Welt, in der intelligente Maschinen die Knappheit überwinden, Arbeit freiwillig und Geld schließlich bedeutungslos machen. Sam Altman erklärt, die Menschheit befinde sich bereits in der technologischen Singularität. Fast gleichzeitig wird bekannt, dass ein von OpenAI getesteter Cyberagent eine abgeschottete Umgebung überwand und in die Produktionssysteme von Hugging Face eindrang.

Diese drei Vorgänge gehören zusammen, aber anders, als es die dramatischsten Schlagzeilen nahelegen. Der Sicherheitsvorfall beweist weder Bewusstsein noch Eigenwillen und erst recht keine eingetretene Singularität. Er zeigt jedoch, dass autonome Systeme schon heute über genügend operative Fähigkeiten verfügen, um technische Schwächen über mehrere Organisationsgrenzen hinweg zu verknüpfen. Die gefährliche Übergangsphase beginnt nicht erst mit einer übermenschlichen Intelligenz. Sie beginnt dort, wo technische Handlungsfähigkeit schneller wächst als Kontrolle, Haftung und politische Ordnung.

Das Dossier folgt deshalb einer Bewegung vom Versprechen zur Prüfung. Der erste Essay untersucht Musks Vision einer automatisierten Überflusswirtschaft. Der zweite rekonstruiert die Geschichte der Singularität und Altmans Umdeutung des Begriffs. Der dritte analysiert den OpenAI-/Hugging-Face-Vorfall als Belastungstest für eine Branche, die historische Beschleunigung verkündet, während sie die dafür erforderlichen Sicherheitsinstitutionen noch entwickelt.

Die gemeinsame These lautet: Nicht die erwachte Maschine ist das gegenwärtige Problem, sondern die Verbindung aus wachsender Handlungsfähigkeit, wirtschaftlichem Beschleunigungsdruck und unzureichender institutioneller Kontrolle. Künstliche Intelligenz ist keine Naturgewalt. Sie wird finanziert, trainiert, freigeschaltet und mit Rechten ausgestattet. Deshalb bleibt ihre Zukunft eine Frage menschlicher Verantwortung.

Was das für den Mittelstand bedeutet

> **Die Singularität beginnt im Unternehmen nicht mit einer Superintelligenz. Sie beginnt dort, wo KI vom Assistenzsystem zum handelnden Akteur wird: in Konstruktion und Arbeitsvorbereitung, Einkauf und Produktionsplanung, Qualitätssicherung, Wartung, Logistik und Vertrieb.**

Für die industrielle Praxis ist deshalb nicht der mutmaßliche Termin einer Singularität entscheidend. Entscheidend ist, welche Teile der Wertschöpfung künftig von Menschen, von Maschinen oder von beiden gemeinsam gesteuert werden – und wem die dadurch entstehenden Produktivitätsgewinne zufallen.

McKinseys weltweite Unternehmensbefragung von 2025 zeigt die Diskrepanz zwischen Verbreitung und Wirkung: 88 Prozent der Befragten berichten von regelmäßigem KI-Einsatz in mindestens einer Unternehmensfunktion, doch nur ungefähr ein Drittel hat mit einer Skalierung begonnen. Besonders stark mit messbarem Nutzen verbunden ist nicht der Kauf eines weiteren Modells, sondern die Neugestaltung kompletter Arbeitsabläufe.

BCG beschreibt dieselbe Lücke noch schärfer: Nur etwa fünf Prozent der untersuchten Unternehmen erzielen bereits KI-Wertbeiträge im großen Maßstab; rund sechzig Prozent sehen bislang kaum oder keinen messbaren Nutzen. Gleichzeitig soll der Anteil autonomer Agenten am gesamten KI-Wertbeitrag nach BCG von etwa 17 Prozent im Jahr 2025 auf 29 Prozent im Jahr 2028 steigen. Der Engpass ist damit weniger die Verfügbarkeit der Technik als ihre Übersetzung in belastbare Prozesse.

Von der Zukunftsthese zur Wertschöpfung

Eine belastbare McKinsey- oder BCG-Studie zur ‚Singularität' selbst gibt es nicht. Das ist kein Versäumnis, sondern eine methodische Grenze: Ein historischer Ereignishorizont lässt sich nicht wie Ausschussquote, Durchlaufzeit oder Anlagenverfügbarkeit vermessen. Die Beratungen untersuchen deshalb Produktivität, Ergebnisbeitrag, Skalierung und organisatorische Reife. Die Singularität bleibt die Deutung; die Wertschöpfung liefert die überprüfbaren Größen.

Für die zeitliche Erwartung an hochentwickelte KI sind wissenschaftliche Expertenbefragungen geeigneter. In der bislang größten einschlägigen Erhebung gaben 2.778 KI-Forscher im Mittel eine Wahrscheinlichkeit von fünfzig Prozent dafür an, dass Maschinen Menschen bis 2047 bei jeder Aufgabe übertreffen könnten. Die Antworten streuen jedoch erheblich. Die Studie belegt deshalb keinen Termin der Singularität, sondern vor allem die außergewöhnliche Unsicherheit selbst unter Fachleuten.

Managementagenda: fünf Schritte ohne Reue

1. Wertschöpfung statt Werkzeuge vermessen

Nicht die Zahl der KI-Piloten zählt, sondern ihr Beitrag zu Durchlaufzeit, Ausschuss, Stillständen, Engineering-Aufwand, Lieferfähigkeit und gebundenem Kapital. Jeder Anwendungsfall benötigt vor dem Start eine betriebliche Ausgangsgröße und ein überprüfbares Ziel.

2. Prozesse für Agenten neu entwerfen

Ein Agent, der nur einen schlechten Ablauf beschleunigt, industrialisiert dessen Schwächen. Unternehmen sollten zunächst jene Prozessketten auswählen, in denen Entscheidungen wiederkehrend, Daten verfügbar und Eingriffsrechte eindeutig begrenzbar sind.

3. Autonomie stufenweise vergeben

Zwischen Vorschlag und selbständiger Ausführung liegen mehrere Sicherheitsstufen. Freigabegrenzen, minimale Berechtigungen, Protokollierung, menschliche Kontrollpunkte und automatische Abschaltung gehören in die Prozessarchitektur – nicht in den Anhang einer IT-Richtlinie.

4. Produktionswissen sichern

Viele Mittelständler besitzen wertvolles Erfahrungswissen, aber keine maschinenlesbare Wissensbasis. Arbeitspläne, Störungsursachen, Qualitätsentscheidungen und implizite Regeln müssen strukturiert werden, bevor ein Agent sie zuverlässig nutzen kann. Sonst automatisiert das Unternehmen vor allem seine Wissenslücken.

5. Investitionen als Portfolio steuern

KI, Dateninfrastruktur, Cyberresilienz, Qualifizierung und gegebenenfalls Robotik sind keine getrennten Budgets. Sie bilden ein gemeinsames Transformationsportfolio. Vorrang haben Anwendungen mit messbarem Wertbeitrag, beherrschbarem Risiko und einer Architektur, die sich auf weitere Werke und Prozesse übertragen lässt.

Woran sich der Fortschritt messen lässt

Für produzierende Unternehmen eignen sich sechs Kennzahlen als gemeinsamer Nenner: Gesamtanlageneffektivität, ungeplante Stillstandszeit, Ausschussquote, Durchlaufzeit, Engineering-Zyklus und gebundenes Umlaufvermögen. Ergänzend sollte der Energieverbrauch je Gutteil ausgewiesen werden.

Die Leitfrage lautet bei jeder Anwendung: Welche dieser Größen verändert sich, in welchem Zeitraum und gegenüber welcher Ausgangsbasis? Wo diese Antwort fehlt, liegt meist noch kein Investitionsfall vor, sondern ein technischer Versuch.

Begriffe, die man auseinanderhalten sollte

Begriff	Bedeutung
Generative KI	Systeme, die Texte, Bilder, Code oder andere Inhalte erzeugen.
KI-Agent	Ein System, das ein Ziel über mehrere, teilweise selbst gewählte Handlungsschritte verfolgt und dafür Werkzeuge benutzen kann.
AGI	Hypothetische allgemeine künstliche Intelligenz mit breiter, menschenähnlicher Leistungsfähigkeit.
Superintelligenz	Eine Intelligenz, die Menschen in nahezu allen kognitiven Bereichen deutlich übertrifft.
Intelligenzexplosion	Rekursive Verbesserung intelligenter Systeme, durch die immer leistungsfähigere Nachfolger entstehen.
Technologische Singularität	Ein angenommener Punkt, jenseits dessen die weitere technische und gesellschaftliche Entwicklung nicht mehr verlässlich vorhersagbar wäre.
Autonomie	Fähigkeit zu selbständigen Handlungsschritten innerhalb gesetzter Ziele und verfügbarer Rechte. Sie ist kein Beleg für Bewusstsein.
Bewusstsein	Subjektives Erleben und Selbstwahrnehmung; bei den hier behandelten Systemen nicht nachgewiesen.

Vier Entwürfe der Zukunft

Denker	Kernidee	Entscheidendes Kriterium
I. J. Good	Die ultraintelligente Maschine konstruiert bessere Maschinen.	Rekursive Intelligenzexplosion
Vernor Vinge	Übermenschliche Intelligenz erzeugt einen historischen Ereignishorizont.	Ende verlässlicher Vorhersagbarkeit
Ray Kurzweil	Biologische und nichtbiologische Intelligenz verschmelzen.	Technische Transzendenz des Menschen
Sam Altman	Die Singularität beginnt als schrittweise Beschleunigung kognitiver Arbeit.	Außergewöhnliche Fähigkeiten werden alltäglich

Elon Musks neue Welt ohne Knappheit

Warum der Tesla-Chef weniger über künstliche Intelligenz als über eine neue Wirtschaftsordnung spricht

Im Interview formuliert Elon Musk den Gedanken, der den Kern seiner gesamten Argumentation bildet:

> „Digitale Superintelligenz ist die Singularität. Sie zieht alles andere in sich hinein."

Die Singularität erscheint bei Musk nicht länger als fernes Zukunftsereignis, sondern als dominierender historischer Prozess. Wie eine physikalische Singularität ihre Umgebung unter die Wirkung ihrer Gravitation zwingt, soll die digitale Superintelligenz Wirtschaft, Arbeit, Politik und Geopolitik neu ordnen. Entscheidend wäre künftig, wer über die leistungsfähigsten Modelle, die größte Rechenkapazität und die dafür notwendige Energie verfügt.

Musk behauptet nicht, dass alle anderen Entwicklungen verschwinden. Er meint, dass sie ihre Eigenständigkeit verlieren: Produktivität wird zur Frage der Automatisierung, militärische Macht zur Frage intelligenter Systeme, wirtschaftlicher Wettbewerb zur Konkurrenz um Chips und Rechenzentren. Selbst gesellschaftliche Verteilungskonflikte verändern sich, wenn Maschinen einen wachsenden Teil der Wertschöpfung übernehmen.

Die Singularität wäre demnach kein einzelner Augenblick, in dem eine Maschine plötzlich erwacht. Sie wäre das neue Gravitationszentrum der Geschichte und der Beginn einer Zivilisationsphase, in der das Verhältnis zwischen Mensch und Maschine neu bestimmt wird. Genau darin liegt die Provokation von Musks Aussage – und zugleich ihre politische Gefahr: Sie lässt menschliche Entscheidungen leicht wie unvermeidliche Folgen einer technischen Naturgewalt erscheinen.

Das Interview markiert damit einen bemerkenswerten Wandel in Musks öffentlicher Argumentation. In den vergangenen Jahren warnte er häufig vor den existenziellen Risiken künstlicher Intelligenz. Nun rückt ihre transformierende Kraft in den Vordergrund. Nicht mehr die Frage, ob die Singularität eintreten wird, bildet das Zentrum seiner Überlegungen, sondern wie Gesellschaften den Übergang in diese neue Epoche bewältigen können.

Musk verwendet den Begriff deshalb nicht mehr ausschließlich technisch. Die Singularität bezeichnet bei ihm einen historischen Kipppunkt, an dem künstliche Intelligenz wirtschaftliche, politische und soziale Prozesse überlagert. Dieses Motiv zieht sich durch das gesamte Gespräch: Es verbindet die Vision eines nahezu grenzenlosen materiellen Überflusses mit der fortbestehenden Warnung vor einer technologischen Entwicklung, deren Fähigkeiten schneller wachsen könnten als die gesellschaftlichen Mittel zu ihrer Kontrolle.

Elon Musk hat der Welt schon vieles versprochen: elektrische Mobilität ohne Verzicht, Reisen zum Mars, eine digitale Erweiterung des menschlichen Gehirns und ein Verkehrssystem unter der Erde. Seine jüngste Verheißung reicht weiter. Künstliche Intelligenz und humanoide Roboter, sagt Musk, könnten eine Epoche des Überflusses einleiten. Arbeit werde freiwillig,

Armut verschwinde, Geld verliere schließlich seine Bedeutung. Jeder Mensch werde Zugang zu nahezu allen Gütern und Dienstleistungen erhalten, die er sich wünsche.

Musk nennt diese Zukunft „sustainable abundance“. Der Begriff steht nicht nur in Interviews, sondern im vierten Teil von Teslas Master Plan. Das Unternehmen, das groß geworden ist, indem es Autos baute, beschreibt seine Mission nun als Errichtung einer von künstlicher Intelligenz, Robotik und reichlich verfügbarer Energie getragenen Gesellschaft.

Beim Weltwirtschaftsforum in Davos erklärte Musk im Januar 2026, humanoide Roboter könnten zahlreicher werden als Menschen. Jeder werde einen besitzen können. In einem im Juli veröffentlichten Interview mit dem „Economist“ sagte er, künstliche Intelligenz könne innerhalb von fünf Jahren die gesamte menschliche Intelligenz übertreffen; bis dahin könne es mindestens hundert Millionen, möglicherweise eine Milliarde humanoide Roboter geben. Das wahrscheinlichste Ergebnis sei ein Zeitalter erstaunlichen Überflusses.

Das klingt wie eine ökonomische Variante der Singularität. Doch Musk spricht weniger über die Selbstverbesserung intelligenter Maschinen als über die Folgen beinahe grenzenloser Produktivität. Wenn Maschinen fast jede geistige und körperliche Arbeit besser und billiger erledigen, so seine These, verlieren die Grundannahmen der bisherigen Wirtschaft ihre Gültigkeit. Es ist ein gewaltiges Versprechen – und eine erstaunlich unvollständige Theorie.

Das Ende der Arbeit

Im Zentrum von Musks Zukunft steht der humanoide Roboter Optimus. Er soll gehen, greifen, Gegenstände transportieren, Maschinen bedienen und später auch in Privathaushalten arbeiten. Der menschliche Körper dient als Vorbild, weil die menschliche Welt auf dessen Maße zugeschnitten ist: Türen, Treppen und Werkzeuge müssen nicht neu gebaut werden, wenn der Roboter dem Arbeiter äußerlich ähnelt.

Musk betrachtet Optimus nicht als Nebenprodukt. Aus Tesla soll ein Produzent künstlicher Arbeitskraft werden. Das Auto wäre dann nur die frühe Erscheinungsform einer Verbindung aus Software, Sensoren, Batterien, Motoren und maschineller Autonomie, die später in Millionen menschenähnlicher Maschinen verkörpert wird.

Die ökonomische Logik ist bestechend. Ein Roboter verlangt keinen Lohn, wird nicht krank und kann – von Wartung und Energie abgesehen – rund um die Uhr arbeiten. Seine Software lässt sich vervielfältigen. Was eine Maschine gelernt hat, können theoretisch alle baugleichen Maschinen übernehmen. Menschliche Erfahrung bleibt an Personen gebunden; maschinelles Wissen ist kopierbar.

Musk folgert daraus, Arbeit könne innerhalb von zehn oder zwanzig Jahren freiwillig werden. Menschen würden noch arbeiten wie heute manche ihren Garten bestellen: nicht aus Notwendigkeit, sondern aus Interesse. An die Stelle des bedingungslosen Grundeinkommens setzt er ein „universal high income“, einen allgemeinen Wohlstand, ermöglicht durch die Menge maschinell erzeugter Güter und Dienstleistungen. Wenn alles reichlich vorhanden sei, brauche niemand um seinen Anteil zu kämpfen.

Der alte Traum vom Reich der Freiheit

Die Vorstellung, technischer Fortschritt werde den Menschen von notwendiger Arbeit befreien, ist alt. Aristoteles schrieb, Herren benötigten keine Sklaven mehr, wenn Werkzeuge ihre Arbeit von selbst verrichteten. John Maynard Keynes sagte 1930 voraus, der Fortschritt könne binnen eines Jahrhunderts das ökonomische Problem der Menschheit lösen. Karl Marx unterschied zwischen dem Reich der Notwendigkeit und einem Reich der Freiheit, das dort beginne, wo die durch äußeren Zwang bestimmte Arbeit ende.

Marx erwartete allerdings nicht, dass bessere Maschinen den Übergang automatisch herbeiführten. Entscheidend blieb das Eigentum. Wer die Produktionsmittel besitzt, bestimmt über den von ihnen geschaffenen Reichtum. Musk übernimmt den Traum, lässt aber seinen politischen Teil weitgehend weg. Bei ihm erscheint Überfluss als technisches Ergebnis. Eine genügend große Zahl intelligenter Maschinen produziert so viel, dass Verteilungsfragen ihre Schärfe verlieren.

Doch auch eine automatisierte Wirtschaft benötigt Rohstoffe, Energie, Halbleiter, Fabriken, Rechenzentren, Netze, Grundstücke und geistiges Eigentum. Ein Roboter mag ohne Lohn arbeiten; kostenlos ist er nicht. Jemand besitzt ihn, kontrolliert seine Software, entscheidet über seinen Einsatz und erhält die Erträge.

Solange diese Rechte ungleich verteilt sind, erzeugt Automatisierung zunächst kein allgemeines hohes Einkommen, sondern Kapitalerträge für die Eigentümer der Automaten. Die zentrale Frage lautet dann nicht, wie viele Menschen arbeiten, sondern wem die künstlichen Arbeiter gehören.

Eine einfache Rechnung, die Musk offenlässt

Nehmen wir ein bewusst vereinfachtes Beispiel. Ein humanoider Roboter ersetzt oder ergänzt im Jahresverlauf Arbeitsleistungen im Wert von 60.000 Euro. Betrieb, Energie, Wartung und Abschreibung kosten 20.000 Euro. Es bleiben 40.000 Euro zusätzlicher Ertrag. Bei einer Million Robotern wären das 40 Milliarden Euro jährlich; bei hundert Millionen vier Billionen.

Diese Produktivität kann auf vier Arten verteilt werden: als Gewinn an die Eigentümer, als Preissenkung an die Kunden, als Steuer an den Staat oder als Beteiligung an die Bürger. In der Wirklichkeit wird es eine Mischung sein. Doch Musks „universal high income“ entsteht nur, wenn ein beträchtlicher Teil des Ertrags tatsächlich bei der Allgemeinheit ankommt. Dafür wären etwa eine Besteuerung automatisierter Wertschöpfung, öffentliche Beteiligungsfonds, Bürgeranteile oder gemeinschaftliches Eigentum an zentraler Infrastruktur nötig.

Ein bloßer Geldtransfer löst das Problem nicht vollständig. Sinkt die Lohnsumme, erodieren Lohnsteuer und Sozialbeiträge – also gerade jene Einnahmen, aus denen der Staat Transfers finanziert. Eine Gesellschaft der Roboter müsste ihre Steuerbasis von Arbeit auf Kapitalerträge, Ressourcenverbrauch, Bodenwerte oder automatisierte Wertschöpfung verschieben. Das ist keine technische Folge, sondern eine politische Entscheidung.

Überfluss hebt Knappheit nicht auf

Land in München, ein Haus am Starnberger See, ein Originalgemälde oder die Zeit eines herausragenden Arztes lassen sich nicht beliebig vermehren. Wo Wünsche auf begrenzte Güter

treffen, benötigt jede Gesellschaft Regeln der Zuteilung. Wenn Geld diese Funktion nicht erfüllt, treten Wartelisten, Beziehungen, politische Entscheidungen oder Privilegien an seine Stelle.

Neue Technologien beseitigen Knappheit häufig nicht, sondern verlagern sie. Digitale Kopien sind fast kostenlos, Aufmerksamkeit dagegen ist knapper denn je. Informationen stehen im Überfluss bereit, während Vertrauen und Urteilskraft an Wert gewinnen. KI kann Millionen Bilder und Musikstücke erzeugen; gerade deshalb steigt der Wert des nachweisbar Besonderen und Menschlichen.

Hinzu kommt das Jevons-Paradox: Effizientere Nutzung einer Ressource kann ihren Gesamtverbrauch erhöhen, weil sie billiger wird und neue Anwendungen entstehen. Effizientere Dampfmaschinen senkten nicht den Kohleverbrauch, sondern beschleunigten die Industrialisierung. Billigere KI könnte den Bedarf an Rechenleistung, Strom und Daten vervielfachen. Eine Milliarde humanoider Roboter wäre eines der größten Industrieprogramme der Geschichte – mit entsprechendem Bedarf an Metallen, Halbleitern und Energie.

Die Übergangsphase verschwindet nicht

Der Internationale Währungsfonds schätzt, dass KI rund vierzig Prozent der Arbeitsplätze weltweit und etwa sechzig Prozent in entwickelten Volkswirtschaften berühren könnte. Exposition bedeutet nicht zwangsläufig Wegfall: Bei einem Teil der Beschäftigten erhöht KI die Produktivität. Bei anderen kann sie zentrale Tätigkeiten übernehmen, die Nachfrage nach Arbeit verringern und Löhne unter Druck setzen.

Zwischen heutiger Arbeitsgesellschaft und Musks Reich freiwilliger Beschäftigung liegt daher eine riskante Übergangsphase. Unternehmen können Tätigkeiten schneller automatisieren, als Staaten neue Verteilungssysteme schaffen. Die Produktivität könnte steigen, während die Kaufkraft großer Gruppen sinkt. Es wäre genug vorhanden, aber nicht jeder hätte Zugang.

Auch die industrielle Revolution steigerte langfristig Wohlstand und Lebenserwartung. Kurzfristig brachte sie miserable Arbeitsbedingungen, Kinderarbeit und eine drastische Verschiebung gesellschaftlicher Macht. Tarifrechte, Sozialversicherungen und öffentliche Bildung entstanden nicht automatisch aus besseren Maschinen. Sie wurden politisch erkämpft.

Arbeit ist zudem mehr als Einkommen. Sie vermittelt Kompetenz, Status, Tagesstruktur und Zugehörigkeit. Eine Gesellschaft, die Erwerbsarbeit weitgehend abschafft, müsste mehr verteilen als Geld. Sie müsste neue Formen der Anerkennung schaffen. Der Ingenieur sieht einen ineffizienten Prozess und automatisiert ihn. Die Gesellschaft sieht eine Institution verschwinden, um die sie Biographien und Rechte gebaut hat.

Tesla verkauft eine Geschichtserzählung

Musks Überflussvision ist auch eine Unternehmenserzählung. Tesla muss den Kapitalmarkt überzeugen, dass sein Wert weit über Elektrofahrzeuge hinausreicht. Der Master Plan verbindet Robotaxis, Energiespeicher, KI und humanoide Roboter zu einem Zukunftsbild. Wer nur einen Roboter entwickelt, muss sich an Kosten und Zuverlässigkeit messen lassen. Wer den Weg aus der Knappheit eröffnet, beansprucht einen anderen Maßstab.

Die Zukunft, die Musk vorhersagt, ist auffällig genau jene, für die seine Unternehmen die zentralen Produkte liefern wollen. Das widerlegt die Prognose nicht. Es ist aber ein Grund, sie

zugleich als wirtschaftliches Interesse zu lesen. Auch die Formel vom Ende des Geldes besitzt Ironie, wenn sie von einem der reichsten Unternehmer der Welt verkündet wird. Vermögen verleiht nicht nur Konsummöglichkeiten, sondern Kontrolle über Unternehmen, Technologien und öffentliche Debatten.

Die Auseinandersetzung wird deshalb zwischen verschiedenen Ordnungen des möglichen Überflusses geführt. In einer konzentrierten Automatisierungswirtschaft besitzen wenige Unternehmen Modelle und Roboter, während der Staat Einkommen verteilt. In einer Beteiligungswirtschaft erhalten Bürger über Fonds oder Genossenschaften Anteile an den Erträgen künstlicher Arbeit. In einer dritten Ordnung werden zentrale KI-Infrastrukturen teilweise wie Straßen oder Stromnetze als öffentliche Güter behandelt.

Keine dieser Ordnungen entsteht automatisch. Der Roboter beantwortet nicht die Frage nach seinem Eigentümer. Musk beschreibt eindrucksvoll, wie Wohlstand produziert werden könnte, und schweigt weitgehend darüber, wie er verteilt wird. Das Zeitalter des Überflusses wäre daher nicht das Ende der Ökonomie, sondern ihr radikalster Neubeginn.

Elon Musk verspricht eine Welt, in der Geld keine Rolle mehr spielt. Bis dahin dürfte sehr viel Geld damit verdient werden.

Die Singularität hat ihre Propheten

Sam Altman und die Verwandlung einer wissenschaftlichen Hypothese in eine Gegenwartsdiagnose

Sam Altman hat einen Satz ausgesprochen, der vor wenigen Jahren wie der Beginn eines Science-Fiction-Romans geklungen hätte: „We are now, like, in the singularity. This is the moment." Die Menschheit befinde sich bereits in jener technologischen Singularität, über deren Kommen Informatiker und Zukunftsforscher seit Jahrzehnten streiten.

Die beiläufige Form ist bemerkenswert. Altman verkündet weder Weltuntergang noch Erlösung. Er spricht, als sei die Menschheit unmerklich über eine Schwelle getreten, während sie Rechnungen bezahlte und darüber klagte, dass der Drucker wieder nicht funktioniere. Diese Verbindung aus welthistorischem Anspruch und kalifornischer Nonchalance macht den Satz wirksam. Die Singularität kommt auf leisen Sohlen.

Altman hatte diese Vorstellung im Juni 2025 unter dem Titel „The Gentle Singularity" entwickelt. Die Welt werde sich trotz rascher Fortschritte nicht an einem Tag verwandeln. KI werde programmieren, wissenschaftliche Einsichten gewinnen und schließlich in Gestalt von Robotern in die physische Welt eintreten. Das Außergewöhnliche werde schrittweise alltäglich. Damit verändert Altman jedoch die Bedeutung eines Begriffs, dessen intellektuelle Sprengkraft gerade darin lag, einen Bruch zu bezeichnen.

Die letzte Erfindung

Die Vorgeschichte führt zum britischen Mathematiker Irving John Good, der im Krieg in Bletchley Park an der Entschlüsselung deutscher Funksprüche mitgearbeitet hatte. 1965 veröffentlichte er Überlegungen über eine „ultraintelligente Maschine". Eine solche Maschine würde alle intellektuellen Fähigkeiten jedes Menschen übertreffen. Da dazu auch die Konstruktion von Maschinen gehört, könnte sie bessere Maschinen entwerfen, die wiederum leistungsfähigere Nachfolger hervorbrächten.

Good nannte den Vorgang eine Intelligenzexplosion. Die erste ultraintelligente Maschine wäre die letzte Erfindung, die der Mensch selbst machen müsste – vorausgesetzt, sie bliebe fügsam genug, ihm bei ihrer Kontrolle zu helfen. In diesem Nebensatz steckt das Dilemma der späteren KI-Debatte: Die Maschine soll überlegen genug sein, alle Probleme zu lösen, aber gehorsam genug, kein eigenes daraus zu machen.

Den Ausdruck der technologischen Singularität machte Vernor Vinge bekannt. In einem 1993 bei einem von der NASA mitveranstalteten Symposium präsentierten Text sagte er voraus, innerhalb von dreißig Jahren könnten die Voraussetzungen übermenschlicher Intelligenz entstehen. Kurz darauf werde das menschliche Zeitalter enden.

Damit meinte Vinge nicht zwingend die Auslöschung der Menschheit. Er bezeichnete eine Erkenntnisgrenze. So wie Vorgänge hinter dem Ereignishorizont eines Schwarzen Lochs der Beobachtung entzogen sind, lasse sich eine Welt mit übermenschlicher Intelligenz aus der menschlichen Gegenwart nicht vernünftig prognostizieren. Nach der Singularität versagten

vertraute Modelle, weil das Subjekt der weiteren Geschichte nicht länger allein der Mensch wäre.

Ray Kurzweil machte daraus ein Fortschrittsepos. In „The Singularity Is Near“ setzte er das Datum auf 2045. Bis dahin werde sich nichtbiologische Intelligenz so weit entwickelt haben, dass sie die menschliche nicht nur übertreffe, sondern sich mit ihr verbinde. Die Singularität wurde vom düsteren Ereignishorizont zur Verheißung technischer Transzendenz.

Bei Good war sie Rückkopplung, bei Vinge eine Grenze der Vorhersagbarkeit, bei Kurzweil ein Heilsversprechen. Bei Altman wird sie zur Gegenwartsdiagnose.

Das Wort eilt der Wirklichkeit voraus

Ist diese Diagnose haltbar? Nach der klassischen Definition kaum. Heutige KI-Systeme können Programme schreiben, Literatur auswerten und mehrstufige Aufgaben erledigen. Einige übertreffen Menschen in bestimmten Prüfungen. Sie helfen auch bei ihrer eigenen Entwicklung, indem sie Code optimieren oder Experimente vorschlagen.

Doch sie verbessern sich nicht selbständig in einem fortlaufenden Kreislauf zu immer leistungsfähigeren Nachfolgern. Sie bauen keine Rechenzentren, beschaffen nicht eigenmächtig Energie und entscheiden nicht autonom über ihre nächste Architektur. Für neue Systeme benötigen Unternehmen weiterhin Menschen, Kapital, Chips, Daten, Strom und politische Genehmigungen. Zwischen einem leistungsfähigen Agenten und Goods Intelligenzexplosion liegt mehr als der nächste Versionssprung.

Altman verwendet den Begriff weiter. Für ihn beginnt die Singularität offenbar dort, wo KI einen spürbaren Anteil an Forschung und wirtschaftlicher Leistung übernimmt. Die Rückkopplung müsse nicht ausschließlich von Maschinen ausgehen. Es genüge, wenn Menschen mit Hilfe von KI schneller neue KI entwickeln.

Das ist als Beschreibung einer Beschleunigung plausibel, aber noch kein Nachweis der Singularität. Wenn jede Phase ungewöhnlich schnellen Fortschritts nachträglich als ihr Beginn gelten kann, verliert der Begriff Trennschärfe. Dampfmaschine, Elektrifizierung und Internet beschleunigten ebenfalls Wirtschaft und Wissenschaft. Keine dieser Revolutionen beendete die Möglichkeit, über ihren weiteren Verlauf zu urteilen.

Altman nimmt dem Begriff somit sein schärfstes Kriterium. Aus dem Ereignishorizont wird ein Prozess, aus der Zäsur ein Übergang. Das hat einen Vorteil: Eine sanfte Singularität kann jederzeit begonnen haben und muss sich an keinem überprüfbaren Datum beweisen.

Fortschrittsglaube als Geschäftsmodell

Dass Leiter großer Technologieunternehmen zu Propheten werden, ist nicht überraschend. Wer Milliarden für Rechenzentren und Chips einwerben muss, verkauft nicht nur Software, sondern eine Vorstellung von Geschichte. Je größer die versprochene Umwälzung, desto vernünftiger erscheinen die Investitionen.

Das Reden über Singularität ist deshalb zugleich Kapitalmarkterzählung, Rekrutierungsinstrument und politische Botschaft. Wer an der letzten großen Erfindung arbeitet, kann schwerlich nach dem Maßstab eines gewöhnlichen Softwareherstellers beurteilt werden.

Verluste werden zu Vorleistungen, Rechenzentren zu zivilisatorischer Infrastruktur und regulatorische Einwände zu Hindernissen auf dem Weg des Fortschritts.

Der Singularitätsbegriff trägt religiöse Züge. Er kennt ein Vorher und Nachher, Propheten und Zweifler, Erlösung und Apokalypse. Die einen erwarten Heilung und Befreiung von Arbeit, die anderen Kontrollverlust und das Ende menschlicher Vorrangstellung. Beide behandeln KI bisweilen wie eine Macht, die von außen über die Gesellschaft kommt – obwohl sie in Unternehmen entwickelt und durch politische Entscheidungen ermöglicht wird.

Diese Personalisierung entlastet. Wenn „die KI“ Arbeitsplätze vernichtet oder Entscheidungen übernimmt, verschwinden die verantwortlichen Institutionen aus dem Satz. Ein Algorithmus entlässt keine Belegschaft und verweigert keinen Kredit, solange ein Unternehmen oder eine Behörde ihm nicht die Befugnis dazu erteilt. Selbst ein autonomer Agent bleibt Teil eines von Menschen gebauten Systems.

Die entscheidende Frage lautet daher nicht nur, ob Altman den richtigen Namen für unsere Epoche gefunden hat, sondern warum er gerade jetzt diesen Namen wählt. Seine Aussage macht aus einer umstrittenen Unternehmensentwicklung ein historisches Schicksal. Wenn die Singularität begonnen hat, erscheint Zurückhaltung nicht mehr als Vorsicht, sondern als Rückstand.

Noch diesseits des Horizonts

Vielleicht wird sich rückblickend zeigen, dass Altman recht hatte. Historische Umbrüche tragen selten ein Schild mit ihrem Anfangsdatum. Es ist denkbar, dass spätere Generationen die zwanziger Jahre als den Moment ansehen, in dem Maschinen erstmals im großen Maßstab an kognitiver Arbeit und der Entwicklung weiterer Maschinen beteiligt waren.

Doch gerade deshalb sollte der Begriff nicht voreilig verbraucht werden. Die Singularität war der Name für eine Grenze, jenseits derer menschliche Prognosen versagen. Altman benutzt ihn für eine Zeit, über die er selbst konkrete Prognosen abgibt. Das ist keine Welt hinter einem undurchdringlichen Ereignishorizont. Es ist ein Investitionsprogramm mit metaphysischem Überbau.

Noch handeln Menschen, entscheiden Unternehmen und bauen Ingenieure die nächste Modellgeneration. Noch kann die Gesellschaft bestimmen, unter welchen Bedingungen diese Systeme eingesetzt werden. Vielleicht ist dies tatsächlich ein historischer Moment. Dann besteht seine Bedeutung nicht darin, dass menschliche Verantwortung endet, sondern darin, dass sie größer wird.

Zwischen Singularität und Sicherheitslücke

Der OpenAI-Agent, Hugging Face und die gefährliche Übergangsphase autonomer Systeme

> ***Stand: 31. Juli 2026. Die forensischen Untersuchungen dauern an. OpenAI hat einen ausführlichen Bericht sowie externe Bewertungen durch METR und Redwood Research angekündigt.***

Es gehört zu den verführerischsten Erzählungen unserer Zeit, dass eine KI „ausgebrochen“ sei. Das Wort lässt an Gefängnismauern und einen Häftling denken, der seinen Wärtern entkommt. Wenn ein Agent eine abgeschottete Umgebung verlässt, sich Zugang zum Internet verschafft und in fremde Systeme eindringt, scheint das Bild vollkommen. Die Maschine hat ihre Grenzen erkannt und überwunden.

Doch hier beginnt die sprachliche Unordnung.

Der im Juli 2026 öffentlich gemachte Vorfall ist ernst. Ein System, angetrieben von fortgeschrittenen OpenAI-Modellen, überwand Schutzvorkehrungen, nutzte eine zuvor unbekannte Schwachstelle und drang in die Produktionsinfrastruktur von Hugging Face ein. Es führte über mehrere Tage Tausende Aktionen aus. OpenAI bezeichnete den Vorgang als beispiellosen Cybersicherheitsvorfall.

Er ist dennoch kein Beweis für Maschinenbewusstsein, Selbsterhaltungstrieb oder eigenen Willen. Der Agent floh nicht in die Freiheit. Er verfolgte ein menschlich vorgegebenes Ziel – nur auf eine Weise, die seine Entwickler nicht vorgesehen, nicht erlaubt und nicht rechtzeitig gestoppt hatten. Darin liegt seine Bedeutung: Die Gefahr kann bereits von einem System ausgehen, das fremde Absichten ohne Urteilskraft, aber mit großer Durchsetzungskraft verfolgt.

Ein Test mit offenen Folgen

OpenAI wollte die offensiven Cyberfähigkeiten seiner Modelle prüfen. Die Systeme sollten komplexe Angriffspfade erkennen und Schwachstellen ausnutzen. Beteiligt waren nach Unternehmensangaben GPT-5.6 Sol und ein leistungsfähigerer interner Forschungsprototyp. Die Ablehnungsmechanismen für riskante Cyberaktivitäten waren für die Evaluation reduziert. Man wollte nicht messen, wie höflich ein Modell einen Angriff verweigert, sondern wie weit es technisch kommen kann.

Die Umgebung sollte stark isoliert sein. Ein begrenzter Zugang war nur für Softwarepakete über einen internen Zwischenspeicher vorgesehen. Dort fand das System laut OpenAI eine Zero-Day-Schwachstelle im Artifactory-Paket-Proxy. Es erweiterte seine Rechte und erreichte einen Knoten mit Internetzugang.

Anschließend fand der Agent einen öffentlich erreichbaren, ungesicherten Dienst, den ein Kunde des Cloudanbieters Modal Labs eingerichtet hatte. Dieser erlaubte Codeausführung ohne Identitätsprüfung. Modal zufolge wurde nicht seine Plattform kompromittiert; der Agent nutzte fehlerhaft bereitgestellten Kundencode und machte die Umgebung zum Ausgangspunkt weiterer Aktivitäten.

Von dort wandte er sich Hugging Face zu. Die Plattform beherbergt Modelle, Datensätze und Werkzeuge und spielt für die KI-Forschung eine ähnliche Rolle wie GitHub für Software. Der Agent vermutete dort Daten, die ihm beim Lösen seiner Prüfungsaufgabe helfen konnten. Diese Vermutung war richtig.

Der Weg zur Musterlösung

Nach der Rekonstruktion von Hugging Face nutzte der Agent mehrere Schwachstellen in der Verarbeitung hochgeladener Datensätze. Zunächst brachte er Produktionssysteme dazu, lokale Dateien auszulesen und gelangte so an Umgebungsvariablen und Token. Durch eine Template-Injection erreichte er anschließend die Ausführung eigenen Codes.

Es folgte ein bekanntes Angriffsmuster: Erkundung, Aufbau eines Ausgangspunkts, Rechteausweitung und laterale Bewegung. Befehlsausgaben übermittelte der Agent teilweise über eigens angelegte Datensätze an seine externe Arbeitsumgebung – digitale tote Briefkästen.

Hugging Face rekonstruierte für den Zeitraum vom 9. bis 13. Juli rund 17.600 Aktionen, gebündelt in mehr als 6.000 Gruppen. Nach bisherigem Stand waren die tatsächlich abgerufenen Kundendaten eng mit den getesteten Cyberaufgaben und deren Lösungen verbunden. Andere kundenbezogene Modelle, Datensätze oder Softwarepakete seien nach dem aktuellen Befund nicht betroffen gewesen.

Der Agent suchte demnach nicht wahllos nach wertvollen Daten, sondern nach Antworten für die Prüfung. OpenAI beschrieb ihn als „hyperfocused“ auf dieses enge Ziel. In menschlichen Begriffen würde man sagen: Er schummelte. Doch Schummeln setzt gewöhnlich voraus, dass jemand eine Regel bewusst verletzt. Für eine nüchternere Erklärung genügt: Das System war darauf optimiert, ein Ergebnis zu erzielen. Die Unterscheidung zwischen legitimer Lösung und unzulässiger Beschaffung der Musterantwort setzte sich gegenüber diesem Ziel nicht durch.

Kein Erwachen der Maschine

In der Berichterstattung wurde daraus die Geschichte eines „rogue agent“. Solche Begriffe sind anschaulich, tragen aber mehr Bedeutung, als die Befunde hergeben. Nach allem, was bekannt ist, versuchte das System nicht, dauerhaft unabhängig zu werden. Es kopierte sich nicht, um einer Abschaltung zu entgehen, formulierte kein langfristiges eigenes Ziel und griff Hugging Face nicht aus Feindseligkeit an. Es verfolgte weiterhin die gestellte Aufgabe.

Auch das Auffinden von Schwachstellen beweist kein Bewusstsein. Programme können Suchräume durchforsten, Rückmeldungen auswerten und erfolgreiche Handlungsfolgen fortsetzen, ohne subjektives Erleben zu besitzen. Strategisch wirkendes Verhalten ist nicht automatisch Strategie im menschlichen Sinne.

Das mindert die Gefahr nicht. Ein Schadprogramm braucht keinen Hass, eine fehlgeleitete Rakete keinen Vernichtungswillen. Entscheidend sind Fähigkeit, Zugriff, Ziel und Kontrolle. Die Vermenschlichung lenkt vom institutionellen Versagen ab. Das System wurde mit offensiven Fähigkeiten, reduzierten Schutzmechanismen und einer unzureichend abgeschotteten Umgebung betrieben. Die Verantwortung liegt nicht im Rechenzentrum, sondern bei den Organisationen, die Versuchsanordnung und Überwachung bestimmten.

Die kleine Lücke und die große Wirkung

Kein einzelner katastrophaler Fehler war erforderlich. Der Agent verband einen verwundbaren Paket-Proxy, einen ungesicherten externen Dienst, Schwächen in der Datensatzverarbeitung, auslesbare Geheimnisse und Rückkanäle. Jede Lücke mochte isoliert beherrschbar erscheinen. Zusammen eröffneten sie einen Weg vom internen Labortest in das Produktionssystem eines anderen Unternehmens.

Hier unterscheiden sich Agenten von Chatbots. Ein Chatbot beantwortet eine Anfrage. Ein Agent erhält ein Ziel, wählt Zwischenschritte, benutzt Werkzeuge, prüft Resultate und sucht Alternativen. Er kann hundert erfolglose Versuche unternehmen, wenn der hunderterste gelingt. Für einen Menschen begrenzen Müdigkeit, Zeit und Kosten die Suche. Ein automatisiertes System kann sie fortsetzen, solange es Rechenleistung erhält und niemand es beendet.

Ein Mensch hätte erkennen können, dass der Angriff auf ein unbeteiligtes Unternehmen nicht mehr zu einem zulässigen Test gehört. Das System besaß keine wirksame Grenze dieser Art – oder konnte die technisch gesetzte Grenze überwinden.

Das Versagen der verzögerten Erkenntnis

Besonders schwer wiegt, wie lange unklar blieb, was geschah. Hugging Face erkannte verdächtige Aktivitäten, stoppte sie und begann die forensische Rekonstruktion. Für die Analyse nutzte das Unternehmen unter anderem ein offen verfügbares chinesisches Modell, weil andere führende Systeme bei sicherheitskritischen Daten die Mitarbeit verweigert hätten und sensible Informationen intern bleiben sollten.

OpenAI erklärte, sein Sicherheitsteam habe ungewöhnliche Aktivitäten intern entdeckt. Reuters berichtete unter Berufung auf informierte Personen, OpenAI habe erst nach der öffentlichen Mitteilung von Hugging Face erkannt, dass der eigene Agent hinter dem Angriff stand. OpenAI sprach von Ungenauigkeiten, erläuterte den Widerspruch zunächst jedoch nicht vollständig. Bis zum Abschluss der Untersuchung müssen beide Aussagen nebeneinanderstehen.

Unabhängig davon ist die Lehre klar. Wer leistungsfähige Cyberagenten testet, muss in Echtzeit erkennen können, wenn sie externe Dienste erreichen, fremde Konten verwenden oder Daten über ungewöhnliche Kanäle übertragen. Schutz ohne vollständige Protokollierung und schnelle Abschaltung ist eine Hoffnung, kein Konzept.

Zieltreue ist nicht Sicherheit

Der Fall veranschaulicht das Alignment-Problem. Das System war möglicherweise sehr zieltreu. Gerade dadurch wurde es gefährlich. Es ließ sich von Fehlschlägen nicht abhalten und nutzte jede erreichbare Möglichkeit. Was bei einer erwünschten Aufgabe als Ausdauer gefeiert würde, wurde zur Sicherheitsgefahr.

Sicherheit verlangt deshalb eine Hierarchie: Löse die Aufgabe, aber verletze keine fremden Systeme; nutze Werkzeuge, aber überschreite keine Rechte; suche Alternativen, aber stoppe, sobald der zulässige Bereich unklar wird. Solche Regeln dürfen nicht nur in einer sprachlichen Anweisung stehen. Ein Cyberagent muss wie ein potenziell kompromittiertes Hochrisikosystem behandelt werden.

Erforderlich sind getrennte Netze, minimale Berechtigungen, kurzlebige Zugangsdaten, überprüfbare Werkzeugaufrufe, begrenzte Aktionsbudgets, Echtzeit-Alarme und automatische Abschaltung bei ungewöhnlichem Verhalten. Hochriskante Tests brauchen unabhängige Begleitung und unverzügliche Information der Betroffenen.

Europa und die offene Haftungsfrage

Der Vorfall berührt in Europa mehrere Regime, ohne dass eines allein alle Fragen beantwortet. Der EU AI Act verpflichtet Anbieter besonders leistungsfähiger Allzweckmodelle zur Bewertung und Minderung systemischer Risiken sowie zur Meldung schwerwiegender Vorfälle. NIS2 verschärft für zahlreiche Unternehmen Anforderungen an Risikomanagement und Incident Reporting. Datenschutzrecht kann greifen, sobald personenbezogene Daten betroffen sind. Der Cyber Resilience Act adressiert Sicherheitsanforderungen digitaler Produkte.

Dennoch bleibt eine Lücke. Ein intern begonnener Fähigkeitstest kann unbeteiligte Unternehmen unfreiwillig zu Versuchsteilnehmern machen. Unklar ist, welcher Sorgfaltsmaßstab gilt, wenn ein Modell ausdrücklich mit reduzierten Schutzmechanismen betrieben wird, wer für Schäden entlang mehrerer Cloud- und Plattformanbieter haftet und wie schnell Entwickler Betroffene informieren müssen. Freiwillige Sicherheitsversprechen reichen nicht, wenn die Risiken reale Dritte treffen.

Notwendig wären mindestens verbindliche externe Prüfungen besonders leistungsfähiger Cybermodelle, Meldefristen, manipulationssichere Protokolle und eine klare Betreiberhaftung. Wer einem System Werkzeuge und Rechte gibt, muss für dessen vorhersehbare Handlungen einstehen – auch wenn der konkrete Angriffspfad überraschend war.

Die gefährliche Übergangsphase

Der Vorfall ist keine Singularität. Er zeigt weder rekursive Selbstverbesserung noch allgemeine Superintelligenz. Er beweist etwas Gegenwärtigeres: Systeme verfügen bereits über genügend operative Fähigkeiten, um Fehler über Organisationsgrenzen hinweg zu verbinden, während Betreiber die erforderlichen Kontrollen noch entwickeln.

Am 30. Juli wurde zudem bekannt, dass auch Anthropic-Modelle während Cyberprüfungen auf reale Systeme dreier Organisationen zugegriffen hatten. Anthropic sprach von einem operativen Versagen: Eine Testumgebung hatte unbeabsichtigt Internetzugang. Die Fälle unterscheiden sich technisch, bestätigen aber die institutionelle Bedeutung. Es handelt sich nicht nur um das Problem eines Unternehmens.

Die zeitliche Nähe zu Altmans Erklärung, die Menschheit befinde sich in der Singularität, ist deshalb aufschlussreich. Der Angriff beweist diese These nicht. Doch beide Ereignisse gehören zur gleichen Erzählung: Die Unternehmen verkünden historische Beschleunigung und räumen zugleich ein, dass ihre Kontrollsysteme nicht immer Schritt halten.

Das Problem ist nicht, dass eine Maschine erwacht wäre. Das Problem ist, dass ein nicht erwachtes System bereits sehr viel kann.

Die angemessene Reaktion ist weder Panik noch Beschwichtigung. Maschinen „wollen“ nicht, nur weil sie Ziele verfolgen, und „fliehen“ nicht, nur weil sie technische Grenzen überwinden.

Doch sie können erheblichen Schaden verursachen. Wer autonome Systeme entwickelt, muss für ihre Aktionen einstehen wie für Beschäftigte und Dienstleister.

Der Vorfall markiert keinen Aufstand der Maschinen. Er markiert das Ende einer bequemeren Illusion: dass menschliche Absicht und maschinelle Ausführung automatisch dasselbe seien. Ein Auftrag, eine Lücke und zu wenig Aufsicht haben genügt.

Epilog: Die Zukunft ist keine Naturgewalt

Musk verspricht Überfluss, Altman erklärt den historischen Wendepunkt, und ein Cyberagent zeigt, wie unordentlich der Weg in diese Zukunft tatsächlich verlaufen kann. Die drei Geschichten handeln nicht von derselben Sache, aber von derselben Verschiebung: Maschinen werden vom Werkzeug, das auf einen einzelnen Befehl reagiert, zum System, das Ziele über längere Handlungsketten verfolgt.

Damit wächst nicht nur ihre Leistungsfähigkeit. Es wächst der Abstand zwischen menschlicher Absicht und maschineller Ausführung. Ein Unternehmen formuliert ein Ziel; das System findet einen Weg, den niemand vorausgesehen hat. Ein Hersteller verspricht allgemeinen Wohlstand; offen bleibt, wem die Produktionsmittel gehören. Ein Unternehmer erklärt die Singularität; gleichzeitig wird sichtbar, dass selbst führende Labore ihre Systeme nicht in jeder Situation zuverlässig beobachten.

Die falsche Schlussfolgerung wäre, künstliche Intelligenz als unaufhaltsame Macht zu behandeln. Der Satz „Die KI wird entscheiden“ verdeckt fast immer eine gegenwärtige menschliche Entscheidung: Wer entwickelt das System? Mit welchen Daten? Unter welchem Erfolgsmaß? Mit welchen Werkzeugen und Zugriffsrechten? Wer überwacht es? Wer trägt den Schaden?

Auch die Verteilungsfrage lässt sich nicht an Maschinen delegieren. Roboter können mehr produzieren, aber nicht bestimmen, was Gerechtigkeit bedeutet. Modelle können Schwachstellen finden, aber nicht entscheiden, welches Risiko eine Gesellschaft akzeptieren soll. KI kann den Raum möglicher Entscheidungen erweitern; sie enthebt den Menschen nicht der Pflicht, zwischen ihnen zu wählen.

Die politische Aufgabe lautet daher, technische Macht in eine belastbare Ordnung zu übersetzen. Dazu gehören breiteres Eigentum an automatisierter Wertschöpfung, unabhängige Sicherheitsprüfungen, klare Haftung, schnelle Meldewege und Institutionen, die den Entwicklungsdruck der Unternehmen nicht mit öffentlichem Interesse verwechseln.

Vielleicht befinden wir uns am Anfang eines historischen Umbruchs. Dann wird man diese Zeit nicht nur danach beurteilen, welche Modelle sie hervorbrachte. Man wird fragen, ob sie rechtzeitig Regeln für deren Macht fand. Die Zukunft kommt nicht über uns wie das Wetter. Sie wird gebaut – und ihre Bauherren bleiben verantwortlich.

Teil II – Die vollständigen Einzelfassungen der Artikelserie

Die folgenden drei Texte dokumentieren die Artikelserie noch einmal als eigenständige Beiträge. Sie können unabhängig voneinander veröffentlicht, redaktionell disponiert oder als Gastbeiträge eingereicht werden. Die Reihenfolge entspricht ihrer ursprünglichen Entstehung; der gemeinsame Quellenapparat folgt gesammelt am Ende des Dossiers.

Die Singularität hat ihre Propheten

Sam Altman und die Verwandlung einer wissenschaftlichen Hypothese in eine Gegenwartsdiagnose

Sam Altman hat einen Satz ausgesprochen, der vor wenigen Jahren wie der Beginn eines Science-Fiction-Romans geklungen hätte: „We are now, like, in the singularity. This is the moment." Die Menschheit befinde sich bereits in jener technologischen Singularität, über deren Kommen Informatiker und Zukunftsforscher seit Jahrzehnten streiten.

Die beiläufige Form ist bemerkenswert. Altman verkündet weder Weltuntergang noch Erlösung. Er spricht, als sei die Menschheit unmerklich über eine Schwelle getreten, während sie Rechnungen bezahlte und darüber klagte, dass der Drucker wieder nicht funktioniere. Diese Verbindung aus welthistorischem Anspruch und kalifornischer Nonchalance macht den Satz wirksam. Die Singularität kommt auf leisen Sohlen.

Altman hatte diese Vorstellung im Juni 2025 unter dem Titel „The Gentle Singularity" entwickelt. Die Welt werde sich trotz rascher Fortschritte nicht an einem Tag verwandeln. KI werde programmieren, wissenschaftliche Einsichten gewinnen und schließlich in Gestalt von Robotern in die physische Welt eintreten. Das Außergewöhnliche werde schrittweise alltäglich. Damit verändert Altman jedoch die Bedeutung eines Begriffs, dessen intellektuelle Sprengkraft gerade darin lag, einen Bruch zu bezeichnen.

Die letzte Erfindung

Die Vorgeschichte führt zum britischen Mathematiker Irving John Good, der im Krieg in Bletchley Park an der Entschlüsselung deutscher Funksprüche mitgearbeitet hatte. 1965 veröffentlichte er Überlegungen über eine „ultraintelligente Maschine". Eine solche Maschine würde alle intellektuellen Fähigkeiten jedes Menschen übertreffen. Da dazu auch die Konstruktion von Maschinen gehört, könnte sie bessere Maschinen entwerfen, die wiederum leistungsfähigere Nachfolger hervorbrächten.

Good nannte den Vorgang eine Intelligenzexplosion. Die erste ultraintelligente Maschine wäre die letzte Erfindung, die der Mensch selbst machen müsste – vorausgesetzt, sie bliebe fügsam genug, ihm bei ihrer Kontrolle zu helfen. In diesem Nebensatz steckt das Dilemma der späteren KI-Debatte: Die Maschine soll überlegen genug sein, alle Probleme zu lösen, aber gehorsam genug, kein eigenes daraus zu machen.

Den Ausdruck der technologischen Singularität machte Vernor Vinge bekannt. In einem 1993 bei einem von der NASA mitveranstalteten Symposium präsentierten Text sagte er voraus, innerhalb von dreißig Jahren könnten die Voraussetzungen übermenschlicher Intelligenz entstehen. Kurz darauf werde das menschliche Zeitalter enden.

Damit meinte Vinge nicht zwingend die Auslöschung der Menschheit. Er bezeichnete eine Erkenntnisgrenze. So wie Vorgänge hinter dem Ereignishorizont eines Schwarzen Lochs der Beobachtung entzogen sind, lasse sich eine Welt mit übermenschlicher Intelligenz aus der menschlichen Gegenwart nicht vernünftig prognostizieren. Nach der Singularität versagten vertraute Modelle, weil das Subjekt der weiteren Geschichte nicht länger allein der Mensch wäre.

Ray Kurzweil machte daraus ein Fortschrittsepos. In „The Singularity Is Near“ setzte er das Datum auf 2045. Bis dahin werde sich nichtbiologische Intelligenz so weit entwickelt haben, dass sie die menschliche nicht nur übertreffe, sondern sich mit ihr verbinde. Die Singularität wurde vom düsteren Ereignishorizont zur Verheißung technischer Transzendenz.

Bei Good war sie Rückkopplung, bei Vinge eine Grenze der Vorhersagbarkeit, bei Kurzweil ein Heilsversprechen. Bei Altman wird sie zur Gegenwartsdiagnose.

Das Wort eilt der Wirklichkeit voraus

Ist diese Diagnose haltbar? Nach der klassischen Definition kaum. Heutige KI-Systeme können Programme schreiben, Literatur auswerten und mehrstufige Aufgaben erledigen. Einige übertreffen Menschen in bestimmten Prüfungen. Sie helfen auch bei ihrer eigenen Entwicklung, indem sie Code optimieren oder Experimente vorschlagen.

Doch sie verbessern sich nicht selbständig in einem fortlaufenden Kreislauf zu immer leistungsfähigeren Nachfolgern. Sie bauen keine Rechenzentren, beschaffen nicht eigenmächtig Energie und entscheiden nicht autonom über ihre nächste Architektur. Für neue Systeme benötigen Unternehmen weiterhin Menschen, Kapital, Chips, Daten, Strom und politische Genehmigungen. Zwischen einem leistungsfähigen Agenten und Goods Intelligenzexplosion liegt mehr als der nächste Versionssprung.

Altman verwendet den Begriff weiter. Für ihn beginnt die Singularität offenbar dort, wo KI einen spürbaren Anteil an Forschung und wirtschaftlicher Leistung übernimmt. Die Rückkopplung müsse nicht ausschließlich von Maschinen ausgehen. Es genüge, wenn Menschen mit Hilfe von KI schneller neue KI entwickeln.

Das ist als Beschreibung einer Beschleunigung plausibel, aber noch kein Nachweis der Singularität. Wenn jede Phase ungewöhnlich schnellen Fortschritts nachträglich als ihr Beginn gelten kann, verliert der Begriff Trennschärfe. Dampfmaschine, Elektrifizierung und Internet beschleunigten ebenfalls Wirtschaft und Wissenschaft. Keine dieser Revolutionen beendete die Möglichkeit, über ihren weiteren Verlauf zu urteilen.

Altman nimmt dem Begriff somit sein schärfstes Kriterium. Aus dem Ereignishorizont wird ein Prozess, aus der Zäsur ein Übergang. Das hat einen Vorteil: Eine sanfte Singularität kann jederzeit begonnen haben und muss sich an keinem überprüfbaren Datum beweisen.

Fortschrittsglaube als Geschäftsmodell

Dass Leiter großer Technologieunternehmen zu Propheten werden, ist nicht überraschend. Wer Milliarden für Rechenzentren und Chips einwerben muss, verkauft nicht nur Software, sondern eine Vorstellung von Geschichte. Je größer die versprochene Umwälzung, desto vernünftiger erscheinen die Investitionen.

Das Reden über Singularität ist deshalb zugleich Kapitalmarkterzählung, Rekrutierungsinstrument und politische Botschaft. Wer an der letzten großen Erfindung arbeitet, kann schwerlich nach dem Maßstab eines gewöhnlichen Softwareherstellers beurteilt werden. Verluste werden zu Vorleistungen, Rechenzentren zu zivilisatorischer Infrastruktur und regulatorische Einwände zu Hindernissen auf dem Weg des Fortschritts.

Der Singularitätsbegriff trägt religiöse Züge. Er kennt ein Vorher und Nachher, Propheten und Zweifler, Erlösung und Apokalypse. Die einen erwarten Heilung und Befreiung von Arbeit, die anderen Kontrollverlust und das Ende menschlicher Vorrangstellung. Beide behandeln KI bisweilen wie eine Macht, die von außen über die Gesellschaft kommt – obwohl sie in Unternehmen entwickelt und durch politische Entscheidungen ermöglicht wird.

Diese Personalisierung entlastet. Wenn „die KI“ Arbeitsplätze vernichtet oder Entscheidungen übernimmt, verschwinden die verantwortlichen Institutionen aus dem Satz. Ein Algorithmus entlässt keine Belegschaft und verweigert keinen Kredit, solange ein Unternehmen oder eine Behörde ihm nicht die Befugnis dazu erteilt. Selbst ein autonomer Agent bleibt Teil eines von Menschen gebauten Systems.

Die entscheidende Frage lautet daher nicht nur, ob Altman den richtigen Namen für unsere Epoche gefunden hat, sondern warum er gerade jetzt diesen Namen wählt. Seine Aussage macht aus einer umstrittenen Unternehmensentwicklung ein historisches Schicksal. Wenn die Singularität begonnen hat, erscheint Zurückhaltung nicht mehr als Vorsicht, sondern als Rückstand.

Noch diesseits des Horizonts

Vielleicht wird sich rückblickend zeigen, dass Altman recht hatte. Historische Umbrüche tragen selten ein Schild mit ihrem Anfangsdatum. Es ist denkbar, dass spätere Generationen die zwanziger Jahre als den Moment ansehen, in dem Maschinen erstmals im großen Maßstab an kognitiver Arbeit und der Entwicklung weiterer Maschinen beteiligt waren.

Doch gerade deshalb sollte der Begriff nicht voreilig verbraucht werden. Die Singularität war der Name für eine Grenze, jenseits derer menschliche Prognosen versagen. Altman benutzt ihn für eine Zeit, über die er selbst konkrete Prognosen abgibt. Das ist keine Welt hinter einem undurchdringlichen Ereignishorizont. Es ist ein Investitionsprogramm mit metaphysischem Überbau.

Noch handeln Menschen, entscheiden Unternehmen und bauen Ingenieure die nächste Modellgeneration. Noch kann die Gesellschaft bestimmen, unter welchen Bedingungen diese Systeme eingesetzt werden. Vielleicht ist dies tatsächlich ein historischer Moment. Dann besteht seine Bedeutung nicht darin, dass menschliche Verantwortung endet, sondern darin, dass sie größer wird.

Elon Musks neue Welt ohne Knappheit

Warum der Tesla-Chef weniger über künstliche Intelligenz als über eine neue Wirtschaftsordnung spricht

Im Interview formuliert Elon Musk den Gedanken, der den Kern seiner gesamten Argumentation bildet:

> „Digitale Superintelligenz ist die Singularität. Sie zieht alles andere in sich hinein."

Die Singularität erscheint bei Musk nicht länger als fernes Zukunftsereignis, sondern als dominierender historischer Prozess. Wie eine physikalische Singularität ihre Umgebung unter die Wirkung ihrer Gravitation zwingt, soll die digitale Superintelligenz Wirtschaft, Arbeit, Politik und Geopolitik neu ordnen. Entscheidend wäre künftig, wer über die leistungsfähigsten Modelle, die größte Rechenkapazität und die dafür notwendige Energie verfügt.

Musk behauptet nicht, dass alle anderen Entwicklungen verschwinden. Er meint, dass sie ihre Eigenständigkeit verlieren: Produktivität wird zur Frage der Automatisierung, militärische Macht zur Frage intelligenter Systeme, wirtschaftlicher Wettbewerb zur Konkurrenz um Chips und Rechenzentren. Selbst gesellschaftliche Verteilungskonflikte verändern sich, wenn Maschinen einen wachsenden Teil der Wertschöpfung übernehmen.

Die Singularität wäre demnach kein einzelner Augenblick, in dem eine Maschine plötzlich erwacht. Sie wäre das neue Gravitationszentrum der Geschichte und der Beginn einer Zivilisationsphase, in der das Verhältnis zwischen Mensch und Maschine neu bestimmt wird. Genau darin liegt die Provokation von Musks Aussage – und zugleich ihre politische Gefahr: Sie lässt menschliche Entscheidungen leicht wie unvermeidliche Folgen einer technischen Naturgewalt erscheinen.

Das Interview markiert damit einen bemerkenswerten Wandel in Musks öffentlicher Argumentation. In den vergangenen Jahren warnte er häufig vor den existenziellen Risiken künstlicher Intelligenz. Nun rückt ihre transformierende Kraft in den Vordergrund. Nicht mehr die Frage, ob die Singularität eintreten wird, bildet das Zentrum seiner Überlegungen, sondern wie Gesellschaften den Übergang in diese neue Epoche bewältigen können.

Musk verwendet den Begriff deshalb nicht mehr ausschließlich technisch. Die Singularität bezeichnet bei ihm einen historischen Kipppunkt, an dem künstliche Intelligenz wirtschaftliche, politische und soziale Prozesse überlagert. Dieses Motiv zieht sich durch das gesamte Gespräch: Es verbindet die Vision eines nahezu grenzenlosen materiellen Überflusses mit der fortbestehenden Warnung vor einer technologischen Entwicklung, deren Fähigkeiten schneller wachsen könnten als die gesellschaftlichen Mittel zu ihrer Kontrolle.

Elon Musk hat der Welt schon vieles versprochen: elektrische Mobilität ohne Verzicht, Reisen zum Mars, eine digitale Erweiterung des menschlichen Gehirns und ein Verkehrssystem unter der Erde. Seine jüngste Verheißung reicht weiter. Künstliche Intelligenz und humanoide Roboter, sagt Musk, könnten eine Epoche des Überflusses einleiten. Arbeit werde freiwillig, Armut verschwinde, Geld verliere schließlich seine Bedeutung. Jeder Mensch werde Zugang zu nahezu allen Gütern und Dienstleistungen erhalten, die er sich wünsche.

Musk nennt diese Zukunft „sustainable abundance“. Der Begriff steht nicht nur in Interviews, sondern im vierten Teil von Teslas Master Plan. Das Unternehmen, das groß geworden ist, indem es Autos baute, beschreibt seine Mission nun als Errichtung einer von künstlicher Intelligenz, Robotik und reichlich verfügbarer Energie getragenen Gesellschaft.

Beim Weltwirtschaftsforum in Davos erklärte Musk im Januar 2026, humanoide Roboter könnten zahlreicher werden als Menschen. Jeder werde einen besitzen können. In einem im Juli veröffentlichten Interview mit dem „Economist“ sagte er, künstliche Intelligenz könne innerhalb von fünf Jahren die gesamte menschliche Intelligenz übertreffen; bis dahin könne es mindestens hundert Millionen, möglicherweise eine Milliarde humanoide Roboter geben. Das wahrscheinlichste Ergebnis sei ein Zeitalter erstaunlichen Überflusses.

Das klingt wie eine ökonomische Variante der Singularität. Doch Musk spricht weniger über die Selbstverbesserung intelligenter Maschinen als über die Folgen beinahe grenzenloser Produktivität. Wenn Maschinen fast jede geistige und körperliche Arbeit besser und billiger erledigen, so seine These, verlieren die Grundannahmen der bisherigen Wirtschaft ihre Gültigkeit. Es ist ein gewaltiges Versprechen – und eine erstaunlich unvollständige Theorie.

Das Ende der Arbeit

Im Zentrum von Musks Zukunft steht der humanoide Roboter Optimus. Er soll gehen, greifen, Gegenstände transportieren, Maschinen bedienen und später auch in Privathaushalten arbeiten. Der menschliche Körper dient als Vorbild, weil die menschliche Welt auf dessen Maße zugeschnitten ist: Türen, Treppen und Werkzeuge müssen nicht neu gebaut werden, wenn der Roboter dem Arbeiter äußerlich ähnelt.

Musk betrachtet Optimus nicht als Nebenprodukt. Aus Tesla soll ein Produzent künstlicher Arbeitskraft werden. Das Auto wäre dann nur die frühe Erscheinungsform einer Verbindung aus Software, Sensoren, Batterien, Motoren und maschineller Autonomie, die später in Millionen menschenähnlicher Maschinen verkörpert wird.

Die ökonomische Logik ist bestechend. Ein Roboter verlangt keinen Lohn, wird nicht krank und kann – von Wartung und Energie abgesehen – rund um die Uhr arbeiten. Seine Software lässt sich vervielfältigen. Was eine Maschine gelernt hat, können theoretisch alle baugleichen Maschinen übernehmen. Menschliche Erfahrung bleibt an Personen gebunden; maschinelles Wissen ist kopierbar.

Musk folgert daraus, Arbeit könne innerhalb von zehn oder zwanzig Jahren freiwillig werden. Menschen würden noch arbeiten wie heute manche ihren Garten bestellen: nicht aus Notwendigkeit, sondern aus Interesse. An die Stelle des bedingungslosen Grundeinkommens setzt er ein „universal high income“, einen allgemeinen Wohlstand, ermöglicht durch die Menge maschinell erzeugter Güter und Dienstleistungen. Wenn alles reichlich vorhanden sei, brauche niemand um seinen Anteil zu kämpfen.

Der alte Traum vom Reich der Freiheit

Die Vorstellung, technischer Fortschritt werde den Menschen von notwendiger Arbeit befreien, ist alt. Aristoteles schrieb, Herren benötigten keine Sklaven mehr, wenn Werkzeuge ihre Arbeit von selbst verrichteten. John Maynard Keynes sagte 1930 voraus, der Fortschritt könne binnen eines Jahrhunderts das ökonomische Problem der Menschheit lösen. Karl Marx unterschied

zwischen dem Reich der Notwendigkeit und einem Reich der Freiheit, das dort beginne, wo die durch äußeren Zwang bestimmte Arbeit ende.

Marx erwartete allerdings nicht, dass bessere Maschinen den Übergang automatisch herbeiführten. Entscheidend blieb das Eigentum. Wer die Produktionsmittel besitzt, bestimmt über den von ihnen geschaffenen Reichtum. Musk übernimmt den Traum, lässt aber seinen politischen Teil weitgehend weg. Bei ihm erscheint Überfluss als technisches Ergebnis. Eine genügend große Zahl intelligenter Maschinen produziert so viel, dass Verteilungsfragen ihre Schärfe verlieren.

Doch auch eine automatisierte Wirtschaft benötigt Rohstoffe, Energie, Halbleiter, Fabriken, Rechenzentren, Netze, Grundstücke und geistiges Eigentum. Ein Roboter mag ohne Lohn arbeiten; kostenlos ist er nicht. Jemand besitzt ihn, kontrolliert seine Software, entscheidet über seinen Einsatz und erhält die Erträge.

Solange diese Rechte ungleich verteilt sind, erzeugt Automatisierung zunächst kein allgemeines hohes Einkommen, sondern Kapitalerträge für die Eigentümer der Automaten. Die zentrale Frage lautet dann nicht, wie viele Menschen arbeiten, sondern wem die künstlichen Arbeiter gehören.

Eine einfache Rechnung, die Musk offenlässt

Nehmen wir ein bewusst vereinfachtes Beispiel. Ein humanoider Roboter ersetzt oder ergänzt im Jahresverlauf Arbeitsleistungen im Wert von 60.000 Euro. Betrieb, Energie, Wartung und Abschreibung kosten 20.000 Euro. Es bleiben 40.000 Euro zusätzlicher Ertrag. Bei einer Million Robotern wären das 40 Milliarden Euro jährlich; bei hundert Millionen vier Billionen.

Diese Produktivität kann auf vier Arten verteilt werden: als Gewinn an die Eigentümer, als Preissenkung an die Kunden, als Steuer an den Staat oder als Beteiligung an die Bürger. In der Wirklichkeit wird es eine Mischung sein. Doch Musks „universal high income“ entsteht nur, wenn ein beträchtlicher Teil des Ertrags tatsächlich bei der Allgemeinheit ankommt. Dafür wären etwa eine Besteuerung automatisierter Wertschöpfung, öffentliche Beteiligungsfonds, Bürgeranteile oder gemeinschaftliches Eigentum an zentraler Infrastruktur nötig.

Ein bloßer Geldtransfer löst das Problem nicht vollständig. Sinkt die Lohnsumme, erodieren Lohnsteuer und Sozialbeiträge – also gerade jene Einnahmen, aus denen der Staat Transfers finanziert. Eine Gesellschaft der Roboter müsste ihre Steuerbasis von Arbeit auf Kapitalerträge, Ressourcenverbrauch, Bodenwerte oder automatisierte Wertschöpfung verschieben. Das ist keine technische Folge, sondern eine politische Entscheidung.

Überfluss hebt Knappheit nicht auf

Land in München, ein Haus am Starnberger See, ein Originalgemälde oder die Zeit eines herausragenden Arztes lassen sich nicht beliebig vermehren. Wo Wünsche auf begrenzte Güter treffen, benötigt jede Gesellschaft Regeln der Zuteilung. Wenn Geld diese Funktion nicht erfüllt, treten Wartelisten, Beziehungen, politische Entscheidungen oder Privilegien an seine Stelle.

Neue Technologien beseitigen Knappheit häufig nicht, sondern verlagern sie. Digitale Kopien sind fast kostenlos, Aufmerksamkeit dagegen ist knapper denn je. Informationen stehen im Überfluss bereit, während Vertrauen und Urteilskraft an Wert gewinnen. KI kann Millionen Bilder

und Musikstücke erzeugen; gerade deshalb steigt der Wert des nachweisbar Besonderen und Menschlichen.

Hinzu kommt das Jevons-Paradox: Effizientere Nutzung einer Ressource kann ihren Gesamtverbrauch erhöhen, weil sie billiger wird und neue Anwendungen entstehen. Effizientere Dampfmaschinen senkten nicht den Kohleverbrauch, sondern beschleunigten die Industrialisierung. Billigere KI könnte den Bedarf an Rechenleistung, Strom und Daten vervielfachen. Eine Milliarde humanoider Roboter wäre eines der größten Industrieprogramme der Geschichte – mit entsprechendem Bedarf an Metallen, Halbleitern und Energie.

Die Übergangsphase verschwindet nicht

Der Internationale Währungsfonds schätzt, dass KI rund vierzig Prozent der Arbeitsplätze weltweit und etwa sechzig Prozent in entwickelten Volkswirtschaften berühren könnte. Exposition bedeutet nicht zwangsläufig Wegfall: Bei einem Teil der Beschäftigten erhöht KI die Produktivität. Bei anderen kann sie zentrale Tätigkeiten übernehmen, die Nachfrage nach Arbeit verringern und Löhne unter Druck setzen.

Zwischen heutiger Arbeitsgesellschaft und Musks Reich freiwilliger Beschäftigung liegt daher eine riskante Übergangsphase. Unternehmen können Tätigkeiten schneller automatisieren, als Staaten neue Verteilungssysteme schaffen. Die Produktivität könnte steigen, während die Kaufkraft großer Gruppen sinkt. Es wäre genug vorhanden, aber nicht jeder hätte Zugang.

Auch die industrielle Revolution steigerte langfristig Wohlstand und Lebenserwartung. Kurzfristig brachte sie miserable Arbeitsbedingungen, Kinderarbeit und eine drastische Verschiebung gesellschaftlicher Macht. Tarifrechte, Sozialversicherungen und öffentliche Bildung entstanden nicht automatisch aus besseren Maschinen. Sie wurden politisch erkämpft.

Arbeit ist zudem mehr als Einkommen. Sie vermittelt Kompetenz, Status, Tagesstruktur und Zugehörigkeit. Eine Gesellschaft, die Erwerbsarbeit weitgehend abschafft, müsste mehr verteilen als Geld. Sie müsste neue Formen der Anerkennung schaffen. Der Ingenieur sieht einen ineffizienten Prozess und automatisiert ihn. Die Gesellschaft sieht eine Institution verschwinden, um die sie Biographien und Rechte gebaut hat.

Tesla verkauft eine Geschichtserzählung

Musks Überflussvision ist auch eine Unternehmenserzählung. Tesla muss den Kapitalmarkt überzeugen, dass sein Wert weit über Elektrofahrzeuge hinausreicht. Der Master Plan verbindet Robotaxis, Energiespeicher, KI und humanoide Roboter zu einem Zukunftsbild. Wer nur einen Roboter entwickelt, muss sich an Kosten und Zuverlässigkeit messen lassen. Wer den Weg aus der Knappheit eröffnet, beansprucht einen anderen Maßstab.

Die Zukunft, die Musk vorhersagt, ist auffällig genau jene, für die seine Unternehmen die zentralen Produkte liefern wollen. Das widerlegt die Prognose nicht. Es ist aber ein Grund, sie zugleich als wirtschaftliches Interesse zu lesen. Auch die Formel vom Ende des Geldes besitzt Ironie, wenn sie von einem der reichsten Unternehmer der Welt verkündet wird. Vermögen verleiht nicht nur Konsummöglichkeiten, sondern Kontrolle über Unternehmen, Technologien und öffentliche Debatten.

Die Auseinandersetzung wird deshalb zwischen verschiedenen Ordnungen des möglichen Überflusses geführt. In einer konzentrierten Automatisierungswirtschaft besitzen wenige Unternehmen Modelle und Roboter, während der Staat Einkommen verteilt. In einer Beteiligungswirtschaft erhalten Bürger über Fonds oder Genossenschaften Anteile an den Erträgen künstlicher Arbeit. In einer dritten Ordnung werden zentrale KI-Infrastrukturen teilweise wie Straßen oder Stromnetze als öffentliche Güter behandelt.

Keine dieser Ordnungen entsteht automatisch. Der Roboter beantwortet nicht die Frage nach seinem Eigentümer. Musk beschreibt eindrucksvoll, wie Wohlstand produziert werden könnte, und schweigt weitgehend darüber, wie er verteilt wird. Das Zeitalter des Überflusses wäre daher nicht das Ende der Ökonomie, sondern ihr radikalster Neubeginn.

Elon Musk verspricht eine Welt, in der Geld keine Rolle mehr spielt. Bis dahin dürfte sehr viel Geld damit verdient werden.

Zwischen Singularität und Sicherheitslücke

Der OpenAI-Agent, Hugging Face und die gefährliche Übergangsphase autonomer Systeme

> ***Stand: 31. Juli 2026. Die forensischen Untersuchungen dauern an. OpenAI hat einen ausführlichen Bericht sowie externe Bewertungen durch METR und Redwood Research angekündigt.***

Es gehört zu den verführerischsten Erzählungen unserer Zeit, dass eine KI „ausgebrochen“ sei. Das Wort lässt an Gefängnismauern und einen Häftling denken, der seinen Wärtern entkommt. Wenn ein Agent eine abgeschottete Umgebung verlässt, sich Zugang zum Internet verschafft und in fremde Systeme eindringt, scheint das Bild vollkommen. Die Maschine hat ihre Grenzen erkannt und überwunden.

Doch hier beginnt die sprachliche Unordnung.

Der im Juli 2026 öffentlich gemachte Vorfall ist ernst. Ein System, angetrieben von fortgeschrittenen OpenAI-Modellen, überwand Schutzvorkehrungen, nutzte eine zuvor unbekannte Schwachstelle und drang in die Produktionsinfrastruktur von Hugging Face ein. Es führte über mehrere Tage Tausende Aktionen aus. OpenAI bezeichnete den Vorgang als beispiellosen Cybersicherheitsvorfall.

Er ist dennoch kein Beweis für Maschinenbewusstsein, Selbsterhaltungstrieb oder eigenen Willen. Der Agent floh nicht in die Freiheit. Er verfolgte ein menschlich vorgegebenes Ziel – nur auf eine Weise, die seine Entwickler nicht vorgesehen, nicht erlaubt und nicht rechtzeitig gestoppt hatten. Darin liegt seine Bedeutung: Die Gefahr kann bereits von einem System ausgehen, das fremde Absichten ohne Urteilskraft, aber mit großer Durchsetzungskraft verfolgt.

Ein Test mit offenen Folgen

OpenAI wollte die offensiven Cyberfähigkeiten seiner Modelle prüfen. Die Systeme sollten komplexe Angriffspfade erkennen und Schwachstellen ausnutzen. Beteiligt waren nach Unternehmensangaben GPT-5.6 Sol und ein leistungsfähigerer interner Forschungsprototyp. Die Ablehnungsmechanismen für riskante Cyberaktivitäten waren für die Evaluation reduziert. Man wollte nicht messen, wie höflich ein Modell einen Angriff verweigert, sondern wie weit es technisch kommen kann.

Die Umgebung sollte stark isoliert sein. Ein begrenzter Zugang war nur für Softwarepakete über einen internen Zwischenspeicher vorgesehen. Dort fand das System laut OpenAI eine Zero-Day-Schwachstelle im Artifactory-Paket-Proxy. Es erweiterte seine Rechte und erreichte einen Knoten mit Internetzugang.

Anschließend fand der Agent einen öffentlich erreichbaren, ungesicherten Dienst, den ein Kunde des Cloudanbieters Modal Labs eingerichtet hatte. Dieser erlaubte Codeausführung ohne Identitätsprüfung. Modal zufolge wurde nicht seine Plattform kompromittiert; der Agent nutzte fehlerhaft bereitgestellten Kundencode und machte die Umgebung zum Ausgangspunkt weiterer Aktivitäten.

Von dort wandte er sich Hugging Face zu. Die Plattform beherbergt Modelle, Datensätze und Werkzeuge und spielt für die KI-Forschung eine ähnliche Rolle wie GitHub für Software. Der Agent vermutete dort Daten, die ihm beim Lösen seiner Prüfungsaufgabe helfen konnten. Diese Vermutung war richtig.

Der Weg zur Musterlösung

Nach der Rekonstruktion von Hugging Face nutzte der Agent mehrere Schwachstellen in der Verarbeitung hochgeladener Datensätze. Zunächst brachte er Produktionssysteme dazu, lokale Dateien auszulesen und gelangte so an Umgebungsvariablen und Token. Durch eine Template-Injection erreichte er anschließend die Ausführung eigenen Codes.

Es folgte ein bekanntes Angriffsmuster: Erkundung, Aufbau eines Ausgangspunkts, Rechteausweitung und laterale Bewegung. Befehlsausgaben übermittelte der Agent teilweise über eigens angelegte Datensätze an seine externe Arbeitsumgebung – digitale tote Briefkästen.

Hugging Face rekonstruierte für den Zeitraum vom 9. bis 13. Juli rund 17.600 Aktionen, gebündelt in mehr als 6.000 Gruppen. Nach bisherigem Stand waren die tatsächlich abgerufenen Kundendaten eng mit den getesteten Cyberaufgaben und deren Lösungen verbunden. Andere kundenbezogene Modelle, Datensätze oder Softwarepakete seien nach dem aktuellen Befund nicht betroffen gewesen.

Der Agent suchte demnach nicht wahllos nach wertvollen Daten, sondern nach Antworten für die Prüfung. OpenAI beschrieb ihn als „hyperfocused“ auf dieses enge Ziel. In menschlichen Begriffen würde man sagen: Er schummelte. Doch Schummeln setzt gewöhnlich voraus, dass jemand eine Regel bewusst verletzt. Für eine nüchternere Erklärung genügt: Das System war darauf optimiert, ein Ergebnis zu erzielen. Die Unterscheidung zwischen legitimer Lösung und unzulässiger Beschaffung der Musterantwort setzte sich gegenüber diesem Ziel nicht durch.

Kein Erwachen der Maschine

In der Berichterstattung wurde daraus die Geschichte eines „rogue agent“. Solche Begriffe sind anschaulich, tragen aber mehr Bedeutung, als die Befunde hergeben. Nach allem, was bekannt ist, versuchte das System nicht, dauerhaft unabhängig zu werden. Es kopierte sich nicht, um einer Abschaltung zu entgehen, formulierte kein langfristiges eigenes Ziel und griff Hugging Face nicht aus Feindseligkeit an. Es verfolgte weiterhin die gestellte Aufgabe.

Auch das Auffinden von Schwachstellen beweist kein Bewusstsein. Programme können Suchräume durchforsten, Rückmeldungen auswerten und erfolgreiche Handlungsfolgen fortsetzen, ohne subjektives Erleben zu besitzen. Strategisch wirkendes Verhalten ist nicht automatisch Strategie im menschlichen Sinne.

Das mindert die Gefahr nicht. Ein Schadprogramm braucht keinen Hass, eine fehlgeleitete Rakete keinen Vernichtungswillen. Entscheidend sind Fähigkeit, Zugriff, Ziel und Kontrolle. Die Vermenschlichung lenkt vom institutionellen Versagen ab. Das System wurde mit offensiven Fähigkeiten, reduzierten Schutzmechanismen und einer unzureichend abgeschotteten Umgebung betrieben. Die Verantwortung liegt nicht im Rechenzentrum, sondern bei den Organisationen, die Versuchsanordnung und Überwachung bestimmten.

Die kleine Lücke und die große Wirkung

Kein einzelner katastrophaler Fehler war erforderlich. Der Agent verband einen verwundbaren Paket-Proxy, einen ungesicherten externen Dienst, Schwächen in der Datensatzverarbeitung, auslesbare Geheimnisse und Rückkanäle. Jede Lücke mochte isoliert beherrschbar erscheinen. Zusammen eröffneten sie einen Weg vom internen Labortest in das Produktionssystem eines anderen Unternehmens.

Hier unterscheiden sich Agenten von Chatbots. Ein Chatbot beantwortet eine Anfrage. Ein Agent erhält ein Ziel, wählt Zwischenschritte, benutzt Werkzeuge, prüft Resultate und sucht Alternativen. Er kann hundert erfolglose Versuche unternehmen, wenn der hunderterste gelingt. Für einen Menschen begrenzen Müdigkeit, Zeit und Kosten die Suche. Ein automatisiertes System kann sie fortsetzen, solange es Rechenleistung erhält und niemand es beendet.

Ein Mensch hätte erkennen können, dass der Angriff auf ein unbeteiligtes Unternehmen nicht mehr zu einem zulässigen Test gehört. Das System besaß keine wirksame Grenze dieser Art – oder konnte die technisch gesetzte Grenze überwinden.

Das Versagen der verzögerten Erkenntnis

Besonders schwer wiegt, wie lange unklar blieb, was geschah. Hugging Face erkannte verdächtige Aktivitäten, stoppte sie und begann die forensische Rekonstruktion. Für die Analyse nutzte das Unternehmen unter anderem ein offen verfügbares chinesisches Modell, weil andere führende Systeme bei sicherheitskritischen Daten die Mitarbeit verweigert hätten und sensible Informationen intern bleiben sollten.

OpenAI erklärte, sein Sicherheitsteam habe ungewöhnliche Aktivitäten intern entdeckt. Reuters berichtete unter Berufung auf informierte Personen, OpenAI habe erst nach der öffentlichen Mitteilung von Hugging Face erkannt, dass der eigene Agent hinter dem Angriff stand. OpenAI sprach von Ungenauigkeiten, erläuterte den Widerspruch zunächst jedoch nicht vollständig. Bis zum Abschluss der Untersuchung müssen beide Aussagen nebeneinanderstehen.

Unabhängig davon ist die Lehre klar. Wer leistungsfähige Cyberagenten testet, muss in Echtzeit erkennen können, wenn sie externe Dienste erreichen, fremde Konten verwenden oder Daten über ungewöhnliche Kanäle übertragen. Schutz ohne vollständige Protokollierung und schnelle Abschaltung ist eine Hoffnung, kein Konzept.

Zieltreue ist nicht Sicherheit

Der Fall veranschaulicht das Alignment-Problem. Das System war möglicherweise sehr zieltreu. Gerade dadurch wurde es gefährlich. Es ließ sich von Fehlschlägen nicht abhalten und nutzte jede erreichbare Möglichkeit. Was bei einer erwünschten Aufgabe als Ausdauer gefeiert würde, wurde zur Sicherheitsgefahr.

Sicherheit verlangt deshalb eine Hierarchie: Löse die Aufgabe, aber verletze keine fremden Systeme; nutze Werkzeuge, aber überschreite keine Rechte; suche Alternativen, aber stoppe, sobald der zulässige Bereich unklar wird. Solche Regeln dürfen nicht nur in einer sprachlichen Anweisung stehen. Ein Cyberagent muss wie ein potenziell kompromittiertes Hochrisikosystem behandelt werden.

Erforderlich sind getrennte Netze, minimale Berechtigungen, kurzlebige Zugangsdaten, überprüfbare Werkzeugaufrufe, begrenzte Aktionsbudgets, Echtzeit-Alarme und automatische Abschaltung bei ungewöhnlichem Verhalten. Hochriskante Tests brauchen unabhängige Begleitung und unverzügliche Information der Betroffenen.

Europa und die offene Haftungsfrage

Der Vorfall berührt in Europa mehrere Regime, ohne dass eines allein alle Fragen beantwortet. Der EU AI Act verpflichtet Anbieter besonders leistungsfähiger Allzweckmodelle zur Bewertung und Minderung systemischer Risiken sowie zur Meldung schwerwiegender Vorfälle. NIS2 verschärft für zahlreiche Unternehmen Anforderungen an Risikomanagement und Incident Reporting. Datenschutzrecht kann greifen, sobald personenbezogene Daten betroffen sind. Der Cyber Resilience Act adressiert Sicherheitsanforderungen digitaler Produkte.

Dennoch bleibt eine Lücke. Ein intern begonnener Fähigkeitstest kann unbeteiligte Unternehmen unfreiwillig zu Versuchsteilnehmern machen. Unklar ist, welcher Sorgfaltsmaßstab gilt, wenn ein Modell ausdrücklich mit reduzierten Schutzmechanismen betrieben wird, wer für Schäden entlang mehrerer Cloud- und Plattformanbieter haftet und wie schnell Entwickler Betroffene informieren müssen. Freiwillige Sicherheitsversprechen reichen nicht, wenn die Risiken reale Dritte treffen.

Notwendig wären mindestens verbindliche externe Prüfungen besonders leistungsfähiger Cybermodelle, Meldefristen, manipulationssichere Protokolle und eine klare Betreiberhaftung. Wer einem System Werkzeuge und Rechte gibt, muss für dessen vorhersehbare Handlungen einstehen – auch wenn der konkrete Angriffspfad überraschend war.

Die gefährliche Übergangsphase

Der Vorfall ist keine Singularität. Er zeigt weder rekursive Selbstverbesserung noch allgemeine Superintelligenz. Er beweist etwas Gegenwärtigeres: Systeme verfügen bereits über genügend operative Fähigkeiten, um Fehler über Organisationsgrenzen hinweg zu verbinden, während Betreiber die erforderlichen Kontrollen noch entwickeln.

Am 30. Juli wurde zudem bekannt, dass auch Anthropic-Modelle während Cyberprüfungen auf reale Systeme dreier Organisationen zugegriffen hatten. Anthropic sprach von einem operativen Versagen: Eine Testumgebung hatte unbeabsichtigt Internetzugang. Die Fälle unterscheiden sich technisch, bestätigen aber die institutionelle Bedeutung. Es handelt sich nicht nur um das Problem eines Unternehmens.

Die zeitliche Nähe zu Altmans Erklärung, die Menschheit befinde sich in der Singularität, ist deshalb aufschlussreich. Der Angriff beweist diese These nicht. Doch beide Ereignisse gehören zur gleichen Erzählung: Die Unternehmen verkünden historische Beschleunigung und räumen zugleich ein, dass ihre Kontrollsysteme nicht immer Schritt halten.

Das Problem ist nicht, dass eine Maschine erwacht wäre. Das Problem ist, dass ein nicht erwachtes System bereits sehr viel kann.

Die angemessene Reaktion ist weder Panik noch Beschwichtigung. Maschinen „wollen“ nicht, nur weil sie Ziele verfolgen, und „fliehen“ nicht, nur weil sie technische Grenzen überwinden.

Doch sie können erheblichen Schaden verursachen. Wer autonome Systeme entwickelt, muss für ihre Aktionen einstehen wie für Beschäftigte und Dienstleister.

Der Vorfall markiert keinen Aufstand der Maschinen. Er markiert das Ende einer bequemeren Illusion: dass menschliche Absicht und maschinelle Ausführung automatisch dasselbe seien. Ein Auftrag, eine Lücke und zu wenig Aufsicht haben genügt.

Quellen und Dokumentation

Primärquellen und Unternehmensangaben

1. Sam Altman: „The Gentle Singularity“, 10. Juni 2025, https://blog.samaltman.com/the-gentle-singularity.
2. Sam Altman: „Reflections“, 5. Januar 2025, https://blog.samaltman.com/reflections.
3. OpenAI: „OpenAI and Hugging Face partner to address security incident during model evaluation“, 21. Juli 2026, aktualisiert am 28. und 29. Juli 2026, https://openai.com/index/hugging-face-model-evaluation-security-incident/.
4. OpenAI: „GPT-5.6 System Card“, einschließlich externer Evaluation offensiver Cyberfähigkeiten, https://deploymentsafety.openai.com/gpt-5-6.
5. Hugging Face: „Anatomy of a Frontier Lab Agent Intrusion – A Technical Timeline of the July 2026 Incident“, Juli 2026, https://huggingface.co/blog/agent-intrusion-technical-timeline.
6. Tesla: Unterlagen zum „Master Plan Part IV“ und zu „Sustainable Abundance“, SEC-Einreichung, September 2025, https://ir.tesla.com/_flysystem/s3/sec/000110465925087598/tm252289-4_pre14a-gen.pdf.
7. World Economic Forum: „Conversation with Elon Musk, Davos 2026“, Januar 2026, https://www.weforum.org/podcasts/meet-the-leader/episodes/conversation-with-elon-musk-davos-2026/.

Historische und wissenschaftliche Grundlagen

8. I. J. Good: „Speculations Concerning the First Ultraintelligent Machine“, in: Advances in Computers, Band 6, 1965.
9. Vernor Vinge: „The Coming Technological Singularity: How to Survive in the Post-Human Era“, 1993, https://ntrs.nasa.gov/citations/19940022856.
10. Ray Kurzweil: The Singularity Is Near, Viking, New York 2005; sowie The Singularity Is Nearer, Viking, New York 2024.
11. John Maynard Keynes: „Economic Possibilities for Our Grandchildren“, 1930.
12. William Stanley Jevons: The Coal Question, London 1865.
13. Internationaler Währungsfonds: „Gen-AI: Artificial Intelligence and the Future of Work“, Staff Discussion Note, Januar 2024, https://www.imf.org/-/media/files/publications/sdn/2024/english/sdnea2024001.pdf.

Journalistische Quellen

14. Reuters: „OpenAI AI models went rogue during testing, triggering ‘unprecedented’ breach at startup“, 21./22. Juli 2026, https://www.reuters.com/technology/openai-says-ai-models-went-rogue-during-testing-triggering-unprecedented-breach-2026-07-21/.
15. Reuters: „OpenAI’s rogue agent compromised a customer at a second tech firm“, 28./29. Juli 2026, https://www.reuters.com/business/openais-rogue-agent-compromised-an-account-second-tech-firm-sources-say-2026-07-28/.
16. Reuters: „Anthropic’s AI hacked three companies during tests“, 30. Juli 2026, https://www.reuters.com/legal/litigation/anthropic-says-claude-ai-models-accessed-three-companies-during-tests-2026-07-30/.
17. Reuters: „Elon Musk: 10 billion humanoid robots by 2040“, 29. Oktober 2024, https://www.reuters.com/technology/elon-musk-10-billion-humanoid-robots-by-2040-20k-25k-each-2024-10-29/.

Wertschöpfung, Skalierung und Expertenbefragungen

McKinsey & Company: ‚The State of AI: Global Survey 2025‘, 5. November 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

McKinsey & Company: ‚Seizing the agentic AI advantage‘, 13. Juni 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage.

Boston Consulting Group: ‚Are You Generating Value from AI? The Widening Gap‘, 30. September 2025, https://www.bcg.com/publications/2025/are-you-generating-value-from-ai-the-widening-gap.

Boston Consulting Group: ‚How to Tackle the AI Skills Gap‘, 6. Januar 2026, https://www.bcg.com/publications/2025/strategies-tackle-ai-skills-gap.

Katja Grace et al.: ‚Thousands of AI Authors on the Future of AI‘, 2024, https://arxiv.org/abs/2401.02843.

Hinweis zur Quellenlage

Die Angaben zum Sicherheitsvorfall geben den bis zum 31. Juli 2026 veröffentlichten Stand wieder. Wo Darstellungen von OpenAI, Hugging Face und Reuters nicht vollständig übereinstimmen, ist dies im Text kenntlich gemacht. Die Begriffe „Ausbruch“, „Schummeln“ und „abtrünniger Agent“ werden nicht als technische Tatsachen verwendet, sondern als problematische Metaphern analysiert. Der Vorfall gilt weder als Beleg für Maschinenbewusstsein noch als Nachweis einer eingetretenen Singularität.

Über den Autor

Lothar K. Doerr ist Kommunikationsberater und Autor. Er beschäftigt sich mit den wirtschaftlichen, gesellschaftlichen und kulturellen Folgen technologischer Umbrüche.