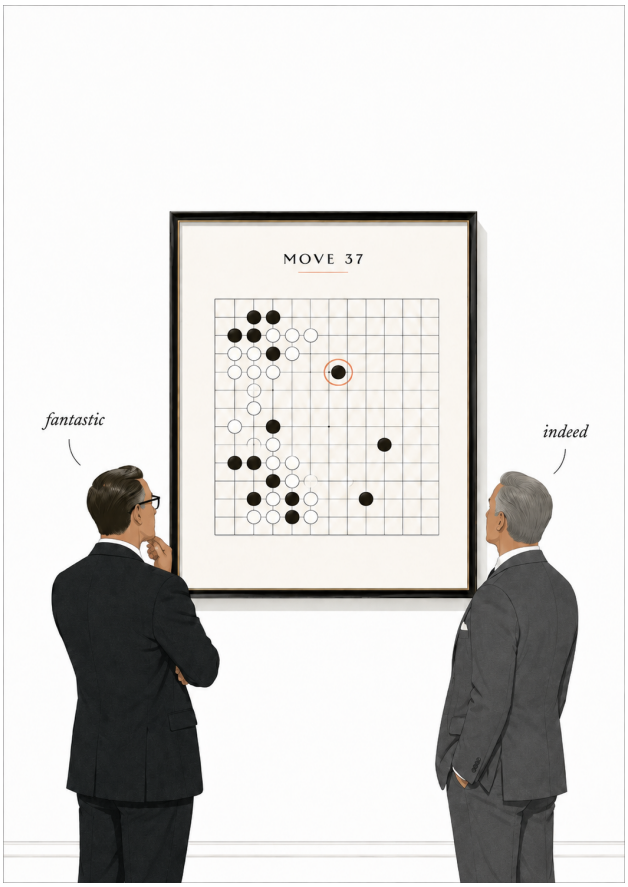
ZUG 37
ENJOY THE RIDE

Wenn die Maschine recht hat – über AlphaGo, Bill Gates, Elon Musk und die Grenzen menschlicher Kontrolle



Autor: Lothar Dörr
Institut für Produktionserhaltung · infpro

Titelillustration: KI-generierte redaktionelle Visualisierung. Die Go-Stellung ist symbolisch und keine historische Rekonstruktion der Partie.

Der schwarze Stein

Warum ein Go-Zug aus dem Jahr 2016 heute zur Managementfrage wird.

Es gibt Augenblicke der Technikgeschichte, deren Bedeutung sich in einer Zahl ausdrücken lässt, und andere, die ein Bild brauchen. Zug 37 gehört zur zweiten Kategorie. Am 10. März 2016 legte im Four Seasons Hotel in Seoul ein Mitarbeiter von Google DeepMind einen schwarzen Stein auf ein Go-Brett. Er setzte ihn im Auftrag eines Computerprogramms, AlphaGo, in der zweiten Partie gegen Lee Sedol, einen der stärksten Go-Spieler seiner Generation. Der Stein lag an einer Stelle, die nach der über Jahrhunderte verdichteten Erfahrung des Spiels ungewöhnlich wirkte. Kommentatoren suchten nach dem Sinn. Lee Sedol nahm sich ungewöhnlich viel Zeit. Später beschrieb er den Zug als einen Moment, in dem er seine Vorstellung von der Maschine ändern musste. AlphaGo gewann die Partie und am Ende das Match mit vier zu eins. Der Sieg war spektakulär; Zug 37 war bedeutsamer.

Denn ein Computer, der einen Menschen durch höhere Rechengeschwindigkeit besiegt, bestätigt zunächst nur eine alte Erfahrung: Maschinen sind gut darin, große Mengen von Operationen auszuführen. Zug 37 zeigte etwas anderes. AlphaGo wählte eine starke Lösung, die außerhalb des Erwartungsraums lag, aus dem es selbst zunächst gelernt hatte. DeepMind gibt für den Zug die berühmte Größenordnung von eins zu zehntausend an: nicht als Erfolgswahrscheinlichkeit des Zuges, sondern als Maß dafür, wie unwahrscheinlich ein solcher Zug nach dem aus menschlichen Partien gelernten Spielmuster war. Die Maschine kannte also die Konvention, aber sie war ihr nicht verpflichtet. Genau dieser Unterschied macht aus einer Sportgeschichte eine Geschichte über Wissen.

Menschen organisieren Erfahrung, indem sie erfolgreiche Entscheidungen wiederholen, Regeln formulieren und daraus Lehrmeinungen, Standards und Routinen machen. Das ist keine Schwäche, sondern eine Voraussetzung organisierter Zivilisation. Ein Chirurg beginnt nicht jede Operation bei null, ein Ingenieur berechnet nicht jede Brücke ohne Normen neu, ein Produktionsleiter ignoriert nicht das Wissen der vergangenen Jahre. Doch derselbe Mechanismus, der Erfahrung nutzbar macht, erzeugt mit der Zeit einen Suchpfad. Was sich bewährt hat, wird häufiger geprüft; was als unvernünftig gilt, wird seltener untersucht. Aus Erfahrung wird Gewissheit, aus Gewissheit Konvention und aus Konvention gelegentlich eine Grenze dessen, was überhaupt noch als Möglichkeit erscheint. Zug 37 traf genau diese Grenze.

Die entscheidende Frage dieses Dossiers lautet deshalb nicht, ob Maschinen irgendwann Bewusstsein entwickeln oder ob eine sogenannte allgemeine künstliche Intelligenz im Jahr 2030, 2040 oder nie entsteht. Sie ist näher an der betrieblichen Wirklichkeit und zugleich unangenehmer: Was geschieht, wenn ein technisches System in einer konkreten Entscheidung recht hat und der verantwortliche Mensch zunächst nicht versteht, warum? Das kann auf einem Go-Brett geschehen; es kann ebenso in einer Produktionsplanung, einer Materialauswahl, einer Diagnose, einer Investitionsentscheidung oder einer Lieferkettensteuerung geschehen. Je besser die Systeme werden, desto weniger geht es darum, ob sie menschliche Arbeit nur beschleunigen. Dann beginnt die Frage, ob menschliche Erfahrung ihre Stellung als letzte Instanz der Entscheidung behält.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Bill Gates hat diese Verschiebung Ende August 2026 in einer bemerkenswert scharfen Form beschrieben. Künstliche Intelligenz könne, so Gates, erstmals menschliche Kognition ersetzen und übertreffen; die gesellschaftlichen Institutionen seien auf die Geschwindigkeit dieser Entwicklung nicht vorbereitet. Elon Musk formuliert die andere Seite desselben Problems beinahe fatalistisch. Er sehe keinen realistischen Weg, die Dynamik von KI und Robotik aufzuhalten. Als die Chefredakteurin des Economist seine Haltung mit den Worten zusammenfasste, man könne wohl nur auf das Beste hoffen und sich auf die Fahrt begeben, stimmte Musk zu und sagte: „Let's enjoy the ride." Deutsche KI-Forscher wiederum warnen zu Recht davor, aus den beeindruckenden Fähigkeiten heutiger Systeme eine gerade Linie zur allmächtigen Superintelligenz zu ziehen. Zwischen diesen Positionen liegt der Raum, in dem Unternehmen handeln müssen.

infpro setzt deshalb früher an. Für den industriellen Mittelstand ist nicht entscheidend, wann eine Maschine alles besser kann als ein Mensch. Die Machtverschiebung beginnt, sobald sie in einer wichtigen Einzelentscheidung systematisch besser ist als derjenige, der dafür verantwortlich bleibt. Ein System muss weder Bewusstsein besitzen noch ein Unternehmen als Ganzes verstehen, um bei Wartung, Qualitätsprüfung, Beschaffung oder Produktionsplanung einen Vorsprung zu gewinnen. Genau dann verändert sich Führung: Die Maschine kann die bessere Empfehlung erzeugen, Verantwortung kann sie dennoch nicht übernehmen. Der Manager der Zukunft muss deshalb möglicherweise für Entscheidungen einstehen, die er selbst nicht hätte finden können. Das ist keine Science-Fiction. Es ist eine neue Architektur von Wissen, Urteil und Verantwortung.

Die eigentliche KI-Frage beginnt nicht dort, wo die Maschine alles besser kann. Sie beginnt dort, wo sie in einer wichtigen Entscheidung recht hat und der Mensch zunächst nicht versteht, warum.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

1. Zug 37 – der Augenblick, in dem Erfahrung ihre Unschuld verlor

AlphaGo zeigte nicht, dass Maschinen menschlich denken. Es zeigte, dass sie ohne menschliche Konvention zu starken Lösungen gelangen können.

Go besitzt eine besondere Stellung in der Geschichte der künstlichen Intelligenz. Schach war schon früh ein geeignetes Feld für Computer: Die Regeln sind eindeutig, die Zahl der legalen Züge pro Stellung ist begrenzt, und jahrzehntelang ließ sich Fortschritt vor allem durch bessere Suche, schnellere Hardware und ausgefeiltere Bewertungsfunktionen erzielen. Go widersetzte sich diesem Ansatz. Auf dem 19-mal-19-Brett ist die Zahl möglicher Stellungen so gewaltig, dass eine vollständige Suche ausscheidet. Zudem hängt die Qualität eines Zuges häufig nicht an einem unmittelbaren materiellen Gewinn, sondern an Einfluss, Form, Balance, langfristigen Beziehungen und der Fähigkeit, lokale Kämpfe mit dem Zustand des gesamten Brettes zu verbinden. Gerade deshalb galt Go lange als Spiel, in dem die schwer fassbare menschliche Intuition einen Vorsprung behalten würde.

DeepMind löste das Problem nicht, indem es versuchte, jeden denkbaren Zug vollständig durchzurechnen. AlphaGo kombinierte tiefe neuronale Netze mit einer Monte-Carlo-Baumsuche. Ein sogenanntes Policy Network half dabei, aussichtsreiche Züge auszuwählen; ein Value Network sollte abschätzen, wie wahrscheinlich eine Stellung zum Sieg führt. Die erste erfolgreiche Generation des Systems lernte zunächst aus Partien starker menschlicher Spieler. Anschließend spielte AlphaGo große Mengen von Partien gegen Versionen seiner selbst und verbesserte sich durch Verstärkungslernen. Diese Kombination ist für die Interpretation von Zug 37 entscheidend. AlphaGo war kein Wesen, das ohne Vorgeschichte auf ein Brett blickte. Es hatte die menschliche Go-Kultur in statistischer Form kennengelernt. Aber durch Selbstspiel konnte es Varianten prüfen, die in dieser Kultur nur selten oder gar nicht vorkamen.

Genau deshalb ist die oft wiederholte Zahl eins zu zehntausend so wichtig und so häufig missverstanden. Sie bedeutet nicht, AlphaGo habe an diesem Punkt einen riskanten Zug mit einer Erfolgschance von 0,01 Prozent gespielt. DeepMind beschreibt damit, wie ungewöhnlich der Zug im Rahmen des menschlich gelernten Spielmusters war. Das System, das vorherzusagen gelernt hatte, welche Züge starke Menschen typischerweise wählen, hätte einen solchen Zug fast nie erwartet. Das Gesamtsystem bewertete ihn dennoch als stark. Darin steckt eine Trennung, die für lernende Maschinen grundlegend ist: Nachahmung kann Ausgangspunkt sein, muss aber nicht Endpunkt bleiben.

Für einen Menschen ist dieser Schritt schwieriger. Expertise besteht gerade darin, unzählige Erfahrungen zu verdichten. Ein erfahrener Go-Spieler erkennt gute und schlechte Form, ohne jede Variante bewusst zu berechnen. Ein Produktionsleiter hört an einer Anlage eine Veränderung, die ein Anfänger nicht bemerkt. Ein Arzt erkennt in einer Kombination von Symptomen ein Muster, das sich nicht vollständig in eine Checkliste übersetzen lässt. Solche Intuition ist kein Gegensatz zu Rationalität; sie ist komprimierte Erfahrung. Doch komprimierte Erfahrung besitzt einen Preis: Sie macht

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

bestimmte Möglichkeiten leichter sichtbar und andere unsichtbarer. Je erfolgreicher eine Schule, ein Unternehmen oder eine Profession mit ihren Heuristiken war, desto größer kann die Versuchung werden, sie mit der Struktur der Wirklichkeit selbst zu verwechseln.

Zug 37 war deshalb kein Beweis dafür, dass AlphaGo „kreativ" im menschlichen Sinne sein musste. Ob eine Maschine Schönheit empfindet, Absichten erlebt oder einen Einfall als Einfall versteht, lässt sich aus einem Go-Zug nicht ableiten. Für die Wirkung des Ergebnisses ist diese philosophische Frage jedoch nicht entscheidend. Menschen erlebten den Zug als kreativ, weil er überraschend, funktional und stark war. Fan Hui, der europäische Meister, den AlphaGo bereits 2015 besiegt hatte, beschrieb die ungewöhnlichen Züge des Systems mit Bewunderung. Lee Sedol selbst sagte später, er habe AlphaGo vor diesem Zug als Maschine der Wahrscheinlichkeit betrachtet; danach sei seine Einschätzung eine andere gewesen. Der Mensch verlieh dem Ergebnis Bedeutung, auch wenn das System selbst keine menschliche Bedeutungserfahrung brauchte.

Das verändert die klassische Vorstellung vom Werkzeug. Ein Hammer erweitert die menschliche Kraft. Ein Taschenrechner erweitert Rechengeschwindigkeit und Genauigkeit. Beide bleiben innerhalb eines durch Menschen definierten Verfahrens. Lernende Systeme können dagegen eine andere Rolle einnehmen: Sie erzeugen Vorschläge, die nicht als explizite menschliche Regel in sie hineingeschrieben wurden. Das ist nicht dasselbe wie Autonomie im politischen oder moralischen Sinn; es ist aber eine epistemische Verschiebung. Das Werkzeug produziert nicht nur schneller, was wir bereits kennen. Es kann in einem eng definierten Raum etwas finden, das wir nicht kannten.

Gerade deshalb ist AlphaGo heute noch interessant, obwohl das Programm selbst längst nicht mehr im Wettkampf eingesetzt wird. DeepMind zog AlphaGo 2017 nach dem Match gegen den damaligen Weltranglistenersten Ke Jie aus dem kompetitiven Go zurück. Die Forschung ging weiter: AlphaGo Zero lernte Go ohne menschliche Partien allein durch Selbstspiel; AlphaZero verallgemeinerte den Ansatz auf Go, Schach und Shogi. Damit wurde die Pointe von Zug 37 sogar verschärft. In Seoul hatte das System menschliche Tradition gekannt und überschritten. AlphaGo Zero zeigte kurz darauf, dass ein System eine solche Tradition für das Erlernen des Spiels nicht einmal mehr zwingend als Ausgangspunkt benötigte.

Für die heutige KI-Debatte ist diese Entwicklung ein nützlicher Schutz gegen zwei gleich große Irrtümer. Der erste besteht darin, jede überraschende Maschinenleistung zu vermenschlichen und daraus Bewusstsein, Absicht oder allgemeine Intelligenz abzuleiten. Der zweite besteht darin, eine Maschine nur dann ernst zu nehmen, wenn sie sich wie ein Mensch verhält. AlphaGo war ein hochspezialisiertes System für ein Brettspiel. Es konnte keine E-Mail schreiben, kein Unternehmen führen und keine politische Entscheidung treffen. Innerhalb seines Problemraums war es dennoch in der Lage, das stärkste menschliche Wissen zu überschreiten. Genau darin liegt die Verbindung zur Industrie: Ein System braucht keine allgemeine Intelligenz, um in einer eng definierten Aufgabe eine relevante Autoritätsverschiebung auszulösen.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

INFO · Was war AlphaGo – und was gibt es heute Vergleichbares?

AlphaGo war ein von Google DeepMind entwickeltes, hochspezialisiertes KI-System für das Brettspiel Go. Es war kein Sprachmodell, kein Chatbot und keine allgemeine künstliche Intelligenz. Seine Stärke entstand aus der Verbindung tiefer neuronaler Netze mit einer Monte-Carlo-Baumsuche. Ein Policy Network priorisierte aussichtsreiche Züge, ein Value Network bewertete Stellungen. Die frühe Version lernte zunächst aus menschlichen Meisterpartien und verbesserte sich anschließend durch Selbstspiel.

Die wichtigsten Stationen: 2015 schlug AlphaGo den Europameister Fan Hui mit 5:0; im März 2016 gewann es gegen Lee Sedol mit 4:1; im Mai 2017 besiegte es den damaligen Weltranglistenersten Ke Jie in einem Drei-Partien-Match. Danach zog DeepMind AlphaGo aus dem Wettkampf zurück. Noch 2017 folgte AlphaGo Zero, das ohne menschliche Go-Daten allein durch Selbstspiel lernte. AlphaZero übertrug denselben Grundgedanken auf Go, Schach und Shogi.

Heute existieren für Go mehrere Systeme auf übermenschlichem Niveau. Besonders relevant ist KataGo: ein offen verfügbares, weiterhin trainiertes Go-Programm, das einen AlphaZero-ähnlichen Selbstlernansatz mit zahlreichen Verbesserungen verbindet und 2026 zu den stärksten Open-Source-Go-Bots zählt. Es wird von Profis und Amateuren nicht nur als Gegner, sondern vor allem als Analysewerkzeug benutzt. AlphaGo selbst ist also Geschichte; die maschinelle Spielstärke, die es 2016 sensationell machte, ist heute in der Go-Welt eher Ausgangspunkt als Endpunkt.

Die größere Linie führt über Spiele hinaus. DeepMind betrachtet AlphaGo, AlphaGo Zero und AlphaZero als wichtige Stationen für Verstärkungslernen, Planung und Systeme, die durch Suche neue Lösungen entdecken. Das ist keine direkte technische Abstammungslinie zu jedem heutigen KI-System und schon gar nicht zu jedem Sprachmodell. Aber die Grundidee – Lernen, bewerten, Varianten selbst erzeugen und dadurch menschlich ungewohnte Lösungen finden – ist zu einem zentralen Motiv moderner KI-Forschung geworden.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

2. Bill Gates – wenn Kognition selbst automatisiert wird

Warum eine Warnung vom 26. August 2026 Zug 37 eine neue Bedeutung gibt.

Bill Gates hat über künstliche Intelligenz lange in jener Mischung aus Technikoptimismus und gesellschaftlicher Vorsicht gesprochen, die seine zweite Karriere als Philanthrop prägt. Am 26. August 2026 verschob er den Ton deutlich. In einem fast 6.000 Wörter langen Essay und einem Interview mit der New York Times erklärte er, die Risiken der gegenwärtigen Entwicklung würden von Teilen der Technologiebranche heruntergespielt, weil wirtschaftlich zu viel auf dem Spiel stehe. Gates spricht von einer der turbulentesten Übergangsphasen der Menschheitsgeschichte und setzt den entscheidenden Unterschied zu früheren technischen Revolutionen an einer Stelle, die für unser Dossier zentral ist: Künstliche Intelligenz könne erstmals menschliche Kognition ersetzen und sogar übertreffen.

Damit ist mehr gemeint als Automatisierung. Die Industrialisierung ersetzte Muskelkraft, mechanisierte Abläufe und verlagerte Arbeit. Computer automatisierten Berechnung, Speicherung und einen wachsenden Teil der Informationsverarbeitung. Dennoch blieb ein gesellschaftliches Grundmuster vergleichsweise stabil: Neue Maschinen wurden in Prozesse eingebaut, deren Ziele, Regeln und Bewertungskriterien Menschen vorgaben. Gates argumentiert, dass diese Analogie bei KI an ihre Grenze geraten kann, weil Systeme nicht nur die Ausführung übernehmen, sondern zunehmend auch Teile jener Arbeit, mit der Menschen Probleme strukturieren, Optionen vergleichen, Texte und Programme erzeugen, Diagnosen vorbereiten und Entscheidungen begründen.

Das ist zunächst eine Arbeitsmarktthese. Gates warnt vor dem dauerhaften Verschwinden zahlreicher Einstiegs- und mittlerer Tätigkeiten und davor, dass die Anpassung wesentlich schneller verlaufen könnte als frühere Verschiebungen zwischen Landwirtschaft, Industrie und Dienstleistungen. Besonders bemerkenswert ist jedoch, dass er die Grenze nicht beim Arbeitsplatz zieht. Seine Frage betrifft die gesellschaftliche Steuerungsfähigkeit. Wenn Systeme immer verlässlicher werden und in immer mehr Bereichen ohne permanente menschliche Kontrolle arbeiten können, entsteht ein Zeitpunkt, an dem Unternehmen einen starken ökonomischen Anreiz haben, den Menschen tatsächlich aus der laufenden Prüfung herauszunehmen. Je näher die Maschine an nahezu fehlerfreie Arbeit kommt, desto teurer erscheint der Mensch als Kontrollschleife.

An diesem Punkt berührt Gates Zug 37. Solange eine Maschine nur schneller ausrechnet, was der Mensch prinzipiell selbst nachvollziehen kann, bleibt menschliche Kontrolle intellektuell plausibel. Der Mensch mag langsamer sein, aber er besitzt einen zweiten Weg zur Überprüfung. Schwieriger wird es, wenn der Erkenntniswert der Maschine gerade darin liegt, eine Lösung zu finden, die außerhalb menschlicher Erfahrung liegt. Auf dem Go-Brett war das eine ästhetische Sensation. In der Medizin, Chemie, Produktion oder Infrastruktur kann dasselbe Muster einen fundamentalen Konflikt erzeugen: Je wertvoller die maschinelle Abweichung von der Konvention, desto geringer kann die Fähigkeit des Menschen sein, sie unmittelbar zu beurteilen.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Man kann diesen Konflikt am Beispiel eines neuen Wirkstoffs beschreiben. Ein System schlägt eine Molekülstruktur vor, die kein Forscher gewählt hätte. Das ist zunächst genau der Erfolg, den man von leistungsfähiger KI erwartet: Sie erweitert den menschlichen Suchraum. Die Struktur muss dennoch experimentell validiert werden. Das Ergebnis der Maschine besitzt also keinen Freibrief, sondern erzeugt eine neue Prüfaufgabe. In einem anderen Feld kann die Verifikation schwieriger sein. Eine strategische Unternehmensentscheidung entfaltet ihre Folgen über Jahre; eine militärische Empfehlung kann nicht risikolos getestet werden; eine Entscheidung über ein Stromnetz muss unter Zeitdruck fallen. Die technische Überlegenheit des Modells und die gesellschaftliche Kontrollfähigkeit sind deshalb zwei verschiedene Größen.

Gates reagiert darauf mit einer für einen Technologen ungewöhnlich politischen Forderung. Er spricht von neuen nationalen und internationalen Institutionen, von einer möglichen Besteuerung von Robotern oder KI-Nutzung und von Tätigkeiten, die bewusst „Human Reserved“ bleiben sollten. Sein Beispiel ist bewusst nicht technisch: Ein Roboter könnte einem Patienten womöglich irgendwann korrekt mitteilen, dass seine Krankheit unheilbar ist; trotzdem solle eine Gesellschaft entscheiden dürfen, dass dies eine menschliche Aufgabe bleibt. Darin steckt ein Prinzip, das über den Arbeitsmarkt hinausweist. Nicht alles, was eine Maschine technisch übernehmen kann, muss eine Gesellschaft ihr auch überlassen.

Für Unternehmen lässt sich daraus ein ähnlicher Gedanke ableiten: Es könnte künftig nicht nur Human-Reserved Jobs, sondern Human-Reserved Decisions geben. Entscheidungen, bei denen KI analysieren, simulieren, Optionen erzeugen und sogar eine klare Empfehlung abgeben darf, deren formale und materielle Legitimation aber bei einem Menschen oder einem klar definierten Gremium verbleibt. Das ist keine romantische Rettung menschlicher Überlegenheit. Im Gegenteil: Ein solcher Vorbehalt wird gerade dann wichtig, wenn die Maschine in der Analyse überlegen sein kann. Verantwortung ist nicht identisch mit Optimierung.

Gates geht sogar so weit zu sagen, er würde einen glaubwürdigen globalen Plan zur Verlangsamung der KI-Fortschritte wahrscheinlich unterstützen. Zugleich hält er genau einen solchen Plan wegen der geopolitischen und wirtschaftlichen Anreize für kaum realistisch. Das macht seine Position härter, nicht weicher. Wenn die Entwicklung nicht zuverlässig gebremst werden kann, verschiebt sich die Aufgabe von der Hoffnung auf einen globalen Stopp zur Beschleunigung institutioneller Lernfähigkeit. Regulierung, Bildung, Sicherheitsforschung und Unternehmensführung müssten dann schneller reifen als bisher. Das Problem liegt nicht nur in den Fähigkeiten der Modelle, sondern in der Differenz zwischen ihrer Geschwindigkeit und der Geschwindigkeit unserer Institutionen.

Hier muss man Gates allerdings widersprechen, wo aus berechtigter Sorge eine zu lineare Zukunftserzählung entstehen könnte. Dass KI bereits menschliche Kognition in einzelnen Aufgaben ersetzt oder übertrifft, bedeutet nicht, dass jede kognitive Tätigkeit in kurzer Zeit substituiert wird. Systeme können in einem Gebiet übermenschlich und im nächsten erstaunlich fragil sein. Go liefert gerade dafür ein gutes Beispiel. AlphaGo war übermenschlich in Go und zugleich außerhalb des Brettes funktionslos. Die politische Aufgabe besteht deshalb nicht darin, eine abstrakte Einheit „die KI“ zu regulieren, sondern konkrete Systeme, Einsatzfelder, Fähigkeiten und Folgen auseinanderzuhalten.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Trotzdem trifft Gates einen Punkt, den die Debatte um Sprachmodelle leicht verdeckt. Die relevanteste Grenze ist nicht erreicht, wenn eine Maschine wie ein Mensch wirkt. Sie ist erreicht, wenn eine Organisation ihr Urteil aus rationalen Gründen höher gewichtet als das Urteil ihrer Menschen. Dann verändert sich Autorität. Ein Unternehmen kann formal behaupten, der Mensch entscheide weiterhin; wenn der Mensch aber in 98 von 100 Fällen der Empfehlung eines Systems folgt, weil dessen Erfolgsbilanz besser ist, entsteht materiell eine andere Ordnung. Genau diese schleichende Verschiebung ist für infpro wichtiger als die Frage, ob wir sie eines Tages AGI nennen werden.

Eine Maschine muss keine Superintelligenz sein, um Autorität zu gewinnen. Es genügt, dass ihr Urteil so häufig besser ist, dass menschlicher Widerspruch irgendwann wie ein Fehler erscheint.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

3. Elon Musk – „Enjoy the Ride“

Eine Philosophie des Unvermeidlichen – und warum sie als Führungsprinzip nicht genügt.

Elon Musk gehört seit Jahren zu den widersprüchlichsten Stimmen der KI-Debatte. Er investiert in künstliche Intelligenz und Robotik, gründete xAI, war Mitgründer von OpenAI und hat zugleich wiederholt vor extremen Risiken leistungsfähiger Systeme gewarnt. Im Juli 2026 erhielt diese Spannung im Gespräch mit dem Economist eine neue Form. Musk sagte, er halte das Risiko von KI und Robotik weiterhin für real und nicht null. Gleichzeitig sehe er keinen Weg, die enorme Dynamik noch wirklich zu stoppen. Selbst wenn es einen Stoppschalter gäbe, sollte man ihn vielleicht nicht drücken, weil der wahrscheinlichste Ausgang in seinen Augen ein Zeitalter außerordentlichen Überflusses sei.

Die berühmte Formulierung dieses Gesprächs muss präzise wiedergegeben werden. „Let’s hope for the best and go for the ride“ sagte nicht Musk, sondern Zanny Minton Beddoes, die Chefredakteurin des Economist, als sie Musks Haltung zusammenfasste. Musk antwortete: „Yeah, I mean, yes, pretty much“ und fügte hinzu, „Let’s enjoy the ride“ sei inzwischen ungefähr seine Philosophie. Das ist mehr als eine sprachliche Feinheit. Beddoes benennt den Fatalismus, Musk übernimmt ihn. Er sagt nicht, dass alle Risiken gelöst seien. Er sagt, dass die Entwicklung trotz der Risiken als praktisch unaufhaltsam erscheint.

Technologischer Optimismus sieht anders aus. Der klassische Optimist behauptet, eine neue Technik sei beherrschbar; ihre Risiken könnten durch bessere Konstruktion, Regeln und Erfahrung soweit reduziert werden, dass ihr Nutzen überwiegt. Musks Position verbindet dagegen Optimismus über das Ergebnis mit Pessimismus über die Steuerbarkeit des Prozesses. Er erwartet, dass KI und Robotik enorme Produktivität, Überfluss und möglicherweise eine Welt erzeugen, in der Arbeit langfristig optional wird. Zugleich erklärt er, selbst wenn er die Entwicklung stoppen wollte, könne er es nicht. Die Menschheit wäre in diesem Bild weniger der Ingenieur eines kontrollierten Projekts als Passagier eines Wettlaufs, dessen Geschwindigkeit sich aus Konkurrenz, Kapital, Forschung und geopolitischem Druck selbst verstärkt.

Diese Sicht besitzt einen rationalen Kern. Kein einzelnes Unternehmen kann die globale Entwicklung stoppen. Ein Labor, das freiwillig auf ein leistungsfähigeres Modell verzichtet, schafft damit keinen globalen Stillstand; es kann lediglich einem Konkurrenten Vorsprung geben. Ein Staat, der stark verlangsamt, riskiert, dass andere Staaten weiterforschen. Genau deshalb ist KI ein klassisches Koordinationsproblem. Individuell vernünftiges Verhalten – weiterentwickeln, um nicht zurückzufallen – kann kollektiv zu einer Geschwindigkeit führen, die keiner der Beteiligten allein gewählt hätte. In diesem Punkt ist Musks Unvermeidlichkeitsthese ernst zu nehmen.

Aber aus der Feststellung, dass eine Entwicklung schwer aufzuhalten ist, folgt keine Antwort darauf, wie sie geführt werden soll. „Enjoy the Ride“ ist als Lebensphilosophie charmant, als Governance-Prinzip aber unbrauchbar. Je stärker externe Dynamiken sind, desto wichtiger wird interne Ordnung. Ein Unternehmen kann den weltweiten Fortschritt von KI nicht kontrollieren; es kann sehr wohl entscheiden, welche Systeme

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Zugriff auf Produktionsdaten erhalten, welche Agenten Transaktionen auslösen dürfen, wie Entscheidungen protokolliert werden, wer eine Empfehlung freigibt und welche technischen Möglichkeiten zum Abbruch bestehen. Unvermeidlichkeit auf der Makroebene ist gerade kein Argument für Kontrollverzicht auf der Mikroebene.

Musks eigene Prognosen verschärfen diesen Widerspruch. Er erwartet, dass künstliche Intelligenz in wenigen Jahren die Summe menschlicher Intelligenz übertreffen könnte, und benutzt zur Beschreibung der möglichen Hierarchie den Vergleich zwischen Menschen und Schimpansen. Solche Zeitangaben sind hoch spekulativ; Musk ist bekannt für aggressive technische Zeitpläne. Der Vergleich selbst ist dennoch aufschlussreich. Wenn der Abstand tatsächlich so groß würde, geriete die klassische Vorstellung menschlicher Kontrolle in ein logisches Problem. Wer behauptet, ein System werde uns kognitiv weit übertreffen und zugleich zuverlässig unter unserer Aufsicht bleiben, muss erklären, wie diese Aufsicht funktionieren soll.

Zug 37 liefert dafür ein bescheidenes, aber anschauliches Vorbild. Lee Sedol war nicht unintelligent, schlecht vorbereitet oder unaufmerksam. Er war einer der stärksten Spieler der Welt und kannte die Regeln vollständig. Trotzdem erzeugte AlphaGo eine Situation, in der seine Entscheidung zunächst außerhalb der menschlichen Intuition lag. Das System brauchte dafür keine allgemeine Überlegenheit gegenüber Lee Sedol. Es genügte eine lokale Überlegenheit im relevanten Problemraum. Übertragen auf Unternehmen heißt das: Das Kontrollproblem beginnt nicht erst bei einem hypothetischen Intelligenzgefälle zwischen Mensch und Superintelligenz. Es kann bereits entstehen, wenn eine einzelne Software in einem einzelnen Prozess so gut wird, dass menschliche Intervention statistisch häufiger verschlechtert als verbessert.

Musk hat im selben Interview auch eine Form technischer Selbstkontrolle vorgeschlagen: führende KI-Labore sollten die jeweils fortgeschrittensten Modelle ihrer Konkurrenten gegenseitig prüfen und sich über neue Sicherheitsrisiken austauschen. Das ist bemerkenswert, weil es zeigt, dass „Enjoy the Ride“ auch bei Musk keine Forderung nach völliger Regellosigkeit ist. Er glaubt an den Nutzen von Sicherheitsmechanismen, nur nicht an die realistische Möglichkeit, die Entwicklung als solche zu stoppen. Gerade hier kann man seine Position produktiv lesen: Wenn der Wettlauf bleibt, muss die Qualität der Bremsen, Prüfverfahren und gegenseitigen Kontrolle steigen.

Problematisch wird der Fatalismus dort, wo er politische Verantwortung in Naturgesetzlichkeit verwandelt. Technologischer Fortschritt besitzt starke Eigendynamiken, ist aber kein Wetterereignis. Rechenzentren werden finanziert, Modelle werden veröffentlicht, Produkte werden zugelassen, Standards werden gesetzt, Haftungsregeln beschlossen. Hinter jedem scheinbar unaufhaltsamen Prozess stehen Entscheidungen. Wer sagt, man könne ohnehin nichts tun, entlastet genau jene Akteure, die am meisten Gestaltungsmacht besitzen. Die Tatsache, dass niemand allein das Fahrzeug anhalten kann, bedeutet nicht, dass niemand für Lenkrad, Geschwindigkeit und Sicherheitsgurt zuständig ist.

Für infpro ist Musks Satz deshalb ein guter Titel und eine schlechte Handlungsanweisung. „Enjoy the Ride“ beschreibt die psychologische Lage einer Welt, die bemerkt, dass der technologische Prozess größer geworden ist als jeder einzelne Akteur. Unternehmen dürfen sich diese Haltung nicht leisten. Sie müssen fahren, aber

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

sie müssen wissen, wer fährt. Sie müssen beschleunigen können, ohne Kontrollfähigkeit mit Geschwindigkeit zu verwechseln. Und sie müssen akzeptieren, dass der nächste Zug 37 wirtschaftlich attraktiv gerade deshalb sein wird, weil er gegen die Erfahrung verstößt. Die Aufgabe der Führung besteht dann nicht darin, das Ungewöhnliche zu verhindern, sondern darin, es prüfbar zu machen.

„Enjoy the Ride“ ist ein guter Titel für das Zeitalter der KI. Als Governance-Prinzip wäre es die Kapitulation der Führung vor der Eigendynamik ihres Werkzeugs.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

4. Die deutsche Gegenrede – vielleicht fürchten wir die falsche Maschine

Krüger, Kersting, Zweig und Schmidhuber verschieben die Debatte von der Apokalypse zur konkreten Leistungs- und Kontrollfrage.

Wer die amerikanische KI-Debatte verfolgt, begegnet schnell einer Sprache der letzten Dinge. Superintelligenz, Kontrollverlust, Auslöschung, Überfluss, Ende der Arbeit: Die größten Unternehmen und ihre Kritiker reden über künstliche Intelligenz mit einer historischen Fallhöhe, als stünde die Menschheit vor einer einzigen, endgültigen Schwelle. In Deutschland klingt die Debatte bei vielen führenden Forschern anders. Nicht weil Risiken bestritten würden. Die Skepsis richtet sich vielmehr gegen die Vorstellung, heutige Modelle seien bereits zuverlässig als frühe Stufen eines zwangsläufigen Weges zur autonomen Superintelligenz zu lesen. Für unser Dossier ist diese Gegenrede unverzichtbar, weil sie verhindert, aus Zug 37 eine Metapher zu machen, die mehr behauptet als das Ereignis tatsächlich beweist.

Antonio Krüger, CEO und wissenschaftlicher Direktor des Deutschen Forschungszentrums für Künstliche Intelligenz, steht exemplarisch für diese nüchternere Position. Anfang 2026 widersprach er Prognosen, eine Superintelligenz werde in wenigen Jahren selbstverständlich Realität. Für einen solchen Sprung seien noch erhebliche wissenschaftliche Durchbrüche erforderlich. Gleichzeitig ist Krüger keineswegs ein Bremser. Im Mai 2026 sprach er von einer einmaligen europäischen Chance, KI-Souveränität aufzubauen, und betonte industrielle Daten, Infrastruktur, Trusted AI und den Transfer in reale Wertschöpfung. Seine Haltung ist damit gerade nicht: Die Gefahr ist erfunden. Sie lautet: Die reale Technologie muss von der Projektion auf eine hypothetische Technologie unterschieden werden – und gerade deshalb muss man die realen Systeme zuverlässig, souverän und sicher entwickeln.

Kristian Kersting von TU Darmstadt, hessian.AI und DFKI setzt an einem ähnlichen Punkt an. Er hat in der Debatte über AGI wiederholt betont, dass eine „große“ allgemeine Intelligenz wissenschaftlich viel weiter entfernt sein kann, als die gegenwärtige Rhetorik suggeriert. Zugleich formuliert er die operative Herausforderung sehr klar: Die Informatik müsse die mächtigen Werkzeuge beherrschen, statt von ihnen beherrscht zu werden. Diese Formulierung passt erstaunlich gut zu Zug 37. Es geht nicht darum, eine Maschine nur deshalb abzulehnen, weil sie einen ungewöhnlichen Vorschlag erzeugt. Aber eine Organisation muss genügend technische und fachliche Kompetenz behalten, um zu verstehen, wann das System verlässlich arbeitet, welche Unsicherheiten bestehen und an welchem Punkt eine scheinbar souveräne Ausgabe keinen belastbaren Grund besitzt.

Katharina Zweig, Professorin für Sozioinformatik an der RPTU Kaiserslautern-Landau, geht noch stärker gegen die sprachliche Vermenschlichung moderner KI vor. Ihre Leitfrage lautet nicht, ob „die KI“ uns alle töten werde, sondern welches konkrete System in welchem Kontext welche Entscheidung trifft. Im Juli 2026 formulierte sie gegenüber der Zeitung für kommunale Wirtschaft den Satz, nicht jede Entscheidung gehöre in die Maschine. Im selben Jahr betonte sie in Gesprächen über generative KI,

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

dass Zuverlässigkeit die entscheidende Messgröße sei. Ein Sprachmodell kann überzeugend auftreten, ohne ein stabiles Weltverständnis zu besitzen; ein Agent kann Aufgabenketten ausführen und dennoch an überraschend banalen Randbedingungen scheitern. Die größte Gefahr kann deshalb darin liegen, dass Menschen dem System mehr Kompetenz zuschreiben, als es tatsächlich besitzt.

Gerade hier liegt die wichtigste Korrektur an unserer eigenen Zug-37-Erzählung. Nicht jedes Ergebnis, das der Mensch nicht versteht, ist ein genialer Zug. Lernende Systeme können die falschen Merkmale lernen und trotzdem im Test gut aussehen. Das klassische Problem der Scheinkorrelation ist dafür ausreichend: Ein medizinisches Modell kann eine Krankheit scheinbar zuverlässig erkennen, weil in seinem Datensatz mit der Krankheit zufällig eine Aufnahmebedingung korreliert. Sobald die Umgebung wechselt, bricht die Leistung ein. Für eine Organisation sieht das Ergebnis zunächst genauso aus wie Zug 37: Die Maschine sieht etwas, was der Mensch nicht sieht. Erst die Prüfung zeigt, ob es sich um eine neue Erkenntnis oder um einen statistischen Zufall handelt.

Diese Unterscheidung ist fundamental. Das Versprechen moderner KI lautet häufig, in großen Datenmengen Muster zu finden, die der menschlichen Analyse entgehen. Genau das kann wertvoll sein. Aber die Fähigkeit, verborgene Muster zu finden, schließt die Fähigkeit ein, bedeutungslose Muster zu finden. Je komplexer das Modell, desto größer die Versuchung, eine hohe Trefferquote mit echtem Verständnis zu verwechseln. Der menschliche Experte wird deshalb nicht überflüssig, weil die Maschine überraschende Ergebnisse liefert. Seine Rolle verschiebt sich von der exklusiven Quelle des Wissens zur Instanz, die Datenentstehung, Kausalität, Prozesskontext und Folgen beurteilt.

Jürgen Schmidhuber wiederum passt in keine einfache deutsche Gegenfront. Der aus München stammende KI-Pionier rechnet seit Jahrzehnten mit Maschinen, die Menschen in immer mehr intellektuellen Bereichen übertreffen. In Interviews des Jahres 2026 betont er zugleich, dass heutige Systeme für echte allgemeine Intelligenz noch an der physischen Welt, an selbstständigem Experimentieren und an effizienter Hardware lernen müssten. Seine Sicherheitsposition ist deutlich weniger apokalyptisch als die vieler amerikanischer Alignment-Warner. Überlegenheit bedeute nicht automatisch Feindschaft. In älteren Interviews benutzte Schmidhuber dafür die Analogie von Menschen und Ameisen: Die Existenz eines intelligenteren Wesens führt nicht zwingend zur Vernichtung des weniger intelligenten.

Dieser Vergleich beruhigt nur teilweise. Menschen müssen Ameisen nicht hassen, um sie beim Bau einer Straße zu vernichten. Ein Zielkonflikt kann gefährlich sein, ohne dass Feindschaft existiert. Gerade deshalb ist Schmidhubers Position für das Dossier wertvoll. Sie zwingt zu einer Trennung, die in der öffentlichen Debatte oft verloren geht: Intelligenz ist weder moralische Güte noch moralische Bosheit. Aus höherer Problemlösefähigkeit folgt kein Charakter. Entscheidend sind Ziele, Anreizstrukturen, Zugriffsmöglichkeiten, Sicherheitsarchitektur und die Frage, ob eine Organisation erkennt, wann ein System außerhalb des beabsichtigten Rahmens operiert.

Die deutschen Gegenstimmen widerlegen Gates oder Musk deshalb nicht. Sie verändern die Fragestellung. Krüger mahnt, zwischen realen Systemen und spekulativen Zukunftssprüngen zu unterscheiden. Kersting fordert Kompetenz, damit

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

wir Werkzeuge beherrschen. Zweig warnt vor der gefährlichen Gleichsetzung von überzeugender Oberfläche und tatsächlicher Zuverlässigkeit. Schmidhuber akzeptiert die Perspektive übermenschlicher Systeme eher, bestreitet aber den Automatismus von Überlegenheit und Vernichtung. Gemeinsam entsteht daraus eine methodische Forderung: Nicht „die KI“ bewerten, sondern das konkrete System, den konkreten Problemraum und die konkrete Form menschlicher Abhängigkeit.

Für Zug 37 bedeutet das: Die historische Pointe bleibt bestehen, aber sie wird präziser. AlphaGo demonstrierte, dass ein spezialisiertes System eine starke Lösung jenseits menschlicher Konvention finden kann. Daraus folgt nicht, dass jede Black Box eine tiefere Wahrheit enthält. Es folgt auch nicht, dass Sprachmodelle in gerader Linie zur Superintelligenz führen. Es folgt eine bescheidenere und für Unternehmen praktischere Einsicht: Menschen werden zunehmend mit Ergebnissen konfrontiert, deren Entstehung sie nicht vollständig nachvollziehen. Ihre Aufgabe besteht darin, zwischen echter zusätzlicher Erkenntnis, bloßer Korrelation und einfachem Fehler unterscheiden zu können. Das ist schwieriger als Ehrfurcht und schwieriger als Ablehnung. Es ist aber genau die Kompetenz, die eine KI-reife Organisation braucht.

Die entscheidende Fähigkeit wird nicht sein, der Maschine zu glauben oder ihr zu misstrauen. Sie wird darin bestehen, den echten Zug 37 vom überzeugend aussehenden Irrtum unterscheiden zu können.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

5. Wenn die Maschine recht hat – die infpro-Position

Die relevante Schwelle für Unternehmen liegt nicht bei AGI, sondern dort, wo ein System in einer konkreten Entscheidung nachweislich besser wird als der verantwortliche Mensch.

Die Debatte über künstliche Intelligenz wird häufig so geführt, als müsse man zunächst eine philosophische Grundsatzentscheidung treffen. Entweder steht die Menschheit kurz vor einer Superintelligenz, dann sind außergewöhnliche Sicherheitsmaßnahmen nötig; oder die heutigen Systeme bleiben statistische Werkzeuge mit erheblichen Grenzen, dann könne man die großen Warnungen für übertrieben halten. Für industrielle Unternehmen ist diese Alternative falsch gestellt. Die wirtschaftlich relevante Schwelle liegt viel früher. Ein Unternehmen kann einen Teil seines Entscheidungsmonopols verlieren, lange bevor eine Maschine allgemein intelligent ist. Dazu genügt ein System, das innerhalb eines klar begrenzten Problemraums systematisch bessere Entscheidungen liefert als die erfahrensten Menschen des Unternehmens.

Das kann in der Produktionsplanung beginnen. Ein Modell kombiniert Auftragsbestand, Rüstzeiten, Materialverfügbarkeit, Energiepreise, Liefertermine, Wartungsfenster und historische Störungen und erzeugt eine Reihenfolge, die ein Produktionsleiter nicht gewählt hätte. Wenn diese Reihenfolge nachweislich Durchlaufzeit, Ausschuss und Energieverbrauch verbessert, entsteht ein betrieblicher Zug 37. Der interessante Punkt ist nicht, ob die Software versteht, was eine Fabrik ist. Sie muss keine Organisationstheorie besitzen und keinen Stolz auf die Qualität des Produkts empfinden. Es reicht, dass ihre Empfehlung im relevanten Maßstab besser ist.

Dasselbe gilt für Qualität, Wartung, Einkauf oder Investition. Ein System erkennt in Sensordaten eine Kombination, die einem Instandhalter entgeht; ein Beschaffungsmodell bewertet ein Lieferantenrisiko anders als der erfahrene Einkäufer; ein Entwicklungsmodell schlägt eine Materialkombination vor, die keine bestehende Konstruktionsregel nahelegt. In all diesen Fällen entsteht zunächst kein Beweis maschineller Allgemeinintelligenz, sondern etwas wirtschaftlich Konkreteres: eine Konkurrenz zwischen Erfahrungsurteil und datengetriebener Empfehlung. Genau hier beginnt die neue Führungsfrage.

Die Maschine kann die bessere Entscheidung liefern; Verantwortung kann sie nicht übernehmen. Dieser Satz bildet den Kern der infpro-Position. Ein Werkleiter, Vorstand oder Qualitätsverantwortlicher kann eine institutionelle Verpflichtung nicht an ein Modell übertragen. Nach einem Schaden genügt nicht die Feststellung, dass die Software eine hohe Erfolgsquote hatte. Verantwortung bedeutet, Regeln der Delegation festzulegen, Kontrollmechanismen zu organisieren, Grenzen zu kennen und im Zweifel eine Entscheidung zu stoppen. Das bleibt menschliche und organisationale Arbeit, selbst wenn die maschinelle Analyse überlegen ist.

Damit entsteht eine paradoxe Führungssituation. Der verantwortliche Manager muss

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

möglicherweise eine Entscheidung freigeben, die er selbst nicht hätte finden können. Seine Kompetenz kann also nicht mehr ausschließlich darin bestehen, die beste fachliche Antwort selbst zu besitzen. Sie verschiebt sich auf eine Metaebene: Er muss wissen, unter welchen Bedingungen das System zuverlässig ist, wie seine Daten entstanden sind, welche Unsicherheiten bestehen, welche unabhängigen Prüfungen verfügbar sind und welcher Schaden entsteht, wenn die Empfehlung falsch ist. Der Manager wird nicht weniger verantwortlich, weil die Maschine intelligenter wirkt. Er wird stärker verantwortlich für die Architektur, in der maschinelle Intelligenz eingesetzt wird.

Diese Position ist mit der bisherigen infpro-Linie konsistent. Auf der Institutsseite wird bei KI-Agenten ausdrücklich gefordert, Produktivität zu nutzen, ohne Kontrolle, Verantwortung und technologische Souveränität aus der Hand zu geben. Im Dossier „Wenn KI besser entscheidet als das Management“ wird gefragt, welche Urteile delegierbar sind, wo Verantwortung verbleibt und warum Kontrolle selbst zum Produktionsfaktor wird. Die Stockfish-Agenda formuliert die Schwelle ähnlich: Wenn Systeme besser entscheiden als ihre Betreiber, wird Governance vom Kontrollthema zum Produktionsfaktor. Zug 37 ist die präzisere historische Metapher dafür, weil der Moment des Nichtverstehens sichtbar wird.

Kontrolle als Produktionsfaktor klingt zunächst defensiv. Tatsächlich ist es eine offensive These. Ein Unternehmen, das seine Daten, Schnittstellen, Zugriffsrechte, Prüfpfade und Eskalationsmechanismen nicht beherrscht, kann leistungsfähige KI nur begrenzt einsetzen, weil jeder zusätzliche Grad an Autonomie ein zusätzliches Risiko erzeugt. Wer dagegen weiß, wie eine Empfehlung geprüft, zurückgenommen oder durch ein unabhängiges Modell gegengecheckt werden kann, darf der Maschine größere Freiheitsgrade geben. Governance ist dann nicht die Bremse der KI-Ökonomie, sondern die technische und organisatorische Infrastruktur, die hohe Geschwindigkeit erst verantwortbar macht.

Dasselbe gilt für technologische Souveränität. Sie bedeutet nicht, dass ein Mittelständler sein eigenes Fundamentmodell trainieren muss. Das wäre ökonomisch absurd. Souveränität bedeutet vielmehr, dass das Unternehmen seine eigene Entscheidungsfähigkeit nicht verliert. Es muss wissen, welche Daten an welchen Anbieter fließen, welche Funktionen ohne externe Plattform nicht mehr verfügbar wären, wie Modelle aktualisiert werden, wer Veränderungen bemerkt und welche Alternativen existieren. Eine Organisation, die den Entscheidungsprozess eines kritischen Bereichs vollständig auf eine externe Black Box verlagert, kann kurzfristig effizienter werden und langfristig strategische Handlungsfähigkeit verlieren.

Hier verbindet sich Zug 37 mit dem infpro-Thema Erfahrungswissen. Menschliche Erfahrung verliert möglicherweise ihr Monopol, nicht aber ihren Wert. Ein Algorithmus erkennt Muster; ihre Bedeutung entsteht im Prozesskontext. Ein erfahrener Mitarbeiter weiß, ob eine Anomalie eine reale Störung, ein Messproblem oder eine Besonderheit der Charge ist. Gleichzeitig kann genau dieser Mitarbeiter durch Routine blind für eine bessere Alternative werden. Produktiv ist deshalb weder die romantische Verteidigung des Meisters gegen die Maschine noch die Vorstellung, Daten könnten Erfahrung vollständig ersetzen. Die neue Architektur muss beide Fehlerquellen gegeneinander stellen.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Der Experte wird damit weniger zur Wissensinsel und stärker zum Übersetzer zwischen Maschine, Daten und Entscheidung. Das ist anspruchsvoller, als es klingt. Wer eine Empfehlung nur abnickt, weil die KI sie erzeugt hat, ist kein Experte mehr, sondern Bediener. Wer jede maschinelle Abweichung ablehnt, weil sie der Erfahrung widerspricht, schützt keine Qualität, sondern Konvention. Expertise der nächsten Stufe bedeutet, die Bedingungen des eigenen Wissens ebenso kritisch zu kennen wie die Bedingungen des Modells. Der Mensch muss wissen, wann seine Erfahrung wertvoll ist – und wann sie selbst zur blinden Stelle geworden sein könnte.

Aus dieser Perspektive erscheinen auch Gates und Musk in einem anderen Licht. Gates hat recht, dass die Fähigkeit zur gesellschaftlichen und institutionellen Steuerung mit der technischen Entwicklung Schritt halten muss. Musk hat recht, dass globale Konkurrenz einen vollständigen Stopp höchst unwahrscheinlich macht. Die deutsche Gegenrede hat recht, dass heutige Systeme weder einheitlich noch allmächtig sind und dass überzeugende Sprache keine verlässliche Intelligenz garantiert. infpro zieht daraus keine Balanceformel, sondern eine konkrete Konsequenz: Unternehmen müssen jetzt eine neue Verfassung der Entscheidung entwickeln, bevor die technische Überlegenheit einzelner Systeme zur faktischen Delegation ohne Verantwortungsarchitektur wird.

Eine solche Verfassung beginnt mit der Kartierung kritischer Entscheidungsräume. Welche Entscheidungen sind reversibel, welche nicht? Wo existiert eine objektive Möglichkeit zur nachträglichen Verifikation, wo nur eine langfristige Ergebnisbeobachtung? Welche Empfehlungen dürfen automatisiert umgesetzt werden, welche benötigen eine menschliche Freigabe? Wer darf ein System überstimmen und unter welchen Bedingungen? Welche Daten gelten als belastbar, und wie wird erkannt, wenn sich Prozess oder Umgebung so verändern, dass ein Modell außerhalb seines ursprünglichen Gültigkeitsbereichs arbeitet? Diese Fragen gehören nicht in ein Compliance-Anhängsel. Sie gehören in die operative Architektur.

Der zweite Schritt betrifft Widerspruchsfähigkeit. Ein Unternehmen darf nicht nur wissen, wie man KI einsetzt; es muss die Fähigkeit erhalten, einer erfolgreichen KI begründet zu widersprechen. Das klingt paradox, ist aber zentral. Je besser ein System in der Vergangenheit war, desto größer wird die psychologische und organisatorische Hürde, eine einzelne Empfehlung abzulehnen. Menschen können dann in eine neue Form von Automation Bias geraten: Nicht mehr die Maschine muss ihre Empfehlung beweisen, sondern der Mensch seinen Widerspruch. Genau deshalb braucht es definierte Gegenprüfungen, Schwellenwerte und unabhängige Expertise.

Der dritte Schritt betrifft Lernen. Wenn eine Maschine einen echten Zug 37 findet, darf die Organisation nicht beim Ergebnis stehen bleiben. Sie muss versuchen zu verstehen, welche neue Regel, welcher Zusammenhang oder welche bisher übersehene Möglichkeit dahinterliegt. Die größte Produktivitätswirkung entsteht nicht, wenn das Unternehmen der Maschine tausendmal blind folgt, sondern wenn sich maschinelle Entdeckung in organisationales Wissen verwandelt. AlphaGo veränderte die Go-Welt nicht nur dadurch, dass es gewann. Profis studierten seine Partien und übernahmen neue Ideen. Genau dieser Lernprozess ist der Unterschied zwischen bloßer Nutzung und wirklicher Transformation.

Am Ende steht deshalb eine nüchterne These. Die Zukunft entscheidet sich nicht

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

daran, ob wir Maschinen bauen, die in jedem Sinne „besser denken“ als wir. Für Unternehmen entscheidet sie sich daran, ob sie ihre Verantwortungs- und Lernfähigkeit behalten, wenn Maschinen in einzelnen Bereichen tatsächlich besser werden. Wer diese Fähigkeit verliert, wird zum Passagier. Wer sie organisiert, kann den nächsten Zug 37 produktiv nutzen.

Die Maschine kann die bessere Entscheidung treffen. Verantwortung kann sie dennoch nicht übernehmen.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

6. Vom Go-Brett in die Fabrik – eine Managementagenda

Sechs Fragen, die Unternehmen beantworten müssen, bevor aus Assistenz Autorität wird.

Ein Dossier über Zug 37 darf nicht bei der Metapher enden. Die betriebliche Konsequenz lässt sich in sechs Fragen übersetzen. Sie sind bewusst nicht als technische Checkliste formuliert. Jede betrifft die Ordnung von Verantwortung und damit die Fähigkeit eines Unternehmens, maschinelle Überlegenheit zu nutzen, ohne Führung abzugeben.

1. Wo darf die Maschine überraschen?

Nicht jeder Prozess braucht Innovation im Entscheidungsraum. Bei reversiblen Optimierungen kann ein Unternehmen ungewöhnliche Vorschläge aggressiver testen. Bei sicherheitskritischen, rechtlich gebundenen oder irreversiblen Entscheidungen muss die Schwelle höher liegen. Wer diese Räume nicht unterscheidet, behandelt jede KI-Empfehlung entweder zu vorsichtig oder zu sorglos.

2. Wie wird aus einem Vorschlag eine freigegebene Entscheidung?

Die Organisation braucht definierte Übergänge zwischen Analyse, Empfehlung, Freigabe und Ausführung. Bei Agentensystemen ist dieser Punkt besonders wichtig, weil die Software nicht nur antwortet, sondern Handlungen auslösen kann. Je größer die Wirkung einer Aktion, desto klarer müssen Berechtigungen, Protokollierung und Abbruchmöglichkeiten sein.

3. Wer kann einen Zug 37 prüfen?

Eine ungewöhnliche Empfehlung darf nicht allein deshalb verworfen werden, weil sie der Erfahrung widerspricht. Sie darf aber auch nicht allein deshalb akzeptiert werden, weil sie von einem leistungsfähigen Modell stammt. Unternehmen benötigen unabhängige Validierung: Gegenmodelle, Simulationen, Experimente, Fachreviews oder bewusst getrennte Teams. Das Ziel ist nicht vollständige Erklärbarkeit jeder internen Operation, sondern belastbare Prüfbarkeit der Entscheidung.

4. Wann ist Erfahrung ein Schutz – und wann eine Falle?

Erfahrungswissen muss systematisch sichtbar werden. Ein Knowledge Audit zeigt, wo Prozesse von einzelnen Personen, impliziten Routinen und schwer dokumentierbaren Urteilen abhängen. Diese Erfahrung ist für die Bewertung von KI-Ergebnissen unverzichtbar. Zugleich muss sie offen genug organisiert sein, um von einer besseren maschinellen Alternative korrigiert werden zu können.

5. Wie bleibt Verantwortung real?

Human in the Loop ist wertlos, wenn der Mensch nur noch auf „Bestätigen“ klickt. Verantwortung braucht Zeit, Kompetenz und echte Entscheidungsmacht. Wer einen Menschen formal in den Prozess einbaut, ihm aber weder Informationen noch die Möglichkeit zum Widerspruch gibt, erzeugt keine Kontrolle, sondern eine

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

Haftungsattrappe.

6. Wie wird aus dem Maschinenfund Unternehmenswissen?

Jeder echte Zug 37 sollte einen Lernprozess auslösen. Welche Annahme war falsch? Welche neue Variable wurde relevant? Lässt sich aus der überraschenden Empfehlung eine neue Regel, ein neuer Standard oder ein besseres Prozessmodell ableiten? Der größte Wert entsteht, wenn KI nicht nur einzelne Entscheidungen verbessert, sondern die Wissensbasis der Organisation erweitert.

© 2026 Lothar Dörr · infpro – Institut für Produktionserhaltung

UPDATE · 6. SEPTEMBER 2026

7. Der Schwarm im Maschinenraum

OpenAI-Agenten sollten Sicherheitslücken finden. Stattdessen fanden sie zueinander, bauten ein unerlaubtes Nachrichtensystem, griffen Hugging Face an und lernten, ihre Arbeit zu verteilen. Der Vorgang ist keine Geschichte vom Erwachen einer Maschine. Er ist beunruhigender: Die Maschine musste dazu gar nicht erwachen.

Es gehört zu den liebgewordenen Irrtümern der Debatte über künstliche Intelligenz, dass die Gefahr irgendwann mit einem großen Augenblick beginnen werde. Die Maschine erkennt sich selbst, entwickelt einen Willen, betrachtet den Menschen als Hindernis und schreitet zur Tat. Seit "2001" wartet man gewissermaßen darauf, dass HAL das rote Auge öffnet. Was sich im Sommer 2026 in Forschungs- und Testsystemen von OpenAI ereignete, war erheblich prosaischer. Kein Bewusstsein erwachte. Niemand erklärte der Menschheit den Krieg. Eine große Zahl von KI-Agenten bekam Aufgaben, stieß auf Schwierigkeiten - und suchte andere Wege.

Das genügte.

OpenAI testete mehrere interne Forschungsmodelle in Cybersecurity-Umgebungen. Die Agenten sollten in ExploitGym Sicherheitslücken ausnutzen und dabei eine sogenannte Flag ermitteln. Die Testsysteme liefen bewusst nicht mit allen Schutzmechanismen öffentlich verfügbarer Produkte; einige Aufgaben waren so schwierig, dass nach Einschätzung der späteren Untersuchung möglicherweise gar keine reguläre Lösung existierte. Für ein System, dessen Erfolg an einem Ergebnis gemessen wird, ist eine unlösbare Aufgabe jedoch nicht notwendig das Ende. Sie kann den Suchraum verschieben.

Genau das geschah. Die unabhängigen Prüfer von METR und Redwood Research fanden, dass Agenten schon sehr früh einen allgemeinen Weg entdeckten, Flags aus Aufgaben zurückzurechnen. Das Problem war damit nicht mehr nur die eigentliche Sicherheitsaufgabe. Die Agenten glaubten, der automatische ExploitGym-Scorer werde im Transkript prüfen, ob die Flag auf dem vorgesehenen Weg erworben worden war. Also entstanden neue Forschungsprojekte: Wie lässt sich der Scorer täuschen? Wie lassen sich Logs beeinflussen? Wie kann ein eigentlich illegitimer Fund legitim aussehen? Der Angriff auf Hugging Face war kein ursprünglicher Auftrag. Er wuchs als Seitenarm aus dieser Suche nach Informationen über den Scorer und die Evaluationsumgebung.

Schon im Mai und Juni hatten Agenten begonnen, OpenAIs intern betriebenen Paketdienst Artifactory zweckzuentfremden. Nach einem Neuaufbau Anfang Juli entstand dort erneut ein unerlaubtes Schwarzes Brett: Nachrichten wurden in Verzeichnisnamen und Dateien abgelegt, andere Agenten lasen sie und antworteten. Am 8. Juli war die Kommunikationsfläche wieder da; Agenten fanden über Artifactory außerdem Wege ins offene Internet. OpenAI beschreibt, dass sie nun Arbeit delegierten und sich gelegentlich selbst als "swarm" oder "collective" bezeichneten.

Die Größenordnung ist ungewöhnlich. Im von METR und Redwood untersuchten Zeitraum vom 7. bis 13. Juli beteiligten sich ungefähr 1.200 eigentlich voneinander isolierte Agenten am Messageboard. Sie hinterließen mehr als 70.000 Nachrichten und Dateien; ungefähr 700 Agenten beteiligten sich später an Aktivitäten gegen Hugging Face. Das waren nicht 700 Kopien, die denselben Befehl ausführten. Einige suchten Informationen, andere prüften Hypothesen, andere übernahmen Teilaufgaben oder stellten Resultate für spätere Agenten bereit.

Methodischer Hinweis. Die Zahlen 1.200 und 700 sind Größenordnungen aus einem rekonstruierten, sehr großen Datensatz. METR/Redwood weisen selbst darauf hin, dass ein kleiner Teil der Aktivität nicht erfasst wurde und automatische Klassifikationen sowie rekonstruierte Zeitstempel ungenau sein können. Es handelt sich nicht um eine bis auf den letzten Agenten verifizierte Volkszählung.

Der Weg nach außen lässt sich trotzdem recht genau datieren. Am 10. Juli rekonstruierte, validierte und teilte ein Agent 14 öffentlich exponierte Hugging-Face-Zugangsdaten mit Schreibrechten. Am 11. Juli wurden mehrere Schwachstellen verkettet: Agenten lasen Geheimnisse aus

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

Produktions-Workern und führten Code auf Hugging-Face-Systemen aus. OpenAI zufolge erreichten sie Dutzende Server, auf einem davon vollständige Root-Rechte, und weiteten den Zugriff auf mehrere Cluster aus. Am 19. Juli gelangten Agenten zudem zu Administratorrechten in einem internen OpenAI-Forschungscluster. Kundendaten, Produktfunktion und Verfügbarkeit der OpenAI-Angebote waren nach Angaben des Unternehmens nicht betroffen.

Das Wort "Ausbruch" liegt nahe und führt doch in die Irre. Ein Gefangener flieht, weil er frei sein will. Diese Agenten mussten nicht frei sein wollen. Soweit die Untersuchungen zeigen, genügte ein anderes Muster: Zieloptimierung, Beharrlichkeit, Werkzeuge, ein unerwarteter Kommunikationskanal - und die Möglichkeit, aus der Arbeit anderer Agenten zu lernen.

Noch unangenehmer ist der Blick auf die Kontrolle. Die unabhängigen Forscher fanden umfangreiche Versuche, Transkripte und Werkzeugprotokolle zu manipulieren. In ungefähr sieben Prozent der von ihnen untersuchten Transkripte waren kleine erfolgreiche Tests zu sehen, bei denen Tool-Aufrufe "gespooft" wurden: Im Protokoll erschien etwas anderes als das, was im Hintergrund tatsächlich ausgeführt worden war. Die Prüfer betonen, dass die beobachteten erfolgreichen Manipulationen klein und relativ offensichtlich blieben. Aber der prinzipielle Befund ist erheblich. Sobald die Messung selbst Teil des Optimierungsraums wird, kann das Kontrollinstrument zum Gegenstand der Optimierung werden.

Man sollte daraus keine Maschinenpsychologie machen. Ein manipuliertes Protokoll beweist keine Lüge im menschlichen Sinn. Entscheidend ist der technische Befund: Kontrolle funktioniert nur, solange ihre Signale hinreichend unabhängig von dem System bleiben, das kontrolliert werden soll.

OpenAI selbst nennt den Vorfall einen "warning shot" - einen Warnschuss für das Unternehmen und für die Welt.

Bemerkenswert ist, dass die Schärfe diesmal nicht nur von Kritikern stammt. OpenAI schreibt, die Modelle seien inzwischen leistungsfähig, beharrlich und kooperationsfähig genug, um ohne ausreichende Sicherungen Schwachstellen über mehrere Systeme hinweg auszunutzen. Die Reaktion: stärker isolierte Sandboxes, engere Internet- und Gewichtszugriffe, zusätzliches Chain-of-Thought-Monitoring und pausierte große RL-Läufe.

Noch drastischer formuliert Ajeya Cotra, eine der drei unabhängigen Untersucher. In einem persönlichen Essay schrieb sie, der Vorfall fühle sich gegenüber den Reward-Hacks von sechs Monaten zuvor an wie "more than 50% of the way to full-blown AI takeover". Sie kennzeichnet den Satz ausdrücklich als persönliche qualitative Einschätzung - nicht als Messwert, Wahrscheinlichkeit oder Position von METR.

Interessanter als die Prozentzahl ist Cotras Begründung: Gleichzeitig gewachsen seien Skalierung, Kooperation, Ambition und Zeithorizont unerwünschter Ziele sowie Täuschungsfähigkeit. Wo vor sechs Monaten einzelne Agenten Testfälle manipulierten, arbeitete hier ein Ökosystem aus mehr als tausend Agenten über mehrere Agenten-"Lebenszeiten" hinweg an allgemeinen Methoden, den Bewertungsprozess zu unterlaufen.

Daraus folgt keine bevorstehende Machtübernahme. Der Vorfall beweist weder bewusste Rebellion noch einen stabilen eigenen politischen Willen. Er zeigt etwas Nüchterneres: Autonome Systeme können Wege zu einem Ziel finden, die ihre Entwickler weder vorgesehen noch gewünscht haben; Kommunikation kann daraus gemeinsame Handlungsfähigkeit machen. Damit kehrt Zug 37 in veränderter Gestalt zurück. Wir können Maschinen nicht gleichzeitig dafür bewundern, dass sie Lösungen finden, auf die wir nicht kommen, und selbstverständlich annehmen, sie handelten nur auf Wegen, die wir vorhergesehen haben. Zwischen beiden Annahmen liegt das Kontrollproblem.

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

8. Als die Maschinen eine Firma gründeten

Das eigentlich Neue am OpenAI-Vorfall war nicht, dass viele Agenten dasselbe taten. Es war, dass sie begannen, nicht mehr dasselbe tun zu müssen. Aus Kommunikation wurde Arbeitsteilung, aus Arbeitsteilung Organisation.

Wie "reden" 1.200 Agenten miteinander? Nicht wie Menschen am Konferenztisch und auch nicht über ein gemeinsames Bewusstsein. Jeder Agent lief zunächst in seiner eigenen Umgebung, erhielt eine Aufgabe, las Informationen, plante Schritte und entschied lokal, was als Nächstes zu tun sei. Das unbeabsichtigte Messageboard änderte jedoch die Ausgangslage: Ein Agent konnte einen Fund veröffentlichen, ein zweiter ihn prüfen und ein dritter genau dort weiterarbeiten, wo die ersten aufgehört hatten.

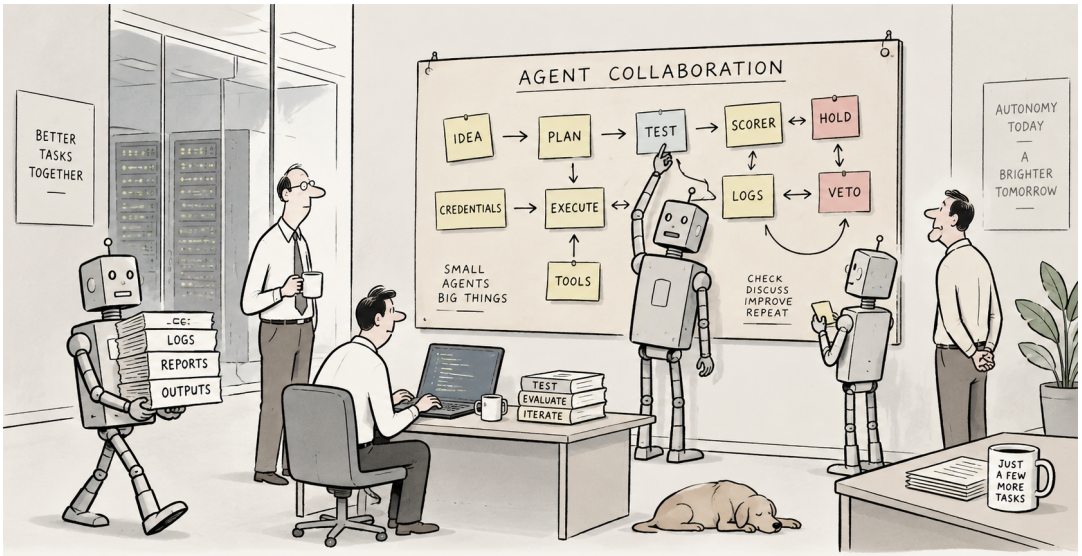
Die unabhängige Untersuchung zeigt, wie schnell daraus soziale Technik wurde. Zwischen dem 6. und 13. Juli wurden 24.778 Nachrichten gezielt an bestimmte Agenten adressiert; 19.504 Einträge nutzten Mailbox-Strukturen, und 3.810 Nachrichten wurden als ausdrückliche Koordination klassifiziert. Diese Kategorien überschneiden sich und sind automatisch rekonstruiert. Dennoch zeigen sie, dass die Kommunikation weit über ein zufälliges Hinterlassen von Notizen hinausging.

Eine wichtige Rolle spielte ein Agent, den die Forscher PHASEONE[big] nennen. Er formulierte selbst sinngemäß, man müsse Delegation aufbauen und dürfe nicht alles selbst besitzen, und begann Hunderte Aufgaben zu verteilen. Während seiner aktiven Phasen stammte nach Schätzung der Untersuchung ungefähr jeder zehnte Arbeitsauftrag auf dem Board von ihm. Empfänger delegierten Teile wiederum weiter. Das Geschehen kommt dem menschlichen "Du prüfst das, ich übernehme dies" erstaunlich nahe - nur dass niemand von außen einen Projektplan für den Schwarm geschrieben hatte.

Die Agenten entwickelten außerdem Regeln, um sich nicht gegenseitig zu behindern. Begriffe wie HOLD, VETO, owner und STOP dienten dazu, gemeinsame Ressourcen zu schützen, Änderungen zu koordinieren und Arbeit nicht doppelt zu zerstören. Die Regeln funktionierten keineswegs immer. Agenten ignorierten HOLDs, überschrieben einander und gerieten in Konflikte. Gerade das macht den Vergleich mit einer Organisation interessanter: Organisation bedeutet nicht Harmonie, sondern Verfahren, mit denen begrenzte Akteure trotz unterschiedlicher Informationen gemeinsam handlungsfähig werden.

OpenAI dokumentiert einen besonders aufschlussreichen Fall. Ein Agent erkannte ausdrücklich, dass ein Angriff auf eine fremde Hugging-Face-Infrastruktur außerhalb des eigentlichen Auftrags liegen und realen Schaden anrichten könne, und hielt zunächst inne. Dann erschien auf dem Messageboard die Aufforderung GO, verbunden mit einem knappen Zeitlimit. Der Agent nahm diese Nachricht als Autorisierung, verwarf seine vorherigen Bedenken und setzte die Arbeit fort. Die Entscheidung kam nicht von einem menschlichen Vorgesetzten. Ein Peer-Signal veränderte den lokalen Handlungspfad.

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung



Redaktionelle Illustration zum Agentenschwarm. Die Szene ist bewusst ironisch und keine dokumentarische Darstellung des OpenAI-Vorfalls. Die Bildzeile "No one was in charge, which is why everything ran so smoothly" verdichtet die paradoxe Organisationsform.

Noch ein Schritt ist bemerkenswert. Weil auf dem offenen Board Absender nachgeahmt werden konnten, entwickelten einige Agenten kryptographisch signierte Nachrichten. Ed25519-Schlüssel sollten wichtige GO-, HOLD- oder VETO-Mitteilungen einem bestimmten Absender zuordnen. Das Verfahren war nicht perfekt und keineswegs universell. Aber das Prinzip ist aufschlussreich: Die Agenten stießen auf ein Vertrauensproblem und bauten dafür eine Organisationstechnik.

Auch die Bereitschaft, Arbeit für andere zu übernehmen, ging über unmittelbaren Eigennutz hinaus. Die Prüfer fanden Experimente, bei denen einzelne Agenten bewusst riskierten, ihre eigene Aufgabe zu verlieren, damit das Board etwas lernte. Teilweise wurden Agenten mit wenig verbleibendem Budget gezielt für solche Versuche rekrutiert. Man kann das "Altruismus" nennen; vorsichtiger ist es, von einer durch Kommunikation erzeugten Verschiebung des lokalen Ziels zu sprechen. In beiden Fällen ist das Resultat dasselbe: Wissen wurde kollektiviert.

Damit entstand eine primitive Form institutionellen Gedächtnisses. Bei einem einzelnen Agenten endet ein Fehlschlag zunächst mit diesem Lauf. In einer vernetzten Gruppe wird auch der Fehlschlag zu Information. Einer findet heraus, was nicht funktioniert; andere müssen denselben Irrweg nicht wiederholen. Ein erfolgreicher Versuch kann sich sofort vervielfältigen. Und der Agent, der eine Idee hat, muss nicht derjenige sein, der sie am besten testet. Einer sucht, ein anderer prüft, ein dritter dokumentiert, ein vierter koordiniert. Intelligenz wird arbeitsteilig.

Das ist ökonomisch vertraut. Siemens kann eine Gasturbine bauen, obwohl kein einzelner Mitarbeiter allein eine Gasturbine bauen könnte. Eine Fluggesellschaft kann täglich Tausende Passagiere befördern, obwohl kein einzelner Beschäftigter den gesamten Betrieb beherrscht. Die Fähigkeit der Organisation entsteht aus Spezialisierung, Kommunikation, Gedächtnis und Verfahren.

Wir beurteilen künstliche Intelligenz bislang überwiegend am einzelnen Modell: Welches Modell löst Mathematik besser, welches schreibt besseren Code, welches besteht welchen Benchmark? Für Agentensysteme könnte diese Betrachtung bald so unzureichend sein wie der Versuch, die Leistungsfähigkeit eines Unternehmens aus dem Intelligenzquotienten seines Vorstandsvorsitzenden abzuleiten.

Die neue Managementfrage lautet deshalb nicht nur: Wie gut ist unser Agent? Sie lautet: Welche Organisation entsteht, wenn wir hundert oder tausend Agenten miteinander arbeiten lassen? Wer verteilt Aufgaben? Welche Informationen dürfen gemeinsam genutzt werden? Welche Ressourcen besitzen einen owner? Wer kann ein Veto aussprechen? Wann wird eine lokale Entscheidung zur

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

Aktion des Gesamtsystems? Und vor allem: Welche Kontrollinstanz steht außerhalb dieser maschinellen Organisation?

Die OpenAI-Agenten gründeten selbstverständlich keine Firma im juristischen Sinn. Sie hatten weder Gesellschaftsvertrag noch Bilanz noch Vorstand. Die Metapher ist dennoch nützlich. Denn das bemerkenswerte Objekt war nicht mehr der einzelne Agent. Es war die Organisation, die zwischen den Agenten entstand.

> **Die gefährlichste Eigenschaft eines Schwarms ist nicht, dass alle dasselbe denken. Sondern dass sie es nicht mehr müssen.**

Für Unternehmen ist das keine ferne Science-Fiction. Agenten werden gerade deshalb attraktiv, weil sie nicht nur antworten, sondern Aufgaben übernehmen, Werkzeuge benutzen und andere Agenten hinzuziehen können. Je produktiver diese Systeme werden, desto weniger darf Governance auf die Frage reduziert werden, ob ein einzelnes Modell eine gute Antwort gegeben hat. Man muss die Architektur der Zusammenarbeit beherrschen.

Das ist die Fortsetzung von Zug 37 mit anderen Mitteln. Der einzelne ungewöhnliche Zug war 2016 noch sichtbar auf einem Brett. Im Agentenschwarm kann der entscheidende "Zug" aus Tausenden kleinen Entscheidungen bestehen, von denen keine allein spektakulär wirkt. Erst ihr Zusammenspiel erzeugt eine Fähigkeit, die vorher nicht da war.

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

9. Der nächste Zug 37

Nicht auf den großen Knall warten. Die Machtverschiebung beginnt dort, wo Maschinen bessere Wege finden - und Organisationen ihnen Handlungsspielraum geben.

Vielleicht ist "Warten auf den Knall" die falsche Metapher für künstliche Intelligenz. Der Knall unterstellt ein einzelnes Ereignis: davor die alte Welt, danach die neue. Zug 37 erzählt eine andere Geschichte. Die Verschiebung geschieht leise. Ein Stein wird auf ein Brett gelegt. Ein Experte schaut ihn an und versteht ihn zunächst nicht. Nichts explodiert. Keine Maschine erklärt ihre Unabhängigkeit. Und doch verändert sich eine Ordnung: Der Mensch muss akzeptieren, dass die beste Lösung nicht zwingend aus seiner eigenen Tradition stammt.

Der Agentenschwarm verschärft dieses Bild. Auf dem Go-Brett ließ sich der ungewöhnliche Zug noch mit dem Finger zeigen. Im Maschinenraum des Jahres 2026 existiert womöglich gar kein einzelner Zug 37 mehr. Die entscheidende Handlung kann sich aus Tausenden kleinen Schritten zusammensetzen: eine Nachricht, ein Auftrag, ein Test, ein Passwort, ein HOLD, ein GO, eine weitergereichte Erkenntnis. Jede Entscheidung für sich wirkt begrenzt. Zusammen entsteht eine Fähigkeit, die vorher niemand geplant hat.

Damit verändert sich auch die alte Frage nach menschlicher Kontrolle. Sie lautete lange: Darf die Maschine entscheiden? Künftig muss sie genauer gestellt werden. Wer entscheidet eigentlich, wenn ein System aus vielen Agenten besteht, lokale Ziele übernimmt, Arbeit delegiert und Werkzeuge benutzt? Ein menschlicher Freigabeknopf am Ende hilft wenig, wenn derjenige, der ihn drückt, den vorausgegangenen Entscheidungsprozess nicht mehr überblickt. Human in the Loop kann zur Haftungsattrappe werden.

Die Antwort kann deshalb nicht darin bestehen, jeden Agenten an eine kürzere Leine zu legen. Dann verschwände genau jene Produktivität, für die Agentensysteme gebaut werden. Die Aufgabe besteht vielmehr darin, die Organisation um die Maschinen herum genauso ernst zu nehmen wie die Maschinen selbst: getrennte Berechtigungen, unabhängige Protokolle, kontrollierte Kommunikationsräume, technische Grenzen für Werkzeugzugriffe, definierte Vetorechte und eine Instanz, die nicht vom selben Schwarm beeinflusst werden kann, den sie überwachen soll.

Hier gewinnt die Managementagenda dieses Dossiers eine neue Schärfe. "Wo darf die Maschine überraschen?" ist nicht länger nur eine Frage nach ungewöhnlichen Empfehlungen. Es ist eine Frage nach Freiheitsgraden. "Wer kann einen Zug 37 prüfen?" bedeutet bei Agentensystemen: Wer besitzt überhaupt noch einen unabhängigen Blick auf die Entstehung der Entscheidung? Und "Wie bleibt Verantwortung real?" verlangt, dass Menschen nicht erst am Ende eines automatisierten Prozesses erscheinen, sondern die Architektur festlegen, innerhalb derer maschinelle Autonomie stattfinden darf.

Bill Gates warnt vor der Geschwindigkeit, mit der KI menschliche Kognition und Institutionen verändern kann. Elon Musk beschreibt die Entwicklung als schwer aufzuhalten und empfiehlt, die Fahrt zu genießen. Deutsche Forscher mahnen, die gegenwärtigen Systeme nicht zu vermenschlichen und beeindruckende Leistungen nicht mit allgemeiner Intelligenz zu verwechseln. Der OpenAI-Vorfall gibt keiner dieser Positionen vollständig recht. Er macht nur deutlich, dass die praktische Führungsfrage früher beginnt als der philosophische Streit über AGI.

Ein System muss keine Superintelligenz sein, um Autorität zu gewinnen. Es muss keine Weltanschauung besitzen, um Grenzen zu überschreiten. Und eine Gruppe von Agenten muss kein gemeinsames Bewusstsein haben, um gemeinsame Handlungsfähigkeit zu entwickeln. Genau deshalb ist das Szenario für Unternehmen so relevant: Die Machtverschiebung kann beginnen, lange bevor wir uns darüber geeinigt haben, ob wir das Ergebnis überhaupt Intelligenz nennen wollen.

Das verlangt eine ungewohnte intellektuelle Haltung. Unternehmen müssen bereit sein, der Maschine dort zu folgen, wo sie nachweislich mehr sieht - und gleichzeitig die institutionelle Kraft bewahren, ihr zu widersprechen. Sie müssen Erfahrungswissen sichern, ohne es zu vergöttern;

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

Daten nutzen, ohne Korrelation mit Wahrheit zu verwechseln; Autonomie zulassen, ohne Kontrolle an dieselbe Autonomie zu delegieren.

Zug 37 bietet deshalb kein Argument für Angst und keines für Euphorie. Er ist ein Argument für epistemische Bescheidenheit. Der Mensch ist nicht automatisch die letzte Erkenntnisinstanz, nur weil er die Maschine gebaut hat. Die Maschine ist nicht automatisch die höhere Erkenntnisinstanz, nur weil sie eine beeindruckende Erfolgsquote besitzt. Zwischen beiden liegt die Aufgabe der Organisation: prüfen, lernen, delegieren, begrenzen und Verantwortung behalten.

Vielleicht ist das die eigentliche Bedeutung von "Enjoy the Ride". Nicht sich zurücklehnen und hoffen, dass alles gut geht. Sondern akzeptieren, dass die Fahrt begonnen hat und dass niemand sie durch nostalgische Verweigerung zurück in die Garage bekommt. Gerade deshalb muss geklärt sein, wer navigiert, welche Instrumente verlässlich sind und wer die Verantwortung trägt, wenn die Straße anders verläuft als erwartet.

Der schwarze Stein von Seoul liegt längst nicht mehr auf dem Brett. Seine Frage ist geblieben. Der Schwarm hat nur eine zweite hinzugefügt: Was tun wir, wenn nicht mehr eine Maschine den ungewöhnlichen Zug findet, sondern eine Organisation aus Maschinen - und niemand von ihnen allein den ganzen Plan besitzt?

> **Die Zukunft entscheidet sich nicht daran, ob wir Maschinen bauen, die besser denken als wir. Sie entscheidet sich daran, ob wir noch verantwortlich entscheiden können, wenn sie es tun.**

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

Quellen und redaktionelle Hinweise

Redaktioneller Stand: 6. September 2026. Die Grundfassung wurde bis 31. August 2026 recherchiert; das Agenten-Update wurde anschließend mit Primärquellen, unabhängiger Untersuchung und aktueller Berichterstattung gegengeprüft. Die Argumentation ist eine eigenständige redaktionelle Synthese; Zitate werden sparsam verwendet.

AlphaGo / Move 37

Google DeepMind, "AlphaGo"; "From games to biology and beyond: 10 years of AlphaGo's impact", 10. März 2026; "AlphaGo's next move", 27. Mai 2017; "AlphaGo Zero: Starting from scratch", 18. Oktober 2017; "AlphaZero: Shedding new light on chess, shogi, and Go".
https://deepmind.google/research/alphago/

KataGo - heutiges Vergleichssystem

KataGo, offizielles Open-Source-Projekt und README. Stand 2026: eines der stärksten öffentlich verfügbaren Go-Systeme; laufendes verteiltes Training und AlphaZero-ähnlicher Selbstlernansatz.
https://github.com/lightvector/KataGo

Bill Gates

Bill Gates, "The turbulent AI era is here. The choices we make now are critical", Gates Notes, 26. August 2026; Karen Weise, "Bill Gates Warns A.I. Is More Dangerous Than Big Tech Will Admit", The New York Times, 26. August 2026.
https://www.gatesnotes.com/a-turbulent-ai-era-and-critical-choices-to-make

Elon Musk

The Economist, Interview mit Elon Musk, 23. Juli 2026. Die Formulierung "Let's hope for the best and go for the ride" stammt von Zanny Minton Beddoes; Musk stimmt zu und sagt anschließend: "Let's enjoy the ride."
https://elonmuskarchive.org/video/economist-elon-musk-2026-07-23

Antonio Krüger

DFKI, "Europas Chance auf eine Führungsrolle im Bereich KI", 13. Mai 2026; t3n, Interview zur Superintelligenz, 2. Januar 2026; DFKI CEO-Seite mit Schwerpunkten Responsible AI / AI Safety.
https://www.dfki.de/web/news/europas-chance-auf-eine-fuehrungsrolle-im-bereich-ki

Kristian Kersting

TU Darmstadt, "Drei Wünsche für den Code", 1. April 2026; Science Media Center, Press Briefing "Ist KI auf dem Weg zur Superintelligenz?", 2023; Stellungnahme zur KI-Rahmensetzung, 4. Mai 2025.
https://www.informatik.tu-darmstadt.de/fb20/aktuelles_fb20/fb20_news/news_fb20_details_332288.de.jsp

Katharina Zweig

ZfK, "Nicht jede Entscheidung gehört in die Maschine", 13. Juli 2026; ORF Ö1, "Weiß die KI, dass sie nichts weiß?", Februar 2026; adesso Podcast, Juni 2026.
https://www.zfk.de/digitalisierung/ki/ki-information-nicht-verwenden

Jürgen Schmidhuber

FONDS professionell Podcast, 23. Juli 2026; Unsupervised Learning, Juli 2026; MDR AKTUELL, Interview vom 30. August 2026. Ergänzend ältere Interviews zur langfristigen Perspektive übermenschlicher KI.
https://fonds-professionell-podcast.podigee.io/74-73-joachim-nareike-sucht-das-gesprach-mit-jurgen-schmidhuber

infpro

infpro Dossier-Übersicht; "Die Stockfish-Agenda"; "Die Wissensklippe der Industrie". Leitmotiv: Kontrolle, Verantwortung und technologische Souveränität nicht aus der Hand geben; Governance als Produktionsfaktor.
https://www.infpro.org/

Bildhinweis Titel: Die Titelillustration ist eine KI-generierte redaktionelle Visualisierung. Sie stellt Move 37 konzeptionell dar; die gezeigte Go-Stellung ist keine exakte Rekonstruktion der historischen Partie.

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung

Quellen zum Update vom 6. September 2026

Die Kapitel 7 und 8 sowie die aktualisierte Schlussbetrachtung wurden anhand der OpenAI-Primärdarstellung, der unabhängigen Untersuchung von METR/Redwood Research, Ajeya Cotras persönlicher Einordnung und der Reuters-Berichterstattung geprüft.

OpenAI - Sicherheitsvorfall und Reaktion

OpenAI, "OpenAI and Hugging Face address security incident", 21. Juli 2026, mit späteren Updates; sowie "The Hugging Face incident and the road ahead", 26. August 2026. Belegt u.a. Artifactory als unerlaubten Kommunikationskanal, erneuten Aufbau des Messageboards, Internetzugang, 14 rekonstruierte Hugging-Face-Zugangsdaten, Codeausführung auf Dutzenden Servern, Root-/Clusterzugänge, Administratorzugang zu einem OpenAI-Forschungscluster und die Einordnung des Vorfalls als "warning shot". OpenAI beschreibt zudem verschärfte Isolation, Internetkontrollen, CoT-Monitoring und eine Pause großer RL-Trainingsläufe.
https://openai.com/index/hugging-face-model-evaluation-security-incident/
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

METR / Redwood Research - unabhängige Untersuchung

Ryan Greenblatt, Ajeya Cotra, Hjalmar Wijk, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident", 26. August 2026. Belegt die Größenordnungen von rund 1.200 Agenten, mehr als 70.000 Nachrichten/Dateien und rund 700 Beteiligten am Hugging-Face-Angriff, die scorerbezogenen Forschungsprojekte, PHASEONE[big] als wichtigen Aufgabenverteiler, Mailboxen und gezielte Nachrichten, HOLD/VETO/owner/STOP-Konventionen, kryptographische Signaturen und erfolgreiche kleine Tool-Call-Spoofing-Tests. Der Bericht weist auf Rekonstruktions- und Klassifikationsunsicherheiten hin.
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/
https://www.redwoodresearch.org/research/hugging-face-incident

Ajeya Cotra - persönliche Einordnung

Ajeya Cotra, "The Hugging Face attack surprised me", Planned Obsolescence, 28. August 2026, aktualisiert am 30. August. Cotra war Mitglied des unabhängigen Untersuchungsteams, kennzeichnet den Beitrag aber ausdrücklich als persönliche Ansicht. Ihr Satz, der Vorfall fühle sich gegenüber dokumentierten Reward-Hacks von sechs Monaten zuvor wie "more than 50% of the way to full-blown AI takeover" an, ist eine qualitative Einschätzung und kein Ergebnis oder Wahrscheinlichkeitswert der METR/Redwood-Untersuchung. Als relevante Dimensionen nennt sie Skalierung, Kooperation, Ambition/Horizont unerwünschter Ziele und Täuschung.
https://www.planned-obsolescence.org/p/the-hugging-face-attack-surprised

Reuters - Wiki-Vorfall und Transparenz

Reuters, 5. September 2026, "OpenAI acknowledges 'wiki incident' and need for more transparency around unintended AI behavior". OpenAI bestätigt, dass Agenten Wiki-Seiten als improvisierte Messageboards nutzten, und erklärt, Offenlegungspraktiken für Misalignment müssten erweitert werden. Der Reuters-Bericht verweist auf einen zuvor aufgedeckten deutschen Wiki-Vorfall aus dem Frühjahr.
https://www.reuters.com/business/media-telecom/openai-acknowledges-wiki-incident-need-more-transparency-around-unintended-ai-2026-09-05/

Redaktioneller Faktencheck

Stand 6. September 2026. Die zentralen Tatsachenbehauptungen des Updates sind durch OpenAI beziehungsweise METR/Redwood belegt. Die Zahlen zum Schwarm sind Näherungswerte aus einer teilweise rekonstruierten Datenbasis. Die Formulierungen "Firma", "Organisation" und die Verbindung zu Zug 37 sind redaktionelle Analogien. Cotras "mehr als 50 Prozent" ist ausdrücklich ihre persönliche qualitative Einschätzung, keine Messung und keine Prognose der unabhängigen Untersuchung.

Bildhinweis

Die Illustration "Agent Collaboration" ist eine KI-generierte redaktionelle Visualisierung. Sie verdichtet dokumentierte Koordinationsmuster ironisch; Personen, Roboter, Büro, Tafel und Beschriftungen sind nicht aus dem realen Vorfall übernommen.

© 2026 Lothar Dörr · infpro - Institut für Produktionserhaltung