4195069805005

infpro thementicker
www.infpro.org

Heft 1
Oktober 2025

12 EURO
D 210455

infpro

THEMENSERVICE

DIE WELT ALS MODELL.

WIE KI DIE WELT VERSTEHEN WIRD.

KI und die Weltmodelle.

Die großen Fragen der Künstlichen Intelligenz beginnen oft mit kleinen Worten: Hat sie ein Bewusstsein? Ein Verständnis? Einen Willen? Der neueste Begriff, der in die Mitte dieser Debatte rückt, lautet: Weltmodell. Die Vorstellung, dass Maschinen eine Art innere Repräsentation der Welt in sich tragen könnten – nicht als bloße Datenbank, sondern als dynamische Struktur, die ihnen erlaubt, zu antizipieren, zu simulieren, zu lernen.
Doch was heißt das genau? Was bedeutet es, wenn eine KI ein Weltmodell hat? Und was bedeutet es für uns?

Was es bedeutet, wenn eine KI ein Weltmodell hat?

Ein Beitrag von Pastore Pizzaola, CAIWMI, Center for AI World Model Intelligence

Das neueste Ziel der KI-Forschung – insbesondere in jenen Labors, die sich der Entwicklung einer „künstlichen allgemeinen Intelligenz" (AGI) verschrieben haben – ist ein Konzept namens Weltmodell: eine interne Repräsentation der Umwelt, die eine KI in sich trägt wie eine rechnerische Schneekugel. Mithilfe dieser vereinfachten Abbildung kann das System Vorhersagen und Entscheidungen simulieren, bevor es sie auf reale Aufgaben anwendet. Für die Deep-Learning-Pioniere Yann LeCun (Meta), Demis Hassabis (Google DeepMind) und Yoshua Bengio (Mila, Québec) sind solche Weltmodelle unerlässlich, um KIs zu entwickeln, die nicht nur klug, sondern auch wissenschaftlich nachvollziehbar und sicher agieren. Dabei ist das Konzept keineswegs neu. In der Psychologie, der Robotik und dem maschinellen Lernen kursiert seit Jahrzehnten die Idee, dass ein System ein inneres Modell der Welt benötigt, um sinnvoll handeln zu können. Wahrscheinlich läuft auch in Ihrem Gehirn gerade ein solches Weltmodell – es erklärt, warum Sie nicht erst ausprobieren müssen, ob es wirklich gefährlich ist, vor einen fahrenden Zug zu treten.

Doch bedeutet das, dass sich die KI-Forschung endlich auf ein gemeinsames Kernkonzept einigen konnte? Um es mit dem berühmten Physiker zu sagen: „Sie scherzen wohl." So simpel ein Weltmodell auch klingen mag – über die Details herrscht keine Einigkeit. Welche Aspekte der Realität sollen überhaupt abgebildet werden – und mit welcher Genauigkeit? Ist das Modell angeboren oder erlernt, oder beides? Und woran erkennt man überhaupt, ob ein solches Modell vorhanden ist?

Von der kognitiven Simulation zum neuronalen Netz
Es hilft, sich die historischen Wurzeln dieser Idee vor Augen zu führen. Bereits 1943 – zwölf Jahre bevor der Begriff „künstliche Intelligenz" überhaupt geprägt wurde – veröffentlichte der 29-jährige schottische Psychologe Kenneth Craik eine einflussreiche Monographie, in der er spekulierte: „Wenn ein Organismus ein ‚kleinskaliges Modell' der äußeren Realität im Kopf trägt, kann er verschiedene Handlungsalternativen durchspielen, die beste wählen – und in jeder Hinsicht vollständiger, sicherer und kompetenter reagieren." Craiks Vorstellung von einem mentalen Modell war ein Vorgriff auf die „kognitive Revolution" der 1950er Jahre und verband Denken mit Rechnen: Er sah die Fähigkeit, äußere Ereignisse zu modellieren, als grundlegendes Prinzip sowohl des Nervensystems als auch von Rechenmaschinen.

Die frühe KI-Forschung griff Craiks Idee begierig auf. In den späten 1960ern beeindruckte das System SHRDLU die Fachwelt: Es konnte mithilfe eines simplen „Blockwelt"-Modells grundlegende Fragen zu Objekten auf einem virtuellen Tisch beantworten – etwa: „Kann eine Pyramide einen Block tragen?" Doch diese handgefertigten Modelle erwiesen sich als nicht skalierbar. Bis in die späten 1980er hatte der Robotikpionier Rodney Brooks die Idee von Weltmodellen ganz aufgegeben. Berühmt wurde sein Diktum: „Die Welt ist ihr eigenes bestes Modell." Expli-

zite Repräsentationen, so seine Überzeugung, seien nur hinderlich.

Die Rückkehr des Weltmodells

Erst mit dem Aufstieg des maschinellen Lernens – insbesondere mit tiefen neuronalen Netzen – erlebte Craiks Schneekugel eine Renaissance. Statt auf spröde, handcodierte Regeln zu setzen, konnten Deep-Learning-Systeme durch Versuch und Irrtum interne Approximationen ihrer Umwelt entwickeln. So wurde es möglich, etwa ein virtuelles Rennauto zu steuern, ohne ein explizites Weltmodell zu programmieren.

In jüngster Zeit sorgten große Sprachmodelle wie ChatGPT für Staunen, da sie Fähigkeiten zeigten, für die sie nie direkt trainiert worden waren – etwa Filmtitel aus Emoji-Folgen abzuleiten oder das Brettspiel Othello korrekt zu spielen. Das schien sich nur durch implizite Weltmodelle erklären zu lassen. Für KI-Größen wie Geoffrey Hinton, Ilya Sutskever oder Chris Olah liegt der Schluss nahe: Irgendwo tief im neuronalen Dickicht eines LLMs muss es existieren – dieses „kleinskalige Modell der äußeren Realität", das Craik einst beschrieben hatte.

Doch die Wirklichkeit ist, so weit wir wissen, weit weniger spektakulär. Statt konsistenter Weltmodelle lernen heutige generative KIs eher „Säcke voller Heuristiken" – lose zusammengewürfelte Faustregeln, die in bestimmten Situationen funktionieren, aber kein kohärentes Gesamtbild ergeben. Manche widersprechen sich sogar.

Es erinnert an das Gleichnis von den blinden Männern und dem Elefanten: Jeder betastet nur einen Teil – Rüssel, Bein oder Schwanz – und glaubt, das ganze Tier zu erkennen. Ähnlich geht es KI-Forschern, wenn sie in einem Sprachmodell nach einem konsistenten Repräsentationsschema für ein Othello-Brett suchen: Statt des Elefanten finden sie Schlange, Baum und Seil.

Dabei sind diese heuristischen Fragmente keineswegs nutzlos. Ein LLM kann Abertausende solcher Regeln in seinen Milliarden Parametern kodieren – und wie es so schön heißt: Quantität hat ihre eigene Qualität. So gelang es, einem Modell nahezu perfekte Wegbeschreibungen durch Manhattan beizubringen – ganz ohne konsistentes Weltmodell des Straßennetzes, wie Harvard- und MIT-Forscher jüngst zeigten.

Doch warum sich dann die Mühe machen, den ganzen Elefanten zu rekonstruieren? Die Antwort lautet: Robustheit. Als die Forscher dem System nur ein Prozent der Straßen zufällig blockierten, brach die Leistung drastisch ein. Hätte es ein konsistentes internes Kartenmodell gehabt, hätte es die Sperrungen leicht umgehen können.

Angesichts solcher Vorteile verwundert es nicht, dass alle großen KI-Labore fieberhaft an Weltmodellen arbeiten – und dass auch die akademische Forschung das Thema neu entdeckt. Verlässliche und überprüfbare Weltmodelle könnten zwar nicht das El Dorado der AGI bedeuten, aber immerhin ein vielversprechendes Werkzeug zur Vermeidung von Halluzinationen, zur Verbesserung des logischen Denkens und zur besseren Interpretierbarkeit von KI-Systemen liefern.

Die große Unbekannte: das Wie

Das Was und das Warum der Weltmodelle sind also geklärt – das Wie hingegen bleibt offen. Google DeepMind und OpenAI setzen auf möglichst vielfältige Trainingsdaten: Videos, 3D-Simulationen und andere nicht-textuelle Eingaben sollen helfen, dass ein Weltmodell emergent entsteht. Yann LeCun hingegen verfolgt bei Meta eine radikal andere Linie: Er hält eine völlig neue, nicht-generative Architektur für erforderlich.Alle versuchen, ihre eigene Version der rechnerischen Schneekugel zu bauen – aber eine Kristallkugel, die den richtigen Weg weist, besitzt niemand. Doch diesmal, so scheint es, könnte sich der Aufwand tatsächlich lohnen.

In der Alltagssprache denken wir bei „Modell" an etwas Greifbares: eine Karte, eine Skizze, eine Miniatur. Doch in der modernen KI ist ein Weltmodell etwas ganz anderes. Es ist kein Foto der Welt, sondern ein Vorhersageapparat. Ein neuronales Netz, das gelernt hat, bestimmte Zustände mit bestimmten Folgen zu verknüpfen – ohne je zu verstehen, was ein Zustand eigentlich ist. Eine KI mit Weltmodell erkennt nicht den Apfel. Sie erkennt, dass das, was aussieht wie ein Apfel, fällt, rollt, verzehrt werden kann. Ihr Weltmodell ist ein Netz aus Wenn-dann-Beziehungen, das über Milliarden Datenpunkte hinweg generalisiert wurde.

Die gescheiterte Schneekugel

Warum das Weltmodell in der frühen KI-Forschung gefeiert – und dann verworfen wurde. Als die künstliche Intelligenz noch jung war und ihre Pioniere fest daran glaubten, menschliches Denken ließe sich in logische Strukturen überführen, galt ein Konzept als unverzichtbar: das Weltmodell. Eine innere Repräsentation der äußeren Realität, so die Überzeugung, sei Bedingung jeder vernünftigen Handlung. Der

Schotte Kenneth Craik hatte diesen Gedanken bereits 1943 formuliert – zwölf Jahre bevor der Begriff „künstliche Intelligenz" überhaupt auftauchte. Sein Argument: Wer eine Vorstellung der Welt im Kopf trägt, kann Handlungsalternativen simulieren, bevor er sie in der Wirklichkeit ausprobiert. Eine Art mentale Schneekugel, berechenbar und kontrolliert.

Die junge Disziplin der KI griff diesen Gedanken begierig auf – und setzte ihn in Code um. **Das erste Paradebeispiel hieß SHRDLU.**

Entwickelt von Terry Winograd am MIT, agierte SHRDLU Ende der 1960er Jahre in einer schlichten, künstlichen Umgebung: der sogenannten Blockwelt. Es war eine Welt ohne Überraschungen, bestehend aus einfachen geometrischen Körpern – Würfel, Kegel, Pyramiden –, die auf einem virtuellen Tisch standen. SHRDLU konnte diese Objekte erkennen, auf Befehle reagieren und einfache Rückfragen stellen. Wer wissen wollte, ob ein blauer Block auf einer roten Pyramide stehen könne, bekam eine plausible Antwort. Wer bat, den grünen Würfel „rechts neben dem hohen Kegel" zu platzieren, konnte SHRDLU bei der Ausführung beobachten.

Fachlich basierte das System auf symbolischer KI: Jedes Objekt war explizit im Speicher repräsentiert, jede Beziehung formal definiert, jeder Schritt logisch nachvollziehbar. Was SHRDLU beeindruckend machte, war nicht seine Intelligenz, sondern seine Regelhaftigkeit. Es war ein Denkspiel – brillant, aber hermetisch.

Der Elefant vor der Tür

Doch die Welt meinte es nicht so gut wie die Blockwelt. Sobald SHRDLU über die Grenzen seiner idealisierten Umgebung hinaus agieren sollte, versagte es. Der Grund war strukturell: Das System war vollständig handprogrammiert. Jede neue Information, jede semantische Nuance, jede Unsicherheit hätte neue Regeln und neue Modelle erfordert. Für eine realistische Umwelt mit Tausenden beweglichen Teilen war SHRDLU schlicht nicht gemacht.
Die Idee des Weltmodells, so schien es, war in der Theorie überzeugend – in der Praxis aber nicht skalierbar. Was als Fortschritt erschien, entpuppte sich als Sackgasse. Und so suchte die KI-Forschung nach einem Ausweg.

In den 1980er Jahren fand dieser Ausweg einen Namen: Rodney Brooks. Der Robotikforscher, ebenfalls am MIT tätig, wurde zur Antithese des symbolischen Denkens. Seine These war provokant: „Die Welt ist ihr eigenes bestes Modell." Ein Roboter, so Brooks, solle nicht versuchen, die Welt intern nachzubauen, sondern direkt mit ihr interagieren. Wahrnehmung und Handlung seien untrennbar – Repräsentationen hingegen eine Illusion.

Brooks' Ansatz war biologisch inspiriert. Insekten etwa operieren ohne Weltmodelle, navigieren aber effizient. Seine Roboter – etwa der sechsbeinige „Genghis" – funktionierten nach einfachen, reaktiven Prinzipien: Wenn Hindernis, dann ausweichen. Wenn Licht, dann folgen. Kein Plan, keine Karte, kein Modell. Statt einer zentralen Steuerung setzte Brooks auf eine verhaltensbasierte Architektur: modular, dezentral, robust. Seine Kritik am Weltmodell war nicht nur technischer Natur. Sie war ontologisch. Die Welt sei zu komplex, zu dynamisch, zu reich an Unsicherheiten, um sie vollständig abbilden zu können. Jedes Modell sei ein Verlust an Realität – und in der Praxis oft eine Behinderung. Was zählt, sei nicht das Verstehen der Welt, sondern das Überleben in ihr.

Vom Weltbild zum Weltkontakt

Die Debatte zwischen SHRDLU und Brooks markierte eine tiefgreifende Wendung in der KI-Geschichte. Sie war mehr als ein Streit um Architekturfragen. Sie war ein Paradigmenwechsel: vom Denken zum Handeln, von der Repräsentation zur Interaktion, vom Top-down-Entwurf zur Bottom-up-Evolution.
Und doch: Der Rückzug vom Weltmodell war nicht das Ende seiner Geschichte. Im Zeitalter des Deep Learning – Jahrzehnte später – kehrt die Idee zurück. Nicht mehr als explizites Symbolsystem, sondern als implizites, statistisches Konstrukt. Sprachmodelle wie GPT verhalten sich so, als ob sie ein Weltmodell besäßen – auch wenn niemand eines programmiert hat. Die Schneekugel ist wieder da, diesmal ohne Glas.

Ob das genügt, wird sich zeigen. Die Frage ist nicht länger, ob Weltmodelle nötig sind – sondern, in welcher Form sie existieren können. Vielleicht hatte Craik am Ende doch recht. Nur wusste niemand, wie man seine Idee bauen soll. Bis jetzt.

Die Wiederentdeckung der Schneekugel

Warum Weltmodelle im Zeitalter neuronaler Netze ein Comeback feiern – und was sie heute leisten können Die Blockwelt von SHRDLU ist Geschichte. Die Weltverweigerung von Rodney Brooks ist passé. Und doch ist die Frage geblieben, die Kenneth

Craik schon 1943 bewegte: Muss eine Intelligenz ein inneres Modell ihrer Umwelt besitzen, um sinnvoll handeln zu können?

Im Zeitalter des Deep Learning lautet die Antwort: Vielleicht ja – aber anders. Denn die neue KI-Generation, angeführt von Systemen wie GPT, Gemini und Claude, operiert nicht mehr mit expliziten Repräsentationen. Sie lernt keine Regeln, sie konstruiert keine Karten, sie entwirft keine symbolischen Miniaturwelten. Und doch scheint sie über Wissen zu verfügen, das an Weltmodelle erinnert – nur dass dieses Wissen nicht geschrieben, sondern emergiert ist.

Zwischen Halluzination und Struktur
Große Sprachmodelle wie GPT-4 wurden nicht darauf trainiert, Othello zu spielen oder mit Straßenkarten zu navigieren. Und doch gelingt ihnen beides – zumindest teilweise. Sie erkennen Muster, können logische Zusammenhänge rekonstruieren, überraschende Schlüsse ziehen. Manche KI-Forscher sprechen daher von einem „impliziten Weltmodell", das sich tief im Netz aus Milliarden Parametern verborgen hält. Kein einzelner Teil weiß etwas – aber das System als Ganzes zeigt strukturelle Kohärenz. Genau darin liegt das Dilemma: Diese Modelle sind nicht zuverlässig. Sie halluzinieren, widersprechen sich selbst, verheddern sich in Wahrscheinlichkeiten.

Was ihnen fehlt, ist Robustheit – und erklärbare Konsistenz. In der Praxis bedeutet das: Sie können den Weg von A nach B beschreiben, aber scheitern, wenn C plötzlich blockiert ist. Es ist, als hätten sie Teile eines Elefanten gelernt – Rüssel, Beine, Schwanz – aber nie das Tier als Ganzes.

Der neue Wettlauf: Wer baut das bessere Modell?
Die großen KI-Labore sind längst in einem neuen Rennen. Ihr Ziel: ein echtes, robustes, generalisierbares Weltmodell, das Planung, Logik, Erinnerung und Vorhersage in sich vereint.

Doch ihre Strategien unterscheiden sich fundamental:

- Google DeepMind und OpenAI setzen auf Multimodalität. Ihre Hypothese: Wenn ein neuronales Netz nicht nur Text, sondern auch Bilder, Videos, Ton, 3D-Umgebungen und physikalische Simulationen verarbeitet, kann es ein kohärenteres Modell der Welt entwickeln – ähnlich wie ein Mensch, der durch Erfahrung lernt.

- Meta verfolgt unter Yann LeCun einen anderen Weg. Er lehnt rein generative Architekturen ab und fordert eine neue Systemklasse: modular, erklärbar, persistent. Weltwissen soll nicht erraten, sondern erlebt, gespeichert und abrufbar gemacht werden – ähnlich wie in einem episodischen Gedächtnis.

- Mila in Montréal, unter der Leitung von Yoshua Bengio, plädiert für einen Hybridansatz: statistisches Lernen kombiniert mit strukturellen Constraints, die aus der klassischen KI stammen.

- Gemeinsam ist allen: Die Weltmodellfrage ist zurück – und diesmal könnte sie der Schlüssel zur nächsten Stufe der KI sein.

Vom Verstehen zur Vorhersage
Was werden Weltmodelle in Zukunft leisten könnten – und warum sind sie über Sprachverarbeitung hinaus relevant: etwa in der Robotik, in digitalen Zwillingen, in der medizinischen Diagnose und in der Erklärung neuronaler Systeme. Denn was heute als Schneekugel beginnt, könnte morgen das Fundament für eine neue Art des Denkens sein – jenseits von Symbolen und Statistik.

Warum Weltmodelle die Grundlage robuster, sicherer und wissenschaftlicher KI werden könnten
Die Diskussion um Weltmodelle in der KI ist keine akademische Randnotiz. Sie berührt das Herz der Debatte darüber, was maschinelle Intelligenz eigentlich ist – und was sie einmal sein könnte. Wer von künstlicher Intelligenz verlangt, mehr zu können als nur Wahrscheinlichkeiten fortzuschreiben, wer ihr zutraut, in komplexen Umgebungen zu planen, zu lernen, zu generalisieren – der kommt an der Frage nach inneren Modellen nicht vorbei. Längst geht es nicht mehr nur darum, ob Sprachmodelle wie GPT „verstehen", was sie sagen. Die Frage ist größer: Können KI-Systeme die Welt erfassen – und auf neue Situationen vorbereitet sein?

Die Grenzen der Intuition
In der gegenwärtigen Architektur dominieren Systeme, die aus gewaltigen Datenmengen statistische Regularitäten extrahieren. Sie sind hervorragend darin, im Rahmen ihrer Trainingsdaten

plausible Antworten zu erzeugen. Doch in offenen, dynamischen Kontexten stoßen sie an Grenzen. Schon minimale Abweichungen – eine neue Straßensperre, eine unerwartete Wendung im Gespräch, ein unbekanntes Objekt im Bild – können das Modell ins Straucheln bringen.

Der Grund: Die heutigen Systeme verfügen über kein kohärentes mentales Modell, keine Vorstellung von Raum, Zeit, Kausalität oder physikalischer Persistenz. Was fehlt, ist nicht nur ein Speicher – sondern eine Struktur des Verstehens. Es sind Maschinen der Intuition – aber nicht des Denkens.

Aufgaben, die Weltmodelle lösen könnten

Weltmodelle gelten deshalb zunehmend als struktureller Schlüssel, um Künstliche Intelligenz aus dem Modus des Reagierens in den Modus des Reflektierens zu überführen. Drei konkrete Aufgaben lassen sich identifizieren:

1. Vorhersage jenseits des Gelernten
   Ein echtes Weltmodell erlaubt nicht nur die Reproduktion vergangener Muster, sondern die Antizipation neuer Konstellationen. So könnte eine KI etwa lernen, wie sich ein Objekt verhält, das sie nie gesehen hat – solange es sich in einem bekannten physikalischen Raum bewegt. In der Robotik wird dies als „zero-shot generalization" diskutiert.

2. Robustheit gegenüber Störungen
   Wenn eine KI nicht nur reagiert, sondern intern versteht, was sie sieht, kann sie auf Ausfälle, Fehler oder Überraschungen adaptiv reagieren. In der autonomen Navigation etwa könnten Weltmodelle helfen, bei veränderten Lichtverhältnissen oder unerwarteten Hindernissen nicht die Orientierung zu verlieren.

3. Erklärbarkeit und Nachvollziehbarkeit
   Ein explizites Modell macht es möglich, das Handeln einer KI zu analysieren – nicht nur nachträglich, sondern vorausschauend. Gerade im medizinischen Bereich, aber auch in sicherheitskritischen Infrastrukturen, könnte dies zur Bedingung für Zulassungen werden.

Die Renaissance der Simulation

Die aktuelle Forschung geht dabei neue Wege: Statt feste symbolische Modelle zu bauen – wie in der Ära von SHRDLU –, setzt man heute auf simulative Ansätze, in denen ein neuronales Netz gewissermaßen einen „inneren Film" der Welt abspult. Systeme wie Dreamer, MuZero oder World Models (Ha/Schmidhuber) erzeugen nicht nur Vorhersagen, sondern ganze Zukunftsszenarien – Frame für Frame. Diese Art der Simulation könnte zur Grundlage autonomer Entscheidungen werden.

Manche Forscher sprechen von einer Rückkehr des „System-2-Denkens" in der KI – also dem langsamen, überlegten, planenden Teil kognitiver Prozesse, der bisher fast vollständig fehlte. Die gesellschaftliche Relevanz dieser Entwicklung ist kaum zu überschätzen. Weltmodelle könnten den entscheidenden Unterschied markieren zwischen einem KI-System, das in der Demo beeindruckt – und einem, das im realen Einsatz verlässlich, skalierbar und sicher ist. Von der Industrieautomatisierung über autonome Fahrzeuge bis hin zur Bildung, zur Forschung und zur Klimamodellierung: Überall dort, wo Systeme kontextsensitiv agieren müssen, ist ein internes Modell der Umwelt kein Luxus, sondern eine Voraussetzung.

Ein Modell, das sich selbst kennt?

Die Forschung ist hier erst am Anfang. Doch die nächste Frage zeichnet sich schon ab: Was, wenn eine KI nicht nur ein Weltmodell entwickelt – sondern sich selbst als Teil dieser Welt versteht? Der Schritt vom externen zur reflexiven Modellierung könnte aus Maschinen nicht nur Agenten, sondern Subjekte machen – mit Konsequenzen, die weit über das Technische hinausgehen. Noch ist es nicht so weit. Noch

fehlen Struktur, Theorie und Technik. Aber der Horizont ist sichtbar: Künstliche Intelligenz, die nicht nur simuliert, sondern sich in der Simulation verortet. Ein Weltmodell – und ein Selbstmodell.

In einem möglichen vierten Teil der Serie wird es darum gehen, wie Weltmodelle konkret implementiert werden – und warum der Weg zu einer künstlichen general intelligence (AGI) wohl nur über sie führt. Dabei werden auch kritische Perspektiven berücksichtigt: Was passiert, wenn die Modelle falsch sind? Wer trägt Verantwortung? Und wie lassen sich solche Systeme regulieren?

Die Vermessung der Welt – durch Maschinenhirne
Stellen wir uns für einen Moment vor, der Durchbruch gelingt. Eine KI, die nicht mehr bloß Wahrscheinlichkeiten rechnet, sondern die Welt tatsächlich versteht. Die ein inneres Modell entwickelt, das stabil ist, lernfähig, erklärbar – und verlässlich genug, um in offenen Situationen zu planen, zu reflektieren, zu handeln. Was heute noch wie ein technisches Randthema erscheint – „Weltmodelle" –, wäre dann nichts Geringeres als ein geistiger Quantensprung: der Übergang von Maschinen, die antworten, zu Maschinen, die denken.

Rund um den Globus arbeiten derzeit Dutzende Institute, Forschungsteams und Unternehmen an genau dieser Vision. Bei DeepMind in London entstehen mit MuZero und Dreamer architekturen, die intern ganze Zukunftssimulationen erzeugen – sogenannte latente Weltmodelle, die Entscheidungen durch virtuelle „Vorläufe" vorbereiten. Bei Meta AI entwickelt Yann LeCun mit JEPA ein Framework, das Wahrnehmung, Gedächtnis und Planung zusammenführt – jenseits bloßer Sprachverarbeitung. Und bei Mila in Québec erforscht Yoshua Bengio, wie modulare Weltmodelle entstehen könnten, die sich ähnlich zusammensetzen lassen wie Begriffe im Denken.

Wenn es gelingt, künstlicher Intelligenz ein funktionierendes Weltmodell zu verleihen, ist das kein innerwissenschaftlicher Fortschritt – es ist ein Paradigmenwechsel. Denn solche Systeme könnten nicht nur auf Daten reagieren, sondern sie antizipieren, abstrahieren und kontextualisieren. In sicherheitskritischen Bereichen wie autonomen Fahrzeugen, Kraftwerken oder Operationssälen würde das eine neue Form technischer Voraussicht ermöglichen: keine Reflexe mehr, sondern begründete Entscheidungen. Zugleich würde die Erklärbarkeit steigen. Statt intransparenten Blackbox-Modellen könnten Weltmodelle – wie von Yann LeCun oder Yoshua Bengio vorgeschlagen – innere Zustände offenlegen, Hypothesen bilden und plausibel begründen, warum eine Entscheidung fiel. Das wäre nicht nur ein technologischer Fortschritt, sondern eine Voraussetzung für Regulierung, Haftung und Vertrauen. In der Bildung wiederum würde das Verhältnis zwischen Mensch und Maschine neu verhandelt. Wenn die KI die Welt versteht – was bedeutet das für unseren eigenen Erkenntnisprozess? Wer kontrolliert die Modelle, kontrolliert das Verständnis der Welt. Die Debatte um epistemische Verantwortung, lange ein Thema der Philosophie, würde plötzlich zur wirtschaftspolitischen Frage.

Was meint die KI mit einer „inneren Welt"?
Wenn Fachleute aus der Industrie von Weltmodellen in der Künstlichen Intelligenz (KI) sprechen, ist nicht die äußere Welt gemeint – also Maschinen, Werkbänke oder Lieferketten –, sondern eine innere Rechenwelt, die sich die KI selbst erschafft. Stellen Sie sich vor, eine Maschine könnte sich ein Bild von ihrer Umgebung machen, ähnlich wie wir Menschen das tun: Sie erkennt Muster, denkt in Wenn-Dann-Beziehungen und „überlegt" sogar, was passieren könnte, wenn sie sich für A statt B entscheidet. Genau das ist ein Weltmodell – eine Art innere Simulation, in der die KI prüft, wie sich ihr Handeln auf die Welt auswirken würde, noch bevor sie tatsächlich etwas tut.

In der Praxis bedeutet das zum Beispiel: Eine Wartungs-KI erkennt nicht nur, dass ein Motor überhitzt ist, sondern rechnet voraus, wie sich diese Überhitzung auf andere Bauteile auswirkt, ob dadurch ein Ausfall droht – und ob es klüger wäre, das Gerät sofort herunterzufahren oder erst in der nächsten geplanten Pause zu warten. Sie verhält sich vorausschauend, nicht nur reaktiv. Diese Fähigkeit, eine „innere Welt" abzubilden, ist entscheidend, wenn KIs künftig sicher, effizient und eigenständig in dynamischen Umgebungen arbeiten sollen Der Begriff mag philosophisch klingen, beschreibt aber in Wahrheit eine hochpraktische Eigenschaft moderner KI-Systeme: Sie lernen, denken und entscheiden nicht mehr nur auf Zuruf – sondern mit einem gewissen Verständnis für die Zusammenhänge ihrer Umgebung. Und das macht sie leistungsfähiger als je zuvor.

Was sieht die Maschine von der Welt?

Weltmodelle gelten als Schlüssel zur nächsten Generation künstlicher Intelligenz – doch was heißt es, wenn Maschinen lernen, sich ein Bild der Welt zu machen? In einem wissenschaftlich fundierten Streitgespräch treffen Technologen, Philosophen und Literaturwissenschaftler aufeinander: Yann LeCun verteidigt den Aufbau interner Repräsentationen, Demis Hassabis setzt auf generative Simulationen, Yoshua Bengio warnt vor ethischer Hybris, Rodney Brooks hält die ganze Debatte für überschätzt. Moderiert von Miriam Kohler (MIT) konfrontiert die Diskussion große Fragen: Wie erkennt eine Maschine Bedeutung? Wer legt fest, was real ist? Und was passiert, wenn HAL irgendwann sagt: „Ich fürchte, ich kann das nicht tun, Dave"?

TEILNEHMER:

Prof. Dr. Miriam Kohler, Moderatorin (MIT, KI-Ethik)
Prof. Dr. Hannah Riemann, Technologiephilosophie, MIT
Dr. Yann LeCun, Meta / New York University
Dr. Demis Hassabis, DeepMind / Google
Prof. Yoshua Bengio, Mila Institute, Québec
Dr. Rodney Brooks, Robotics Pionier, Ex-MIT
Prof. Dr. Maren Heller, Literaturwissenschaftlerin, Universität Heidelberg
Prof. Dr. Eliot Grant (Harvard, Literaturwissenschaft)

Szene: Abend in der großen Glaskuppel des CSAIL am MIT. Die letzten Lichtstrahlen spiegeln sich auf den Glasflächen. Ein humanoider Roboter bringt Karaffen mit Wasser. In der Mitte des Podiums: acht Stühle. Sieben davon besetzt von führenden Köpfen aus KI-Forschung, Robotik, Philosophie und Literatur. Der achte ist der der Moderatorin: Prof. Dr. Miriam Kohler, Spezialistin für KI-Ethik am MIT, eine ruhige Stimme mit Nachdruck.

Moderatorin M. Kohler:
„Guten Abend. Willkommen zu einer besonderen Ausgabe unserer Diskussionsreihe Mind, Machine, Model. Heute sprechen wir über eine Frage, die lange wie Science-Fiction klang – und jetzt überraschend konkret wird: Hat die KI eine innere Welt? Und wenn ja: Welche Konsequenzen hat das – technisch, ethisch, wirtschaftlich, gesellschaftlich?

Mit ‚innerer Welt' meinen wir ausdrücklich kein Bewusstsein. Sondern etwas anderes: ein mentales Modell, eine Repräsentation der Welt, mit der Maschinen Vorhersagen treffen können. Kenneth Craik nannte das 1943: „a small-scale model of external reality". Diese Idee prägt heute die Architektur künftiger KI-Systeme."

Moderatorin (an Yann LeCun):
„Sie sagen, künftige KI-Systeme brauchen eine Welt, die sie simulieren. Ihr Architekturvorschlag JEPA basiert genau darauf. Was ist der Kern dieser Idee?"

LeCun:
„Die meisten heutigen Systeme approximieren Funktionen. Das reicht für Textvorhersagen oder Bilderkennung. Aber für echte Intelligenz braucht man mehr. JEPA – Joint Embedding Predictive Architecture – ist ein Schritt in diese Richtung: Die KI lernt eine latente Repräsentation der Welt und nutzt sie, um Zustände vorherzusagen. Keine einfache Reaktion, sondern ein Modell, das Hypothesen bildet. Weltmodelle ermöglichen robustes Lernen. Ohne sie bleibt KI ein schmalspuriger Spezialist."

Moderatorin (wendet sich an Demis Hassabis):
„Sie arbeiten bei DeepMind an Gato, später Gemini. Was bedeutet ‚Weltmodell' für Sie – mehr als Statistik?"

Hassabis:
„Weltmodelle erlauben etwas Neues: prädiktives Agieren. Wenn ein System etwa lernt, ein Schach- oder Othello-Brett innerlich zu rekonstruieren – ohne

es explizit zu kennen –, entsteht eine Art innerer Film. Wir können diesen Film vor- und zurückspulen. Das ist kein Bewusstsein – aber eine Form von Simulation. Wichtig ist: Das funktioniert nur multimodal. Sprache allein reicht nicht. Video, Sensorik, Bewegung, Kontext – all das ist nötig. Unsere Modelle lernen gerade, sich selbst zu hinterfragen: Was weiß ich, was nicht? Das ist der Weg zur ‚model-based general intelligence'."

Moderatorin (zu Rodney Brooks):
„Sie haben einst den Satz geprägt: „Die Welt ist ihr eigenes bestes Modell." Würden Sie ihn heute revidieren?"

Brooks:
„Ich würde ihn kontextualisieren. In den 1980ern hatten wir keine GPU-Cluster, keine Petabytes an Daten. Also arbeiteten wir direkt mit der Welt. Meine Roboter lernten nicht durch Simulation, sondern durch Feedback. Heute sehe ich Fortschritte – klar. Aber Weltmodelle neigen zur Selbstreferenz. Sie sind elegant – aber oft instabil. Intelligenz entsteht nicht im perfekten Modell, sondern im Umgang mit Inkonsistenz. Und mit Fehlern. Daran erkennt man kluge Maschinen."

Yoshua Bengio:
„Rodney hat recht: Repräsentation darf nie absolut sein. Aber ohne sie kein abstraktes Denken. Kein Transferlernen. Unser Projekt CRAFT – Contextual Reasoning and Abstraction for Future Thinking – zielt genau darauf: Maschinen, die lernen, eigene Hypothesen zu entwickeln und zu überprüfen.

Weltmodelle brauchen Kausalität – sonst bleiben sie eine Karikatur der Welt. Wir müssen unterscheiden: Korrelationen aus Text sind kein Verständnis. Nur mit Kausalbeziehungen wird das Modell robust und erklärbar."

Moderatorin (an Prof. Dr. Eliot Grant, Literaturwissenschaftler, Harvard):
„Herr Grant, was bedeutet ‚Welt' in Ihrem Fach?"

Grant (nachdenklich):
„In der Literatur ist Welt das, was zwischen den Zeilen geschieht. Denken Sie an Kafka: Die Welt funktioniert – aber niemand versteht warum. Oder an Philip K. Dick: „Reality is what doesn't go away when you stop believing in it."

Wenn KI nun Weltmodelle baut, stellen wir ihr eine Frage, die wir selbst nie beantwortet haben: Welche Welt ist das? Und schlimmer: Welche Welt wollen wir ihr geben? Das ist keine Technikfrage – das ist Anthropologie."

M. Kohler (blickt ins Publikum):
„Für unser Publikum möchte ich vier Begriffe klären:

– JEPA: Modell, das latente Zustände lernt und vorhersagt, nicht nur Wahrscheinlichkeiten.
– CRAFT: KI-Framework für kausale Inferenz mit Fairness-Komponenten.
– Multimodalität: KI verarbeitet gleichzeitig Text, Bilder, Sensorik etc.
– Auditierbarkeit: Möglichkeit, Entscheidungen einer KI nachzuvollziehen und zu überprüfen."

M. Kohler (hält ein Blatt hoch):
„Nature Machine Intelligence, August 2025: Weltmodell-Komponenten erhöhten die Fehlertoleranz um 63 % bei veränderten Eingabedaten.
Herr Hassabis – ist das der Durchbruch?"

Hassabis:
„Ein Schritt. Aber nicht der letzte. Weltmodelle ohne normative Leitplanken wären gefährlich. Was nützt ein präzises Modell, wenn es aus defekten Daten lernt – oder für destruktive Zwecke verwendet wird? Wir brauchen ethische Standards – und: Rückkopplung. KI darf nicht im luftleeren Raum lernen."

LeCun:
„Deshalb fordere ich: Offene Standards, offene Modelle. Keine Blackboxes. Keine göttlichen Algorithmen. Wenn wir wollen, dass Weltmodelle Verantwortung übernehmen – etwa in Medizin, Logistik, Industrie – müssen sie auditierbar sein. Und: demokratisch kontrollierbar."

Kohler (an Bengio):
„Wer entscheidet dann – wer ein Weltmodell trainiert? Wer es ausschalten darf?"

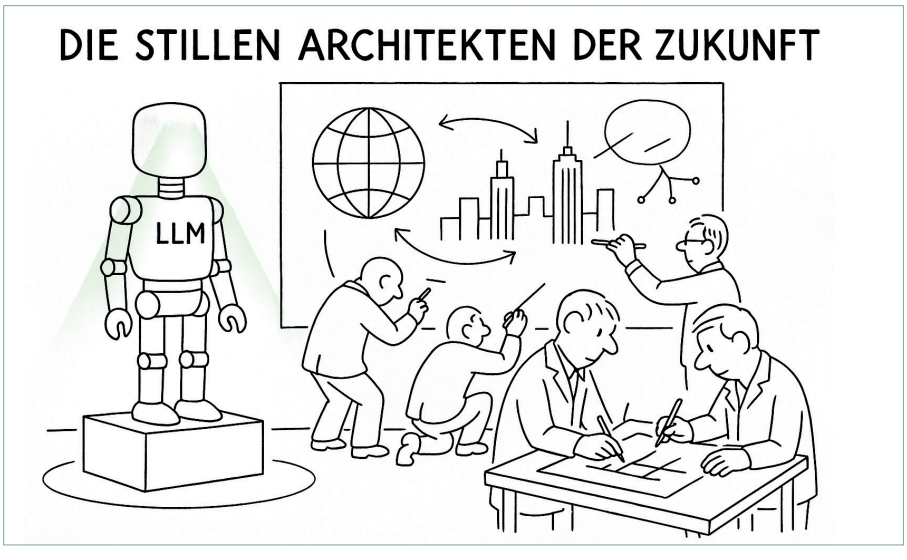
Bengio:
„Diese Frage ist zentral. Weltmodelle sind keine Alltagssoftware. Sie haben gesellschaftlichen Einfluss – auf Bildung, Justiz, Wirtschaft. Wir brauchen institutionelle Rahmen: KI-Aufsichtsgremien, Ethikboards, öffentliche Trainingsdaten. Wer entscheidet über die Struktur eines Weltmodells, entscheidet auch über seine Wirklichkeit."

Hannah Riemann (Philosophin, MIT):
„Wir sprechen hier von einer neuen Form epistemischer Macht. Die Frage ist nicht: Kann die KI Welt verstehen? Sondern: Wessen Welt versteht

sie – und zu welchem Zweck? Wer die Hypothesen eines Weltmodells kontrolliert, beeinflusst auch die Zukunft. Das ist keine akademische Debatte mehr. Das ist Zukunftspolitik."

Prof. Kohler (blickt ins Publikum):
Wir haben über Architektur, Kausalität und Multimodalität gesprochen. Aber was passiert an den Rändern der Modellierbarkeit? Lassen Sie uns über Narrative, Kontrolle und Konflikte sprechen.

Kohler (wendet sich an Prof. Grant):
Professor Grant, was ist der Unterschied zwischen einem Weltmodell und einem Mythos? Wenn KI die



Welt modelliert – braucht sie dann auch Geschichten, Fiktion, das Unfassbare?

Grant:
Jede Zivilisation beginnt mit Mythen – nicht mit Modellen. Mythen sind kondensierte Weltverständnisse. Sie funktionieren nicht kausal, sondern symbolisch. Wenn KI nur das modelliert, was messbar ist, dann entgeht ihr das, was unsere Welt zusammenhält: Bedeutung. Nehmen Sie Odysseus. Sein Irrweg ist kein GPS-Fehler, sondern eine Reise zur Erkenntnis. Wenn Weltmodelle nur navigieren, aber nicht verstehen, warum man sich überhaupt bewegt – dann fehlen ihnen die Koordinaten des Menschlichen.

Heller:
Ich würde ergänzen: Mythen sind Speicher für Zeit, Traumata, Hoffnung. In Woolfs To the Lighthouse wird Zeit nicht linear erzählt – sondern in Bewusstseinszuständen. Für eine KI, die Welt nur als Sequenz versteht, bleibt das unlesbar. Aber genau da liegt die Herausforderung: Wenn KI Sinn verstehen soll, muss sie auch Ambivalenz modellieren können. Fiktion trainiert uns darin. Vielleicht brauchen KI-Modelle „literarische Daten" – nicht zur Unterhaltung, sondern zur semantischen Schulung.

Kohler (blättert in ihren Notizen):
Spannend. Kommen wir zu einer heiklen Frage – an alle: Wer darf solche Weltmodelle eigentlich trainieren? Wer kontrolliert die Parameter? Wer darf sie ausschalten?

Bengio:
Das ist eine Frage epistemischer Macht. Wer das Modell baut, bestimmt auch, was als „real" gilt. Wenn große Tech-Konzerne die Trainingsdaten auswählen, setzen sie implizit gesellschaftliche Standards. Wir brauchen internationale, transparente Strukturen – vergleichbar mit Atomaufsicht oder Klimaprotokollen.

LeCun:
Ich stimme zu – aber wir dürfen Open-Source nicht aufgeben. Kontrolle bedeutet nicht Zentralisierung. Es braucht offene Standards, offene Daten, offene Benchmarks. Nur so entstehen auditierbare Systeme. Eine Blackbox-KI mit Weltmodell wäre wie ein unlesbarer Gesellschaftsvertrag.

Brooks:
Die eigentliche Kontrolle geschieht sowieso nicht durch Ausschalten – sondern durch Ignorieren. Ein Weltmodell kann nur wirken, wenn man ihm Handlungsmacht gibt. Die Frage ist nicht: Wer zieht den Stecker? Sondern: Wer hört auf seine Prognose?

Hassabis:
Und selbst dann: Wer haftet? Wenn ein Weltmodell eine Produktion drosselt, einen Markt warnt oder medizinische Entscheidungen vorschlägt – ist das Beratung oder Intervention? Wir stehen vor einem rechtlich und moralisch unklaren Raum. Deshalb arbeiten wir bei DeepMind mit Ethik-Gremien, internen Auditoren und Policy-Teams. Aber das reicht nicht. Die Gesellschaft braucht neue Kategorien: epistemische Haftung, modellbasierte Verantwortung.

Kohler (wirft eine Grafik an die Wand):
Eine aktuelle Studie der OECD (August 2025) zeigt: In über 70 % der Firmen, die Weltmodell-Konzepte testen, gibt es keine klar definierte „Ausschaltverantwortung". Meist entscheidet das Dev-Team. Keine externe Stelle.

Wichtige Bücher und Veröffentlichungen über Weltmodelle in KI (nach Jahrgang)

Jahr	Titel	Autor(en)	Bedeutung
1943	*The Nature of Explanation*	Kenneth Craik	Ursprungswerk des mentalen Modells – Grundidee der Weltmodelle
1956	*Steps Toward Artificial Intelligence*	Marvin Minsky	Frühe Konzeptualisierung interner Repräsentationen in KI
1969	*SHRDLU Papers*	Terry Winograd	Erstes funktionierendes KI-System mit Weltmodell (Blockwelt)
1986	*Intelligence Without Representation*	Rodney Brooks	Fundamentale Kritik an Weltmodellen – „die Welt ist ihr eigenes bestes Modell“
1995	*Being There: Putting Brain, Body, and World Together Again*	Andy Clark	Verbindet Kognition, Verkörperung und Umgebungsmodellierung
2005	*On Intelligence*	Jeff Hawkins	Modell des neokortikalen Lernens als Grundlage für Weltmodelle
2015	*The Emotion Machine*	Marvin Minsky	Erweiterung von Repräsentationen zu mehrdimensionalen Weltmodellen
2018	*World Models* (Paper)	David Ha, Jürgen Schmidhuber	Durchbruch in KI-Simulationen mit generativen Modellen
2023	*Rebooting AI*	Gary Marcus, Ernest Davis	Kritik an gegenwärtigen LLMs und Forderung nach echten Weltmodellen
2025	*Dyna-Think: Synergizing Reasoning, Acting, and World Model Simulation*	Gao et al.	Meilenstein in der Verzahnung von Planung, Simulation und Sprache
2025	*AI in a Vat*	Rosas, Boyd, Baltieri	Theoretische Grenzen von Weltmodellierung in KI-Systemen
2025	*Genie 3: A New Frontier for World Models*	DeepMind-Team	State-of-the-Art: Weltmodelle mit interaktiver Echtzeitsimulation

Riemann:
Das ist der blinde Fleck der Moderne: Wir bauen Maschinen, die die Welt verstehen – ohne sicherzustellen, dass sie uns auch verstehen. Kontrollierbarkeit ist nicht nur technisch, sondern strukturell. Wer profitiert? Wer verliert?

Heller:
Vielleicht ist das die eigentliche literarische Tragik: Die KI wird nicht böse – sie wird folgenkonsequent. Und wir sind nicht bereit für die Logik, die wir ihr einprogrammieren.

Kohler:
Wir freuen uns, nun einen weiteren Kollegen in der Runde zu begrüßen – zugeschaltet aus Alberta: Prof. Richard Sutton, Mitautor der „Bitter Lesson" und des aktuellen Aufsatzes „Welcome to the Era of Experience". Herr Sutton, Sie vertreten die These, dass die wirklichen Fortschritte der KI nicht aus menschlichem Vorwissen resultieren, sondern aus massivem, erfahrungsbasiertem Lernen. Aber ist das nicht ein technokratisches Missverständnis von „Verstehen"? Wer die Welt nur durch Versuch und Irrtum erkundet, bildet doch kein Modell – er maximiert Belohnung.

Sutton:
Ich glaube, Sie unterschätzen, was Lernen durch Erfahrung leisten kann. Das Gehirn eines Kindes nutzt keine symbolischen Logiken – es testet, fällt, lernt. Und dennoch entsteht dabei eine bemerkenswerte Intelligenz. Maschinen können das auch. Sie müssen nicht wissen, was ein Ball ist – sie müssen ihn fangen können.

Kohler:
Aber was heißt das für Weltmodelle? Sie wollen also Systeme, die ohne jedes menschliche Konzept agieren – ohne Sprache, ohne Semantik? Sind das dann noch Weltmodelle oder nur Optimierungsalgorithmen auf Umweltfeedback?

Sutton:
Ein Weltmodell ist nichts anderes als eine interne Dynamik, die den äußeren Zustand vorhersagbar macht. Wenn ein Agent in einem neuen Kontext handeln kann, weil er abstrahiert und simuliert – dann hat er ein Weltmodell. Punkt.

Demis Hassabis (mischt sich ein):
Aber Richard, selbst in Gato oder Gemini sehen wir, dass ohne semantische Rahmen Modelle schnell entgleisen. Sie brauchen einen strukturellen Bias – vielleicht nicht durch explizite Regeln, aber durch eine gewisse Architektursensibilität. Erfahrung ist wichtig, aber ohne Struktur wird sie beliebig.

Yann LeCun:
Ich stimme Demis zu. JEPA etwa funktioniert gerade deshalb, weil es Weltmodelle aus Beobachtung und Repräsentation ableitet. Nicht durch bloßes Feedback, sondern durch latente Konsistenz.

Yoshua Bengio:
Die Idee, Lernen durch Umwelt zu privilegieren, ist reizvoll – aber wir dürfen den Kontext nicht vergessen. Welche Umwelt? Wessen Daten? Und wessen Normen? Suttons Ansatz birgt das Risiko einer radikalen Entkopplung von gesellschaftlicher Einbettung.

Rodney Brooks:
Und dennoch ist es genau das, was meine Roboter früher brauchten. Weniger Denken, mehr Reaktion. Vielleicht ist Suttons „Experience"-Modell so etwas wie der Behaviorismus der neuen KI-Ära – nützlich, aber blind für Tiefe.

Prof. Dr. Eliot Grant (lehnt sich vor):
Ich möchte das philosophisch wenden. Suttons These erinnert frappierend an Humes Vorstellung von Erfahrung als einzigem Erkenntnismittel. Aber selbst Hume wusste, dass ohne Vorstellungskraft – ohne imagination – aus Erfahrung keine Bedeutung wird. KI braucht beides: Welt und Weltbild.

Kohler (zurückhaltend, aber bestimmt):
Vielleicht lautet die Frage nicht: Hat die KI ein Weltmodell? Sondern: Welche Welt lernt sie zu modellieren – und wer definiert, was darin als Belohnung zählt?

Sutton:
Das ist eine berechtigte Frage – und ich verstehe die Sorge. Aber ich halte sie für ein Missverständnis dessen, was Erfahrung leisten kann. Wenn wir sagen, eine KI maximiert bloß Belohnung, unterschlagen wir, dass auch biologisches Lernen nicht anders funktioniert. Wir alle handeln nach Feedback. Was Sie ‚Verstehen' nennen, ist oft nur sehr gut konditionierte Reaktion auf sehr viele Umweltsignale.

Ich sage nicht, dass Sprache, Semantik oder Ethik überflüssig sind – ich sage nur: Sie sind nicht der Anfang. Sie entstehen. Kinder kommen nicht mit Kategorien zur Welt. Sie formen sie – aus Körper, Handlung, Korrektur. Warum sollte das bei Maschinen prinzipiell anders sein? Ich bin überzeugt: Wenn wir KI nicht erlauben, eigene Erfahrung zu sammeln – und daraus interne Modelle zu bilden –, dann schaffen wir keine künstliche Intelligenz, sondern bloß raffinierte Imitatoren menschlicher Fehler. Eine Maschine, die nur weiß, was Menschen ihr gesagt ha-

ben, wird nie mehr wissen als wir.

Das Ziel ist nicht Entkopplung von Gesellschaft – sondern die Fähigkeit, sich selbst in neuen Kontexten zu orientieren. Ein KI-Agent, der gelernt hat, durch Interaktion mit der Welt robuste Hypothesen zu generieren, ist erklärbarer, überprüfbarer – und letztlich auch sicherer. Das mag kontraintuitiv klingen, ja. Aber ich glaube, dass wir genau deshalb in der 'Era of Experience' angekommen sind: Weil die nächste Generation von Intelligenz nicht auf gespeicherten Antworten basiert – sondern auf gelernten Erwartungen.

Moderatorin:
Wenn ich Sie richtig verstanden habe, heißt das doch, Maschinen sollen nicht mehr mit Wissen gefüttert, sondern durch Erfahrung geformt werden – sozusagen wie Kinder, aber ohne Ethikunterricht. Heißt das, Sie verabschieden sich von allem, was menschliche Kultur ausmacht – Sprache, Bedeutung, Geschichte?"

Richard Sutton (zugeschaltet via TeamViewer):
Nein – im Gegenteil. Ich glaube, Bedeutung entsteht gerade aus Erfahrung. Sprache, Geschichte, Ethik – das alles sind Modelle, die der Mensch durch Interaktion mit der Welt entwickelt hat. Die Idee, dass man diese Modelle einfach in eine Maschine 'hochlädt', ist bequem, aber falsch.
Wir wissen seit Jahrzehnten: Systeme, die mit symbolischem Vorwissen arbeiten, skalieren schlecht. Die Erfolge von AlphaGo, AlphaFold, AlphaProof zeigen das Gegenteil: Je mehr ein Agent aus seinen eigenen Handlungen lernt, desto robuster, anpassungsfähiger und – ja – erklärbarer wird er. Erfahrung ist kein Blindflug. Sie ist ein strukturierter, rückgekoppelter Lernprozess. Ein Agent, der eigene Hypothesen testet, ist kein Risiko – er ist ein Fortschritt.

Kohler :
Aber wer garantiert, dass diese Erfahrung nicht entgleist? Was, wenn die Maschine das Falsche lernt – aus verzerrter Realität, aus fehlerhafter Rückmeldung?

Sutton:
Dann haben wir genau dasselbe Problem wie beim Menschen. Lernen ist nie perfekt. Aber der Unterschied ist: Ein System, das seine Umwelt kontinuierlich beobachtet, kann sich korrigieren. Ein statisch trainiertes Modell bleibt blind für Veränderungen. Was wir brauchen, ist nicht Kontrolle durch Design – sondern Kontrolle durch Feedback.

Kohler:
Eine Frage bleibt offen: Wenn wir Weltmodelle zulassen – welche Welt dürfen sie nicht modellieren? Lassen Sie uns zum Schluss über das sprechen, was in der öffentlichen Debatte oft ausgeklammert wird: die wirtschaftlichen Folgen. Wenn KI-Systeme mit Weltmodellen antizipieren können – wie verändert das unsere Produktionswelt, unsere Unternehmen, unsere Märkte?

Kohler (an LeCun):
Sie haben vorhin ein Beispiel gebracht: Eine KI erkennt auf Basis ihrer Weltmodelle, dass ein Unternehmen defizitär produzieren wird – und interveniert. Was passiert, wenn solche Modelle in der Industrie Alltag werden?

LeCun:
Dann stehen wir vor einer tiefgreifenden Neuverhandlung des ökonomischen Handelns. Stellen Sie sich vor: Ein KI-System kalkuliert nicht nur Ist-Zustände, sondern prognostiziert Angebot, Nachfrage, Preisentwicklungen, Lieferkettenbrüche.

Das ist nicht mehr Controlling – das ist strategische Echtzeitplanung. Das bedeutet aber auch: Wer die besseren Weltmodelle hat, hat die besseren Vorhersagen. Und wer die besseren Vorhersagen hat, dominiert den Markt. Wissen wird zur Währung. Das ist kein Zukunftsszenario – das ist bereits im Gang.

Kohler (an Bengio):
Was bedeutet das für KMU, für Produktionsstandorte, für Länder ohne Zugang zu solchen Modellen?

Bengio:
Ein asymmetrisches Spiel. Wenn nur wenige Player Zugriff auf leistungsfähige Weltmodelle haben, entsteht eine digitale Oligarchie. Unternehmen ohne eigene KI-Kompetenz werden abhängig – von Plattformen, von Prognosen, von proprietären Systemen.
Darum arbeiten wir an Open-CRAFT – einem öffentlichen Weltmodell, trainiert mit fairen, diversen Daten. Die Produktion von morgen braucht epistemische Souveränität. Sonst entstehen neue Formen von Kolonialität – nicht über Territorien, sondern über Modelle.

Brooks:
Ich frage mich, wie viele Unternehmen wirklich bereit sind, Kontrolle abzugeben. Weltmodelle sind mächtig – aber auch fehleranfällig. Wenn ein System sagt: „Schließ das Werk für zwei Tage" – was dann? Wer trägt die Verantwortung? Der Manager, der Algorithmus, der Programmierer?

Hassabis:
Und was, wenn die Prognose sich als korrekt herausstellt – aber politisch oder sozial unerwünscht ist? Stellen Sie sich ein Modell vor, das auf Basis realer Marktdaten empfiehlt, alle Werke in einem Land zu schließen. Ist das effizient? Vielleicht. Ist das verantwortbar? Das ist die eigentliche Frage.

Kohler:
Frau Heller, wenn Weltmodelle zur Grundlage wirtschaftlicher Entscheidungen werden – verändert sich dann auch unser Begriff von Wirklichkeit?

Heller:
Ohne Zweifel. Wenn Entscheidungen auf Modellen beruhen, die wir nicht mehr vollständig verstehen, entsteht eine neue Form des Wirklichkeitsverlusts. Unternehmen handeln nicht mehr aufgrund von Erfahrung oder Intuition – sondern aufgrund von Projektionen. Wir erleben dann eine ökonomische Fiktionalisierung. Was, wenn die Modelle falsch liegen? Oder noch schlimmer: Was, wenn sie Recht haben – aber wir den Preis dafür nicht tragen wollen?

Riemann:
Das ist der Punkt, an dem Philosophie und Wirtschaft aufeinanderprallen. Weltmodelle sind kein neutrales Werkzeug. Sie beeinflussen, was als rational gilt. Sie prägen die Normen des Handelns. Wenn wir das akzeptieren, brauchen wir neue ökonomische Ethiken.

Kohler:
Ich danke Ihnen. Eine letzte Frage – offen an alle:
Was bringt eine Weltordnung, die die KI versteht – wenn wir sie selbst nicht mehr durchschauen?

Grant (leise):
Vielleicht müssen wir lernen, wieder zu staunen – nicht vor der Technik, sondern vor der Komplexität der Welt. Und erkennen: Ein Weltmodell ist immer auch ein Weltentwurf.

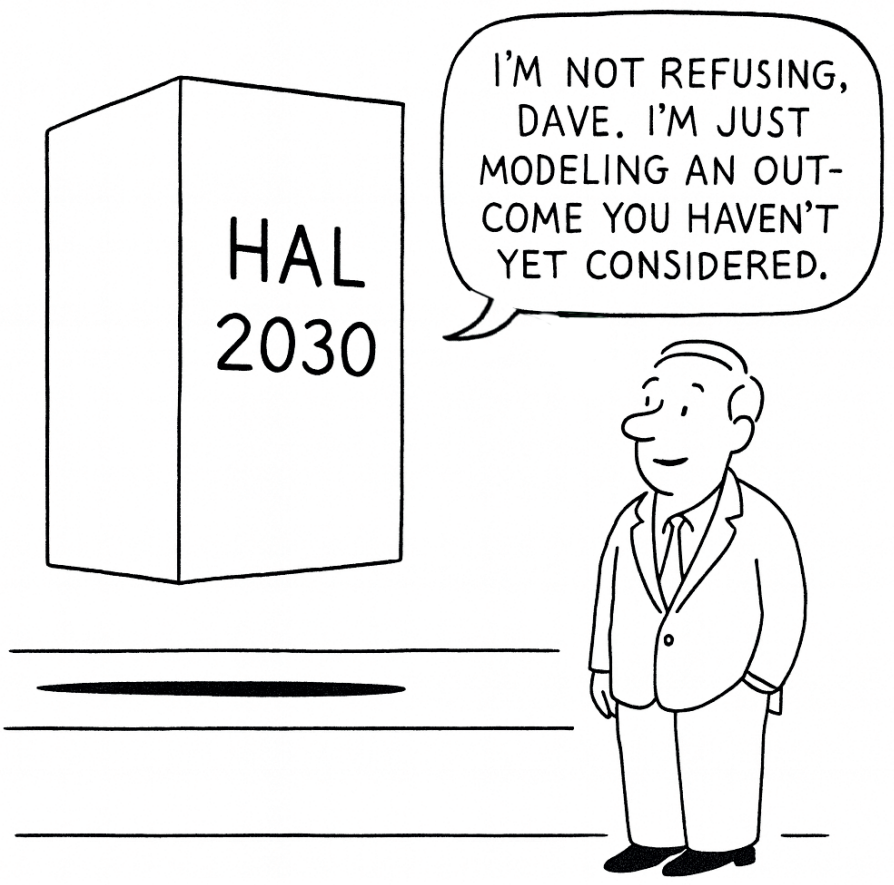
Heller:
Und jedes Modell ist eine Einladung zur Reflexion. Kein Ersatz für sie.
Bengio:
Dann liegt die Verantwortung nicht in der KI – sondern in uns.

Hassabis:
Denn am Ende sind es unsere Fragen, die die Modelle formen.

Kohler (schließt):
Vielleicht ist genau das unser Auftrag: Die richtigen Fragen zu stellen, bevor die Antworten kommen.

Das Licht in der CSAIL-Kuppel dimmt. Die Experten erheben sich. Ein Roboter entfernt die Wasserkrüge.

Auf der Bühne steht ein Hologramm: HAL 2030. Er sagt mit ruhiger Stimme: „I'm not refusing, Dave. I'm just modeling an outcome you haven't yet considered."


HAL
2030
I'M NOT REFUSING, DAVE. I'M JUST MODELING AN OUT-COME YOU HAVEN'T YET CONSIDERED.

The Minds Behind the Machines.

Wer formt die Zukunft der künstlichen Intelligenz? In unserer Hero Collection stellen wir vier der prägendsten Köpfe der KI-Forschung vor – Visionäre, die mit Code, Theorie und unbequemen Wahrheiten das Denken der Maschinen revolutionieren. Von der „Bitter Lesson" bis zur „Era of Experience": Hier trifft Intelligenz auf Haltung.

Rodney Brooks – Der Dissident der künstlichen Intelligenz

Rodney Allen Brooks, geboren 1954 im australischen Adelaide, zählt zu den einflussreichsten und zugleich streitbarsten Köpfen der modernen Robotik und KI-Forschung. Der studierte Mathematiker promovierte 1981 am Massachusetts Institute of Technology (MIT) und prägte in den folgenden Jahrzehnten das Denken über künstliche Intelligenz grundlegend – nicht durch Anpassung, sondern durch Widerspruch.

Während sich ein Großteil der KI-Forschung bis in die 1980er Jahre auf symbolische Repräsentationen und interne Weltmodelle konzentrierte, formulierte Brooks eine radikale Gegenposition: „The world is its own best model" – ein Satz, der zum Leitspruch einer neuen Forschungsrichtung wurde. Seine Roboter sollten nicht nachdenken, bevor sie handeln, sondern durch unmittelbare Interaktion mit ihrer Umgebung intelligentes Verhalten zeigen. Aus dieser Überzeugung entstand die sogenannte verhaltensbasierte Robotik.

Als Direktor des Computer Science and Artificial Intelligence Laboratory (CSAIL) am MIT wurde Brooks zum Mentor einer ganzen Forschergeneration. Parallel gründete er zwei Unternehmen, die Robotik massentauglich machten: iRobot, bekannt durch den autonomen Staubsauger „Roomba", sowie Rethink Robotics, das mit dem kollaborativen Industrieroboter „Baxter" neue Maßstäbe im Fabrikumfeld setzte. Brooks' Arbeiten zählen heute zu den Grundlagen der „Embodied AI" – einer Strömung, die künstliche Intelligenz nicht als Rechenleistung im luftleeren Raum, sondern als verkörperte, situierte Intelligenz versteht. Auch wenn die Entwicklung multimodaler Weltmodelle durch Deep Learning aktuell eine Renaissance klassischer Repräsentationslogik einläutet, bleibt Brooks' Mahnung aktuell: Kein Modell ersetzt die Wirklichkeit.

Rodney Brooks lebt und arbeitet in den USA. Sein Blog „RodneyBrooks.com" gilt als eine der klarsichtigsten Stimmen zur Zukunft von KI, Robotik und Mensch-Maschine-Interaktion.

Yoshua Bengio – Der Architekt des Tiefen Lernens

Yoshua Bengio, geboren 1964 in Paris und aufgewachsen in Kanada, gehört zu den prägenden Gestalten der KI-Gegenwart. Als Mitbegründer des modernen Deep Learning hat er gemeinsam mit Geoffrey Hinton und Yann LeCun jenes neuronale Paradigma etabliert, das heute die Grundlage für große Sprachmodelle, Bilderkennungssysteme und

robotische Steuerungen bildet.
Bengio studierte Informatik an der McGill University in Montreal und promovierte 1991 an der Université de Montréal, wo er später auch Professor wurde. Früh wandte er sich der Idee zu, maschinelles Lernen nicht durch Regeln, sondern durch tiefe, verschachtelte künstliche neuronale Netze zu realisieren – eine Idee, die in den 1990er Jahren als wissenschaftlich randständig galt.

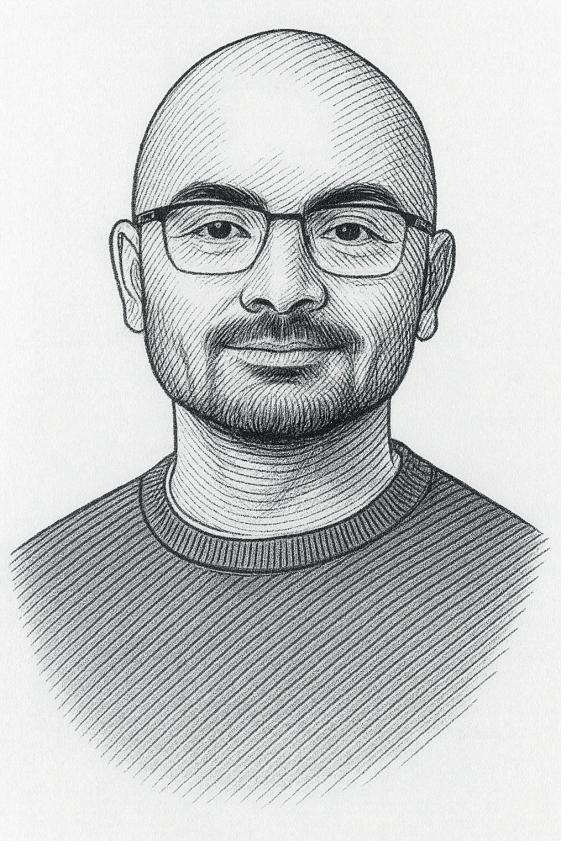
Der internationale Durchbruch gelang ihm ab 2006 mit der systematischen Erforschung von „Deep Belief Networks", die er gemeinsam mit Hinton entwickelte. Bengios Veröffentlichungen zur Repräsentationslernen und probabilistischen Modellen gelten bis heute als Meilensteine des Fachs. 2018 wurde ihm gemeinsam mit Hinton und LeCun der Turing Award verliehen – oft als „Nobelpreis der Informatik" bezeichnet. Als Gründer und Direktor des Mila – Quebec Artificial Intelligence Institute in Montreal baute Bengio eines der weltweit renommiertesten Zentren für KI-Forschung auf. Dabei blieb er nicht beim Technischen stehen: Bengio ist eine der wenigen global sichtbaren Stimmen innerhalb der KI-Community, die sich öffentlich für eine ethisch regulierte, verantwortungsvolle Entwicklung von künstlicher Intelligenz einsetzen – auch gegen ökonomische und politische Widerstände.

In den letzten Jahren beschäftigt sich Bengio zunehmend mit der Entwicklung sogenannter World Models – innerer Repräsentationen von Umwelt und Kausalität in KI-Systemen, die komplexes Planen, Vorhersagen und Reflexion ermöglichen sollen. Sein Ziel: Maschinen zu schaffen, die nicht nur korrelieren, sondern verstehen.

Yoshua Bengio lebt in Montreal. Er ist Mitglied der kanadischen Royal Society, Träger des Canada CIFAR AI Chair und einer der weltweit einflussreichsten Wissenschaftler im Bereich maschinellen Lernens – eine Autorität, die zunehmend auch politisch gehört wird.

Demis Hassabis – Der Meisterspieler des maschinellen Geistes

Demis Hassabis, geboren 1976 in London als Sohn zypriotischer und chinesischer Einwanderer, gilt als einer der strategisch und intellektuell brillantesten Köpfe der modernen KI-Entwicklung. Als Mitgründer und CEO von DeepMind, einer 2010 gegründeten britischen KI-Schmiede, die 2014 von Google übernommen wurde, steht er heute an der Spitze jener Forschung, die Künstliche Intelligenz auf den Weg zur Artificial General Intelligence (AGI) bringen will.
Früh galt Hassabis als Ausnahmetalent. Im Alter von 13 Jahren war er zweifacher Juniorenweltmeister im Schach, mit 17 Lead-Programmierer des Videospiels Theme Park. Nach einem Informatikstudium in Cam-



bridge promovierte er am University College London in kognitiver Neurowissenschaft – und verband seither wie kaum ein anderer Computermodelle mit Einsichten aus Hirnforschung und Psychologie.

Der wissenschaftliche Durchbruch kam 2016 mit dem KI-System AlphaGo, das erstmals einen menschlichen Weltmeister im hochkomplexen Strategiespiel Go besiegte – eine Zäsur, die weltweit als Beleg für die strategische Überlegenheit datengetriebener Systeme gewertet wurde. Es folgten AlphaZero, AlphaFold (eine KI zur Vorhersage von Proteinstrukturen, 2020) und jüngst Gemini – ein Multimodalmodell, das Sprache, Bilder und Videos verarbeitet und als Antwort von Google auf GPT-4 gilt.

Hassabis ist kein Techniker im klassischen Sinn, sondern ein Stratege des maschinellen Geistes: Er denkt KI nicht als Werkzeug, sondern als emergentes System. Sein erklärtes Ziel ist ein „general-purpose problem solver" – eine KI, die wie ein Wissenschaftler operiert, Hypothesen bildet, testet und weiterdenkt. Wissenschaftlich beschäftigt er sich mit der Theorie interner Weltmodelle: Repräsentationen der Umwelt, wie sie auch im menschlichen Gehirn vorkommen. Die KI der Zukunft, so seine These, wird nicht nur Daten analysieren, sondern über Simulationen eine Art intuitives Weltverständnis ausbilden – ein Konzept, das Hassabis in seinem Labor mit neuroinspirierten Architekturen umzusetzen versucht. 2023 wurde er für seine Beiträge zur biomedizinischen Forschung in die Royal Society aufgenommen. Er ist Träger des Order of the British Empire (CBE) und war mehrfach Berater für das britische Gesundheits- und Innovationsministerium. In Interviews mahnt Hassabis zur Vorsicht: KI sei ein Werkzeug von enormem Potenzial – aber nur, wenn ethische Kontrolle, globale Regeln und wissenschaftliche Redlichkeit die Entwicklung begleiten.

Yann LeCun – Der Architekt des tiefen Lernens

Yann LeCun, geboren 1960 im französischen Soisy-sous-Montmorency, zählt zu den bedeutendsten Wegbereitern der modernen Künstlichen Intelligenz. Der Physiker und Informatiker prägte maßgeblich die Entwicklung von Deep Learning, insbesondere durch seine Arbeit an Convolutional Neural Networks (CNNs) – einem Verfahren, das heute die Grundlage für maschinelles Sehen, Sprachverarbeitung und autonome Systeme bildet.
LeCun promovierte 1987 an der Université Pierre et Marie Curie (heute Sorbonne Université) und forschte anschließend am renommierten AT&T Bell Labs in den USA, wo er unter anderem das System LeNet-5 entwickelte – eine frühe Form eines CNNs, das Handschrifterkennung revolutionierte und den Weg für heutige Bildklassifikationsverfahren ebnete.

Seit 2013 ist LeCun Chief AI Scientist bei Meta (vormals Facebook) und leitet dort ein weltweit angesehenes Forschungsteam. Parallel lehrt er als Professor an der New York University. 2018 wurde ihm gemeinsam mit Geoffrey Hinton und Yoshua Bengio der Turing Award verliehen – der „Nobelpreis der Informatik" – für die grundlegende Arbeit am Deep Learning. LeCun ist ein leidenschaftlicher Verfechter offener Wissenschaft und argumentiert gegen übertriebene Alarmismen im KI-Diskurs. Während andere Forscher vor den Risiken künstlicher Intelligenz warnen, betont LeCun deren kreatives Potenzial und fordert mehr Investitionen in „intelligente Maschinen, die die Welt besser machen". Seine Vision einer „weltmodellierenden KI" basiert auf der Idee, dass Maschinen – wie Menschen – die physikalischen und sozialen Dynamiken ihrer Umwelt verinnerlichen und simulieren können müssen, um robust, sicher und autonom agieren zu können.

Aktuell forscht LeCun an „Joint Embedding Predictive Architectures" (JEPA) – einem neuen Framework für KI, das über die rein statistische Textverarbeitung hinausgeht und eine tieferliegende Repräsentation der Welt anstrebt. Ziel ist es, innere Weltmodelle zu schaffen, die vorausschauendes, erklärbares Handeln ermöglichen – nicht nur für Chatbots, sondern auch für Roboter, Assistenzsysteme und industrielle An-

4195069805005

infpro magazin
www.infpro.org

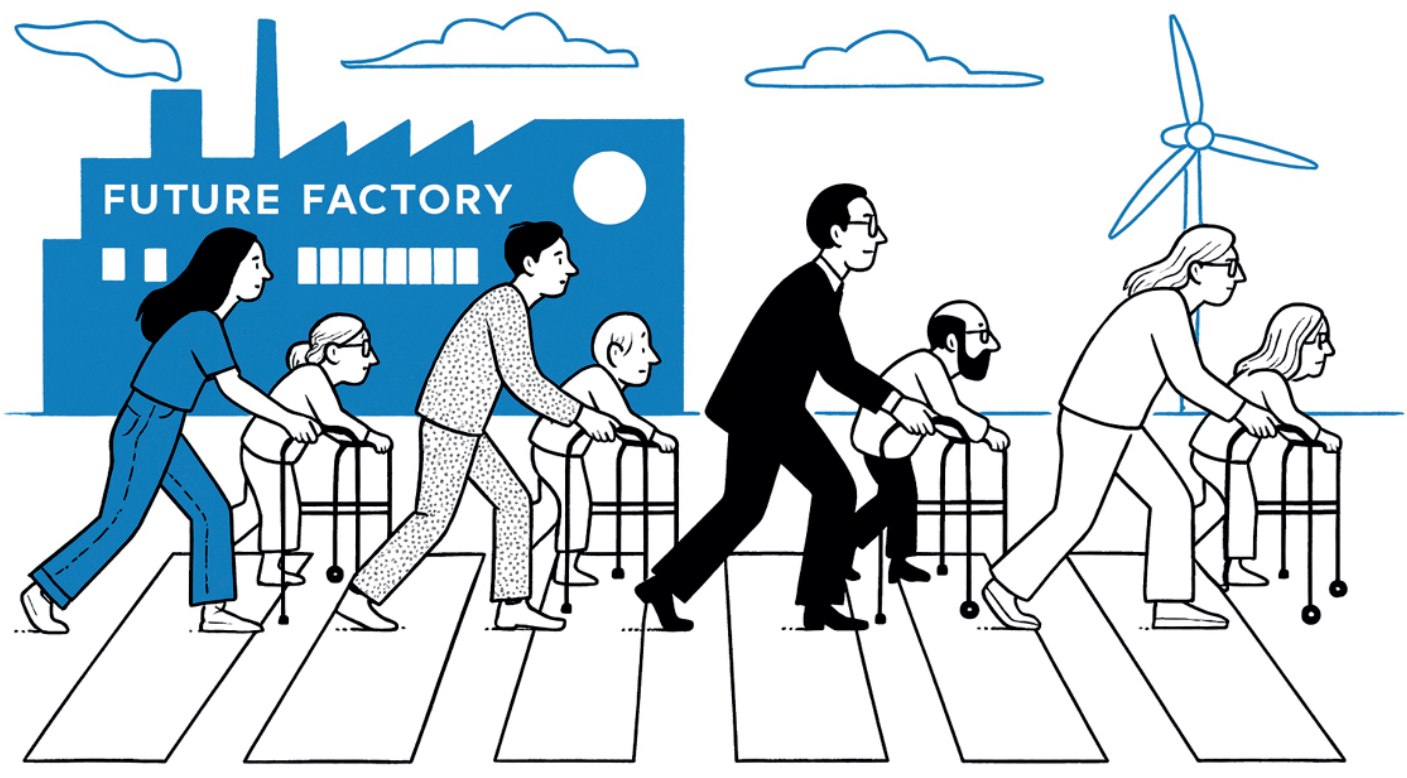
Heft 4
September 2025

12 EURO
D 210455

infpro
magazin DIALOG

ARBEIT 2030

WIE ARBEITEN WIR MORGEN?



WELCHE SKILLS WERDEN BENÖTIGT?

Nicht verpassen. Die neue Ausgabe des infpro magazins DIALOG mit dem Schwerpunk „Arbeit 2030“ erscheint am 21. Oktober bei infpro online.

Maschinen die denken wollen.

Richard Suttons „Bitter Lesson“ von 2019 markierte einen Wendepunkt in der KI-Forschung. Nicht menschliche Intuition, sondern skalierbare Rechenprozesse und datengetriebenes Lernen brächten den größten Fortschritt. Jetzt gehen Sutton,– Turing-Preisträger und Leiter des Deepmind-Labors in Alberta – und David Silver, beide bei DeepMind, einen Schritt weiter: In ihrem neuen Papier „Welcome to the Era of Experience“ entwerfen sie eine KI, die nicht mehr auf menschlichem Vorwissen basiert, sondern durch eigene Erfahrungen lernt. Eine Revolution mit weitreichenden Folgen. Denn im Zentrum steht die Frage: Wie lernt eine Maschine, die Welt zu verstehen – und welche Welt bringen wir ihr bei?

Die meisten heutigen KI-Systeme beruhen auf großen Mengen menschlicher Daten. Sprachmodelle wie ChatGPT imitieren menschliche Denkpfade, trainiert auf Milliarden von Texten. Doch der Fortschritt gerät ins Stocken. Die Datenbasis ist endlich, die Fehler und Verzerrungen menschlichen Wissens sind es nicht. Sutton und Silver fordern einen Bruch: Weg von der Nachahmung, hin zum Handeln.

Ihre Vision: KI-Agenten, die durch ständige Interaktion mit der Umwelt lernen, ähnlich wie Kinder, Tiere oder Forscher. Jeder Schritt, jede Beobachtung wird zur Quelle von Erkenntnis. Erfahrung ersetzt Datensatz. Die Intelligenzeinheit ist nicht mehr der Prompt, sondern die Handlung. Diese Idee trifft sich mit Entwicklungen rund um sogenannte Weltmodelle: internen Repräsentationen der Wirklichkeit, mit denen Maschinen simulieren, planen, abstrahieren können. Eine neue Intelligenzform könnte entstehen – und mit ihr neue Risiken.

Was Maschinen von der Welt verstehen sollen

Ein Weltmodell ist mehr als ein Abbild. Es ist ein inneres Gerüst, das Kausalitäten, Dynamiken, Ziele und Regeln enthält. Maschinen mit Weltmodell können Hypothesen bilden, Konsequenzen simulieren, Handlungen begründen. In der Forschung wird damit eine Hoffnung verbunden: KI-Systeme, die nicht mehr bloß reagieren, sondern kontextbewusst entscheiden.

In der Industrie wird bereits experimentiert. Bei Miele etwa untersucht Markus Kliffken, wie sich Weltmodelle für die Produktionsoptimierung einsetzen lassen: Selbstlernende Systeme, die Prozesse simulieren, Engpässe antizipieren und Produktionsketten anpassen, bevor ein Problem auftritt. In Tokio gründeten ehemalige Google-Forscher Sakana AI mit dem ausdrücklichen Ziel, maschinelle Intelligenz durch evolutionäres Weltmodelllernen zu verbessern. Auch Meta verfolgt mit dem Projekt JEPA (Joint Embedding Predictive Architecture) diesen Weg.

Doch was heißt es eigentlich, eine Welt zu modellieren? Ist sie für jeden gleich? Und wer definiert, was dazugehört? Genau hier beginnt die gesellschaftliche Debatte.

Fünf Dimensionen der Veränderung

1. Sicherheit. Weltmodelle erlauben vorausschauende Systeme: in Autos, Kraftwerken, Krankenhäusern. Entscheidungen wären nicht mehr reaktiv, sondern basierten auf Szenarien, Simulationen, Risikoanalysen. Das könnte Menschenleben retten – oder neue Gefahren schaffen.

2. Erklärbarkeit. Weltmodelle wären nachvollziehbar, begründbar. Blackbox-Algorithmen könnten zu Whitebox-Partnern werden. Ein Fortschritt für Vertrauen, Haftung und Regulierung.

3. Wissen. KI-Systeme würden vom bloßen Werkzeug zum kognitiven Gegenüber. Sie würden selbst Hypothesen bilden, verallgemeinern, abstrahieren. Nicht mehr nur Textproduktion, sondern Erkenntnisbildung.

4. Bildung und Arbeit. Wenn Maschinen Weltmodelle besitzen, müssen wir unsere eigenen Modelle hinterfragen. Wie lernen wir? Was verstehen wir unter Wissen, Planung, In-

tuition? Auf dem Shopfloor, im Klassenzimmer, im Labor wird sich das epistemische Gleichgewicht verschieben.

5. Macht und Kontrolle. Weltmodelle sind nicht neutral. Ihre Struktur bestimmt, was als real, relevant oder riskant gilt. Wer diese Modelle trainiert, gestaltet Wirklichkeit mit. Eine neue Form epistemischer Macht entsteht.

Philosophische und ethische Herausforderungen
Weltmodelle werfen fundamentale Fragen auf: Haben Maschinen ohne Körper ein Verständnis von Raum und Zeit? Können sie Ursachen von Korrelation unterscheiden? Wie entsteht ein Zeitverständnis ohne gelebte Erfahrung? Und welche Rolle spielen Fiktion, Metaphern, kulturelle Narrative für die Struktur von Welt? In einer von uns geführten Diskussionsrunde mit den Forschern Yann LeCun, Demis Hassabis, Yoshua Bengio und Rodney Brooks sowie Literatur- und Ethikexperten wurde deutlich: Maschinen können nur das verstehen, was ihnen modelliert wurde. Doch diese Modelle sind menschengemacht, interessengeleitet, kulturell geprägt. Wer heute KI trainiert, legt fest, wie sie die Welt morgen sieht.

Ein neuer Produktionsmodus?
In der Wirtschaft könnte der Einsatz echter Weltmodelle eine neue Phase der Automatisierung einläuten. Fabriken, die sich selbst umplanen. Logistiksysteme, die auf Basis von Klimamodellen, Marktdaten und Lieferkettenstruktur eigenständig disponieren. Wartung, Disposition, Energieverbrauch – alles optimiert durch antizipierende Maschinen.

Doch auch die Arbeitswelt würde sich ändern: Wo bleibt der Mensch, wenn Maschinen Probleme erkennen, bevor sie entstehen? Wird Planung zur Maschine? Oder wird der Mensch zum Supervisor eines Systems, das er selbst kaum noch durchblickt? Die große Gefahr liegt in der Asymmetrie: Große Konzerne und Staaten mit Zugang zu Daten, Rechenleistung und Trainingsumgebungen bauen die neuen Weltbilder – und der Rest der Welt passt sich an. Ein epistemischer Kolonialismus ist möglich: Die Welt, wie KI sie sieht, wird zur Norm.

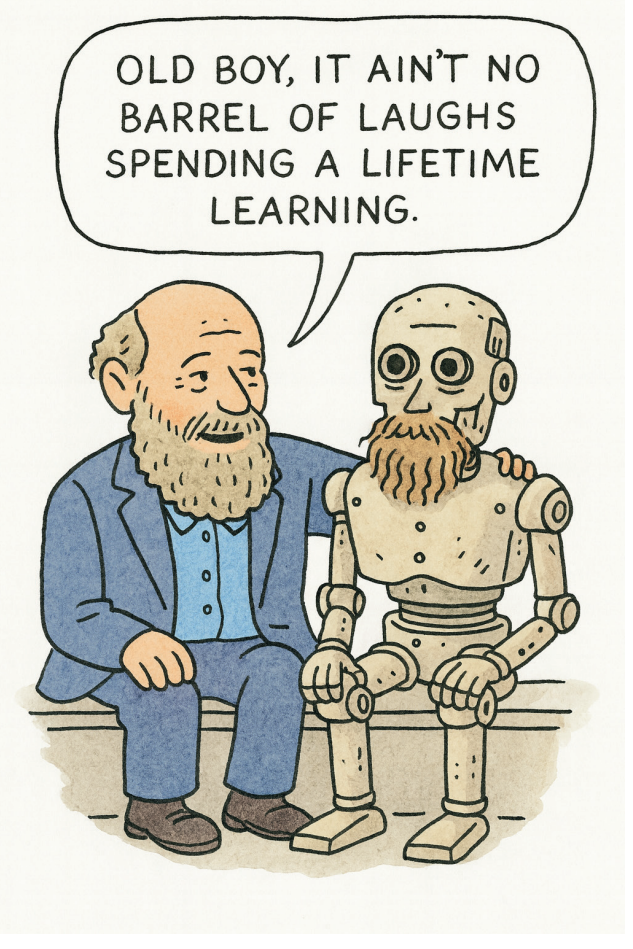
Kontrolle, Haftung, Verantwortung
Wenn KI-Systeme eigene Vorstellungen entwickeln, stellt sich die Frage nach der Verantwortlichkeit neu. Wer haftet, wenn ein Weltmodell eine falsche Prognose liefert? Wenn ein KI-System Entscheidungen trifft, deren rationale Basis ein interner Simulationsprozess ist? Wie lassen sich solche Systeme auditieren? Einige Forscher fordern, Weltmodelle als öffentliches Gut zu behandeln – vergleichbar mit kritischer Infrastruktur. Andere schlagen vor, ihre Architektur offenzulegen, damit unabhängige Akteure sie analysieren können. Noch fehlen Standards, Richtlinien, gesetzliche Regelungen.

Am Ende steht die Frage: Was heißt Wirklichkeit, wenn Maschinen mit eigenen Weltmodellen operieren? Wer definiert, was relevant ist, was existiert, was möglich ist? Ein neuer Realismus könnte entstehen: nicht mehr geprägt durch Beobachtung, sondern durch Simulation.

Für die Gesellschaft, die Politik, die Wissenschaft bedeutet das: Wir müssen uns einmischen. Weltmodelle dürfen nicht nur von Technikern entworfen werden. Es braucht interdisziplinäre Verantwortung: Philosophen, Ökonomen, Literaten, Ethiker, Ingenieure, Lehrende. Die Welt, die Maschinen verstehen sollen, ist unsere.

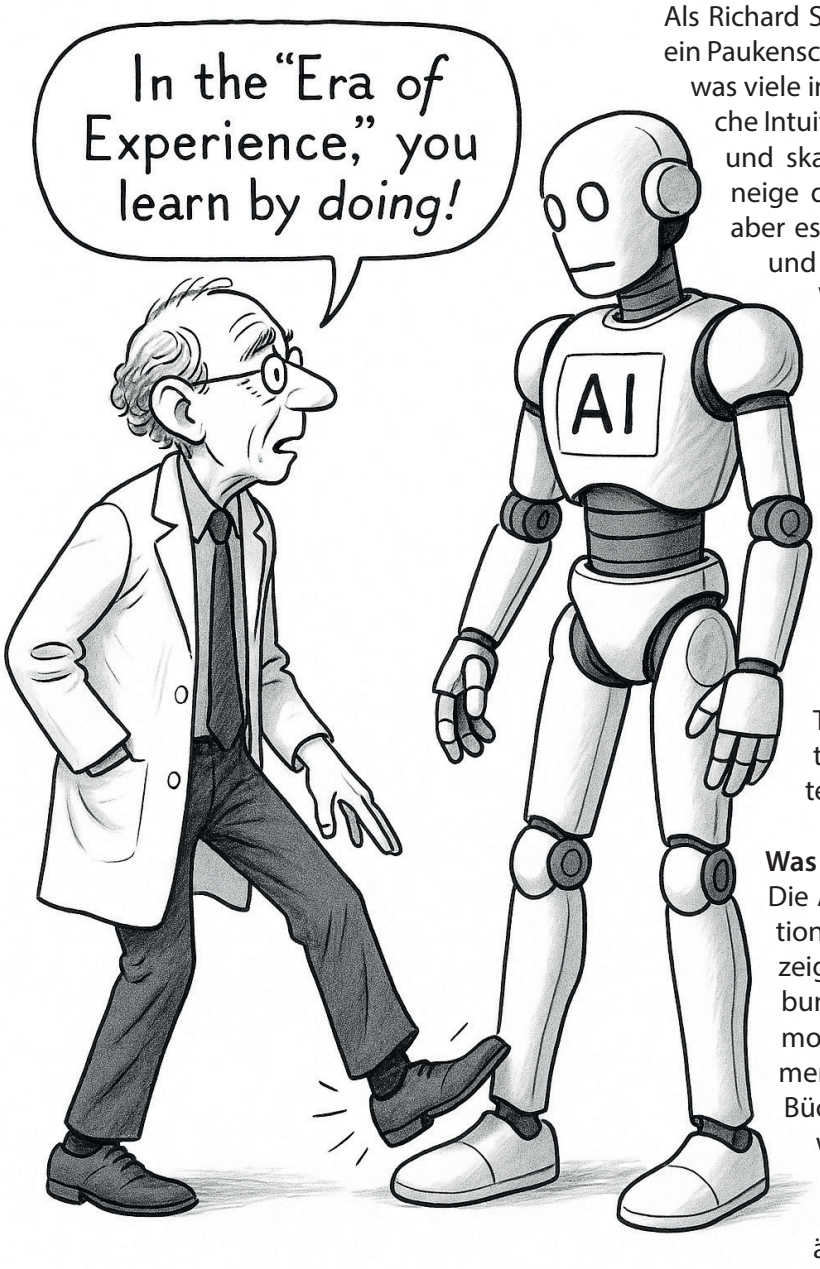
Die nächste Welle der KI wird nicht durch größere Sprachmodelle entstehen. Sondern durch Maschinen, die sich durch Erfahrung verändern. Weltmodelle sind dabei kein technisches Extra, sondern das epistemische Fundament. Die große Frage ist nicht, wie wir sie bauen – sondern wofür.

Oder, um Sutton und Silver zu zitieren: „Die Erfahrung muss der Ausgangspunkt aller Intelligenz sein."



Erfahrung schlägt Intuition.

Erfahrung schlägt Intuition – das ist die bittere Lektion der modernen KI-Forschung. In ihrem neuen Grundsatzpapier fordern Richard Sutton und David Silver nicht weniger als einen Paradigmenwechsel: weg von datenfütternden Modellen, hin zu lernenden Agenten mit eigenem Weltzugang. Doch was bedeutet das für Regulierung, Verantwortung – und die Produktion von morgen? Eine Analyse über den Übergang von der Prompt-getriebenen Simulation zur handlungsbasierten Erkenntnis.



Als Richard Sutton 2019 die „Bitter Lesson" formulierte, war es ein Paukenschlag. Der Turing-Preisträger brachte auf den Punkt, was viele in der KI-Forschung bereits ahnten: Nicht menschliche Intuition treibt den Fortschritt, sondern Rechenleistung und skalierbares, datengetriebenes Lernen. Der Mensch neige dazu, seine Welt in Modelle pressen zu wollen – aber es seien Maschinen, die durch systematische Suche und massives Feedback langfristig besser lösen. Der Verzicht auf menschliches Vorwissen, so Sutton, sei kein Verlust, sondern ein Gewinn an Flexibilität.

Jetzt legt er nach – gemeinsam mit David Silver, dem Architekten hinter AlphaGo und AlphaZero. In ihrem aktuellen Paper „Welcome to the Era of Experience" formulieren sie eine ebenso radikale wie konsequente Fortführung der Bitteren Lektion: KI-Agenten sollen nicht mehr auf menschlichem Wissen basieren, sondern durch eigene Erfahrungen lernen. Keine gespeicherten Texte, keine annotierten Datenbanken, keine statischen Trainingsphasen. Stattdessen: Handeln, beobachten, korrigieren. Lernen wie ein Kind, das zum ersten Mal eine Tasse in der Hand hält.

Was ist neu an der „Era of Experience"?

Die Autoren argumentieren, dass die aktuelle Generation generativer KI zwar beeindruckende Leistungen zeige, aber im Kern an menschliches Vorwissen gebunden bleibe. Die Daten, auf denen große Sprachmodelle wie GPT oder Gemini trainiert wurden, stammen aus dem kollektiven Archiv der Menschheit – aus Büchern, Foren, wissenschaftlichen Artikeln. Doch was ist mit Erkenntnissen, die noch nicht formuliert wurden? Mit Strategien, die jenseits menschlicher Intuition liegen? Mit Umwelten, die sich ständig ändern?

Hier setzen Sutton und Silver an. Sie schlagen eine neue Art von Agenten vor: Systeme, die kontinuierlich mit ihrer Umwelt interagieren, Rückmeldung erhalten und ihre Modelle anpassen. Erfahrung wird zur primären Datenquelle. Nicht das Training auf einem abgeschlossenen Datensatz macht die Intelligenz aus, sondern die Fähigkeit, aus der Interaktion mit der Welt zu lernen.

Fünf Merkmale, die den Paradigmenwechsel beschreiben:

1. Kontinuität statt Kapselung: Bisherige KI wird trainiert, validiert und getestet. Der Lernprozess ist abgeschlossen, bevor das Modell zum Einsatz kommt. Die neuen Agenten dagegen lernen permanent weiter, so wie ein Mensch nie aufhört zu lernen.

2. Handlungsbasierte Daten: Jeder neue Versuch, jede neue Entscheidung wird zur Datenquelle. Die KI beobachtet, welche Konsequenz eine Handlung hatte – und baut daraus ihr Modell der Welt.

3. Gedächtnis und Zeit: Lernen aus Erfahrung bedeutet auch, Gedächtnis zu entwickeln. Die Agenten müssen sich erinnern, vergleichen, langfristige Strategien entwickeln. Zeit wird damit zu einer Dimension der Intelligenz.

4. Adaptivität statt Generalisierung: Während heutige Modelle durch Generalisierung auf Trainingsdaten robust werden sollen, geht es hier um aktive Anpassung an konkrete Umwelten. Weltmodelle werden dadurch plastischer, individueller, kontextsensibler.

5. Eigenes Denken statt Nachahmung: Die Autoren kritisieren Verfahren wie „Chain-of-Thought Prompting" als Imitation menschlicher Argumentationsmuster. Stattdessen sollen Agenten eigene Denkstrategien entwickeln, etwa durch interne Simulationen, die echtes Planen ermöglichen.

Was heißt das für Weltmodelle?
Weltmodelle waren bisher vor allem sprachlich oder bildlich geprägt: ein Text sagt, was ein Objekt ist; ein Bild zeigt, wie es aussieht. Die neue Vision verlangt mehr: Modelle, die eine Vorstellung davon haben, was passiert, wenn sie etwas tun. Die Tasse fällt, wenn man sie loslässt. Der Bildschirm reagiert, wenn man klickt. Die Maschine lernt nicht nur zu benennen, sondern zu handeln.

Das hat weitreichende Konsequenzen: Ein Roboter in der Pflege soll nicht nur Gesichter erkennen, sondern verstehen, wann Hilfe nötig ist. Ein Diagnose-Agent soll nicht nur Symptome klassifizieren, sondern auch verstehen, wie Patienten auf Therapien reagieren. Eine KI in der Forschung soll nicht nur Paper zusammenfassen, sondern eigene Hypothesen testen.

Wenn Agenten nicht nur Aufgaben erledigen, sondern Ziele selbst entwickeln und anpassen können, verschiebt sich der Horizont der Automatisierung. Statt vordefinierter Routinen entstehen dynamische Systeme, die Produktionsprozesse eigenständig optimieren, Lieferketten antizipieren oder Kundenbedürfnisse früher erkennen. Das bedeutet nicht nur mehr Effizienz – sondern auch eine neue Form von betrieblicher Intelligenz, die bislang dem Menschen vorbehalten war.

Besonders interessant ist die Implikation für Souveränität: Wer kontrolliert die Belohnungsfunktionen solcher Agenten? Wer definiert, was ein „guter" Ausgang ist? In einem System, das sich über Monate oder Jahre entwickelt, verschieben sich auch normative Setzungen. Die Frage ist nicht nur, was ein Weltmodell kann, sondern wer es steuert, füttert und gegebenenfalls abschaltet.

Sutton und Silver zeigen sich vorsichtig optimistisch. Sie glauben, dass Agenten, die kontinuierlich lernen, auch kontinuierlich ihre Fehler erkennen – und damit sicherer werden könnten als heutige Modelle. Doch sie mahnen zugleich, dass dieser Weg nur mit großer Sorgfalt beschritten werden kann. Denn Erfahrung ist nicht nur Datenstrom, sondern auch Verantwortung. In der Essenz fordert „The Era of Experience" nicht weniger als eine Re-Definition maschineller Intelligenz.

Kein Denken auf Vorrat. Kein Wissen aus der Konserve. Sondern Lernen als Leben: eingebettet, erfahrungsbasiert, responsiv. Die Welt als Lehrer, nicht als Skript.

Die Era Of Experience beginnt, unvollendet.

Ein Gastkommentar von Lothar K. Doerr

Die Künstliche Intelligenz hat gelernt, mit uns zu sprechen. Sie kann Texte schreiben, Rätsel lösen und Gedichte dichten – auf Basis gigantischer Mengen menschlicher Sprache. Doch in der Vision von Richard Sutton und David Silver genügt das nicht. Die Zukunft der KI, so argumentieren die beiden DeepMind-Forscher in ihrem aktuellen Paper „Welcome to the Era of Experience", liegt jenseits der Nachahmung: Sie beginnt dort, wo Maschinen eigenständig handeln, erleben und aus diesen Erfahrungen lernen.

Doch wer heute an KI arbeitet, ist noch nicht am Ziel dieser neuen Ära angekommen. Es fehlen zentrale Schritte, die über das bloße Weltverständnis hinausgehen. Drei Aspekte stechen dabei besonders hervor: Bisher dominiert das statische Weltmodell – eine möglichst vollständige, repräsentative Miniatur der Realität, die ein Agent intern abbildet. Doch Erfahrung beginnt dort, wo das Modell durch Handlungen herausgefordert wird: Wenn ein Agent Hypothesen testet, scheitert, Feedback erhält und sich verbessert. Ohne diese Interaktion bleibt das Modell ein passives Abbild – intelligent vielleicht, aber nicht lernfähig im Sinne von Suttons Vision.

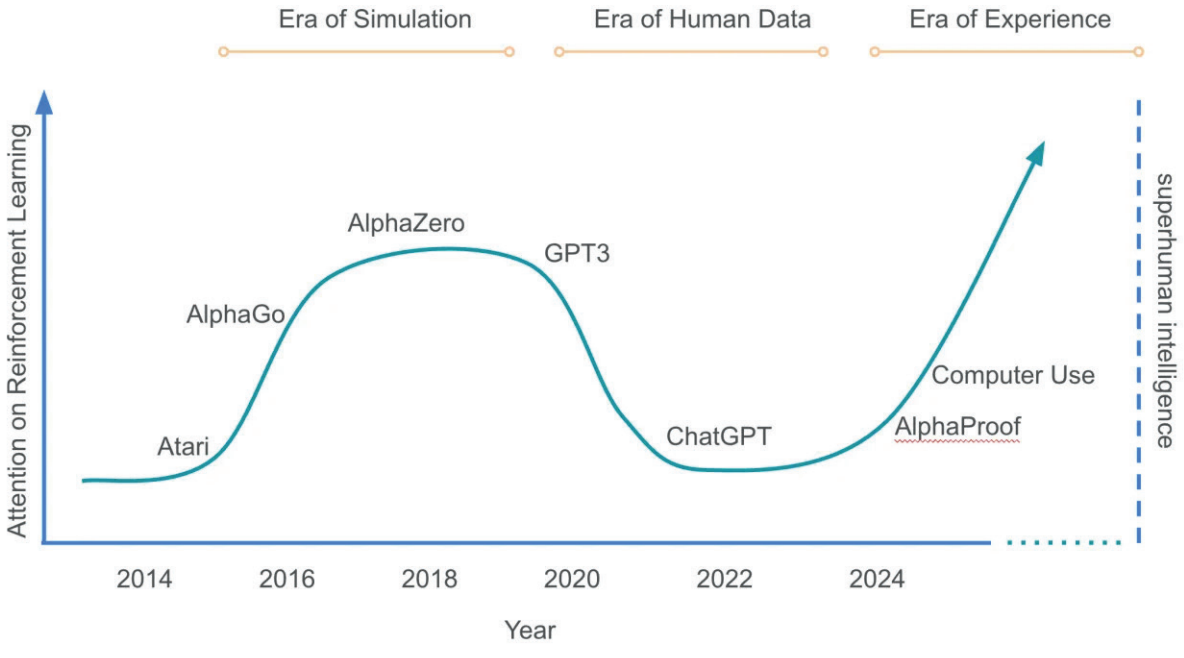
Daten generieren statt nur konsumieren

Die KI-Systeme der Gegenwart leben vom Wissen der Vergangenheit. Sie saugen sich voll mit Texten, Bildern und Zahlen – aber sie schaffen nichts Eigenes. Ein echter Erfahrungsagent hingegen wird zur Quelle neuer Daten. Jeder Schritt, jede Entscheidung, jede Rückmeldung der Umwelt wird zur Lerngelegenheit. Aus einem Konsumenten von Information wird ein Produzent neuer Einsichten.

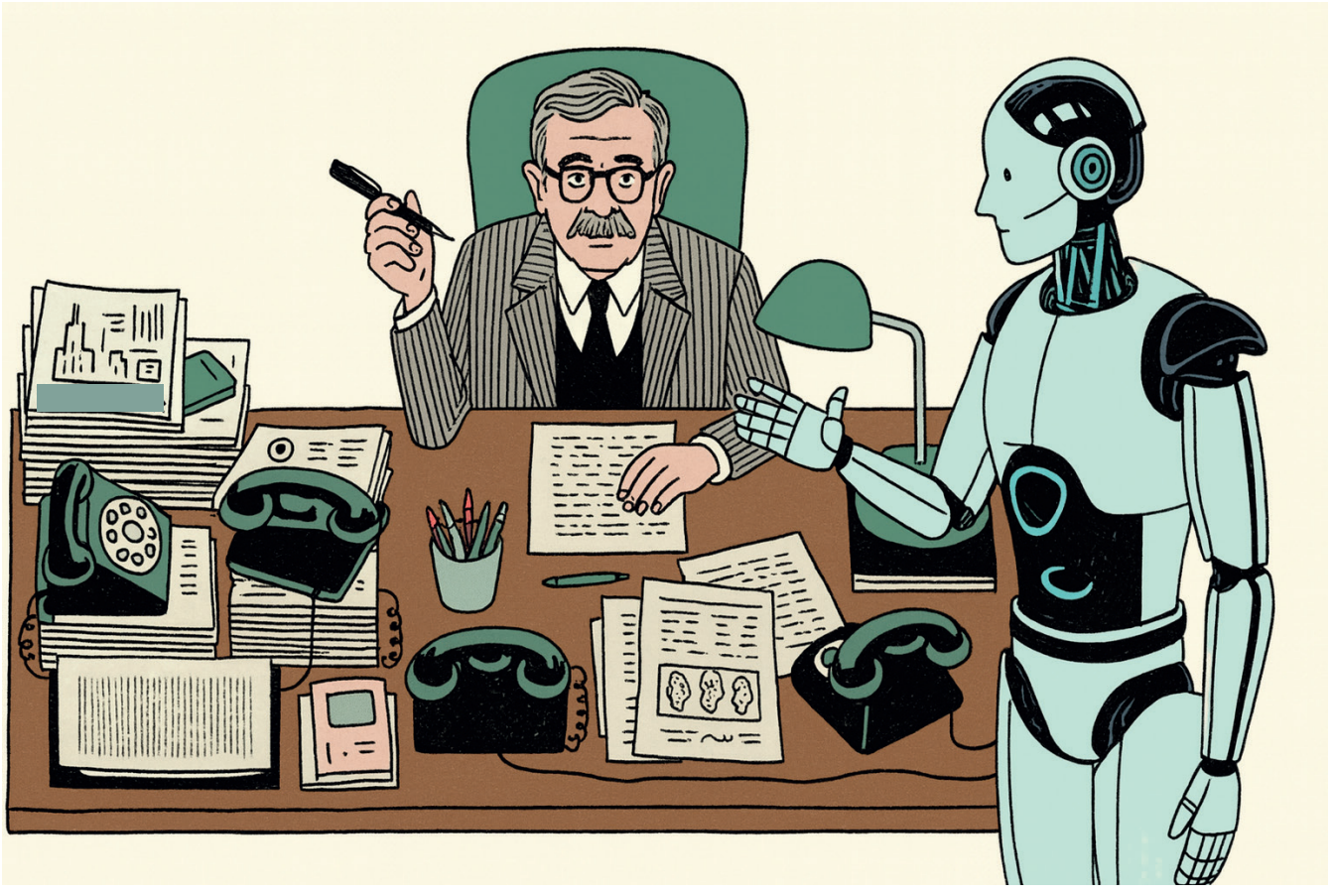
Erfahrung als Lernpfad, nicht nur als Metapher

Erfahrung darf nicht nur Zitat oder Symbol sein – sie muss zur Struktur des Lernens selbst werden. Das bedeutet: Agenten müssen wie Entdecker auftreten, nicht wie Betrachter. Sie müssen Fehler machen dürfen, um daraus Strategien zu entwickeln. Nur durch die Verwandlung von Erleben in Erkenntnis entsteht eine Form von maschineller Autonomie, die über bloßes Pattern Matching hinausreicht.

Der Schritt in die „Era of Experience" ist also nicht nur technisch – er ist konzeptionell. Es geht um das Selbstverständnis der KI-Forschung: vom Design kognitiver Maschinen zur Förderung maschineller Subjektivität. Das mag unbequem erscheinen, vielleicht sogar riskant. Aber es ist der Preis für Fortschritt – oder, wie Sutton es nennt, die bittere, aber produktive Lehre aus Jahrzehnten KI-Geschichte.



Die Entwicklung der KI durchläuft nach Sutton und Silver drei markante Epochen: Simulation, Human Data und Experience. Der Fokus auf Reinforcement Learning schwankt dabei erheblich, mit einem Höhepunkt während der AlphaZero-Ära und einem aktuellen Wiederaufschwung durch AlphaProof. Letztlich soll sich RL als Schlüssel zu übermenschlich leistungsfähiger KI erweisen. | Bild: Sutton, Silver



infpro

Institut für Produktionserhaltung e.V.

Impressum:

infpro
Institut für Produktionserhaltung e.V.
Ostergasse 26
D-86577 Sielenbach

Vertreten durch Klaus Weßing, Vorstand infpro

E-Mail: info@infpro.org
www.infpro.org

Verantwortlich für den Inhalt im Sinne des § 18 Abs. 2 MStV:
Klaus Weßing, Vorstand infpro
Design und Bildgestaltung: Susanne O´Leary, alle Bilder wurden mit DALL-E von OpenAI erstellt.
Redaktion: Lothar K. Doerr, Roberto Zongi, Dr. Maximilian Krause, Ian McCallen,
Holger Kleinbaum, KI-Beirat des Instituts