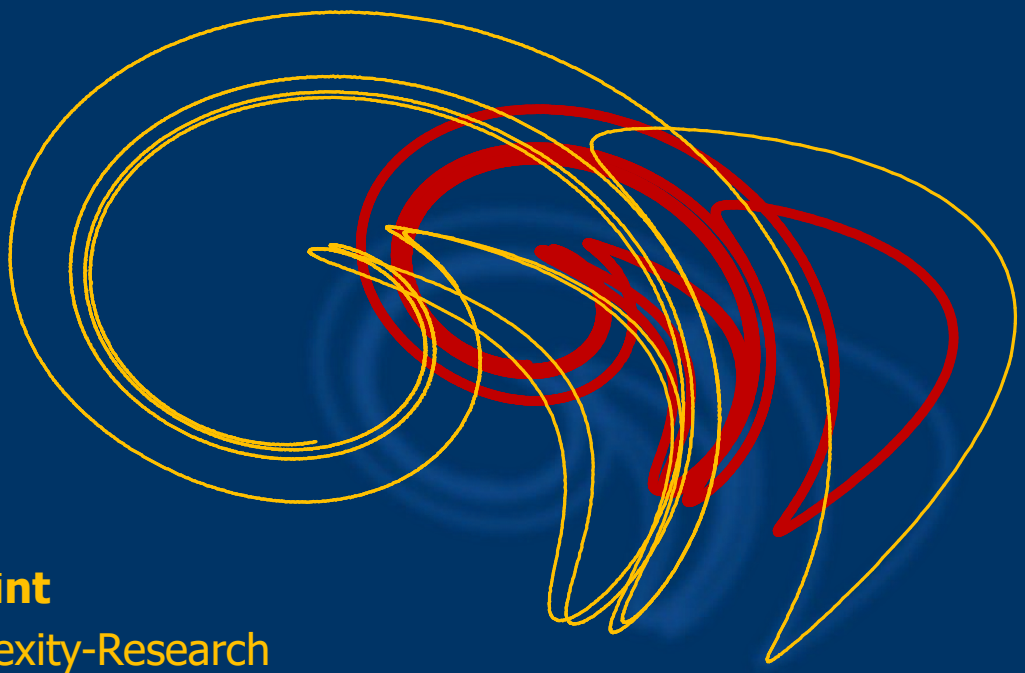

Preprint
Complexity-Research

Guido Strunk

Über die Unvereinbarkeit von KI mit einer vollständigen Überprüfbarkeit ihrer Algorithmen

Preprint, erschienen in:

Guido Strunk (2025) Über die Unvereinbarkeit von KI mit einer vollständigen Überprüfbarkeit ihrer Algorithmen. The Defence Horizon Journal. Special Edition, Current Security and Democracy Challenges for Europe, 1 (2025), 28-33.

Schönbrunner Str. 32 / 20
A-1050 Wien
guido.strunk@complexity-research.com
www.complexity-research.com

Über die Unvereinbarkeit von KI mit einer vollständigen Überprüfbarkeit ihrer Algorithmen

Guido Strunk

Autor: FH-Hon.Prof. Priv.-Doz. Dr. Dr. Dipl.-Psych. Guido Strunk; Forschungsschwerpunkte: Methoden und Anwendungen der Komplexitätsforschung in Psychologie und Ökonomie. Zahlreiche für den vorliegenden Artikel relevante Publikationen, u. a.: Guido Strunk, *Systemische Psychologie: Grundlagen einer allgemeinen Systemtheorie der Psychologie* (Complexity-Research, 2024). *Free Hugs. Komplexität verstehen und nutzen* (Complexity-Research, 2021). *Leben wir in einer immer komplexer werdenden Welt? Methoden der Komplexitätsmessung für die Wirtschaftswissenschaft.* (Complexity-Research, 2019). Bei den in dem vorliegenden Artikel vertretenen Ansichten handelt es sich um die des Autors. Diese müssen nicht mit jenen der FH Campus Wien, der TU Dortmund oder Complexity-Research Wien übereinstimmen.

Abstract: Die Bedeutung und Nutzung von Künstlicher Intelligenz (KI) hat in den letzten Jahren rasant zugenommen. Neben den bereits länger bekannten Herausforderungen, die mit dem Einsatz technischer Entscheidungsunterstützung in unübersichtlichen Situationen verbunden sind (z. B. Reliabilität, Sensitivität, Spezifität), ist im Zusammenhang mit dem Einsatz von KI auch mit neuartigen Herausforderungen zu rechnen. Diese betreffen vor allem die Nachvollziehbarkeit und Überprüfbarkeit der Mechanismen, die eine KI zu einer konkreten Entscheidungsleistung veranlassen. Dabei geht es u. a. um die Objektivität, Reliabilität, Validität, ethisch-moralische Integrität und Evidenzbasierung der Entscheidungspfade innerhalb der Black Box einer KI.

Abstract (Englisch): The importance and use of artificial intelligence (AI) has grown rapidly in recent years. In addition to the well-known challenges associated with the use of technical decision support in uncertain situations (e.g. reliability, sensitivity, specificity), new challenges are expected in connection with the use of AI. These relate primarily to the traceability and verifiability of the mechanisms that lead an AI to make a particular decision. This includes the objectivity, reliability, validity, ethical and moral integrity and evidence-based nature of the decision paths within the black box of an AI.

Bottom-line-up-front: Aus der Perspektive der neueren Komplexitätsforschung und der darauf aufbauenden Systemischen Psychologie folgt eine weitgehende Unvereinbarkeit von KI und vollständiger Vorhersagbarkeit bzw. Kenntnis ihrer Algorithmen.

Problemdarstellung: Ist es möglich, die Entscheidungsregeln innerhalb einer KI so gut zu verstehen, dass eine Beurteilung der Gültigkeit ihrer Entscheidungen möglich ist?

Was nun?: Bei Entscheidungen über den Einsatz von KI-Systemen ist zu berücksichtigen, dass eine vollständige Vorhersehbarkeit und Kenntnis ihrer Algorithmen grundsätzlich nicht möglich ist. Solange keine psychologischen Zuverlässigkeitstests für KI-Systeme vorliegen, sollte ihr Einsatz auf risikoarme Bereiche beschränkt bleiben, die von ausreichend geschulten und verantwortlichen Personen überwacht werden können. Die Grundhaltung gegenüber KI-Systemen sollte von der Einsicht getragen sein, dass sich Künstliche Intelligenz hinsichtlich ihrer Autonomie und Undurchschaubarkeit deutlich von herkömmlichen technischen Hilfsmitteln unterscheidet.

Einleitung

In Brechts „Leben des Galilei“[1] fordert Galilei die Gelehrten auf, sich mit einem Blick durch das Fernrohr von der Existenz der Jupitermonde zu überzeugen. Der Mathematiker unter den Gelehrten weigert sich mit den folgenden Worten: „Man könnte versucht sein zu antworten, dass Ihr Rohr, etwas zeigend, was nicht sein kann, ein nicht sehr verlässliches Rohr sein müsste, nicht?“[2] Dann konkretisiert er: „Wenn man sicher wäre, dass Sie sich nicht noch mehr erregten, könnte man sagen, dass, was in Ihrem Rohr ist und was am Himmel ist, zweierlei sein kann.“[3] Tatsächlich erregt sich Galilei gerade wegen der Weigerung der Gelehrten, naturwissenschaftliche Beweise auch nur eines Blickes zu würdigen. Elegant spielt Brecht den empirischen Beweis gegen die bornierte Ideologie aus. Dabei hilft, dass für Brechts Publikum längst klar ist, wessen Weltsicht die richtige ist.

Descartes[4] hatte etwa zu der gleichen Zeit wie Galilei den Zweifel zur zentralen Methode der Wissenschaften erklärt. Vor diesem Hintergrund erscheint die Kritik des Mathematikers durchaus berechtigt. Erst wenn die Funktionsweise eines Fernrohrs vollständig verstanden ist und alle relevanten Details seiner Abbildungsleistung und möglicher Artefakte bekannt sind, kann es als „verlässliches“ Beobachtungsinstrument gelten. Wenn heute über den Einsatz neuer, gemeinhin als „Künstliche Intelligenz“ bezeichneter Technologien diskutiert wird, mit deren Hilfe z. B. Bedrohungslagen eingeschätzt, verdächtige Aktivitäten erkannt oder Symptommuster in medizinischen Befunden interpretiert werden, stellen sich die gleichen Fragen: Wie funktioniert das Beobachtungsinstrument, wie kommt es zu seinen Urteilen, produziert es Artefakte, beruht es gar auf sachfremden, diskriminierenden Annahmen?

Forschungsmethodische Probleme, die typischerweise mit Vorhersage- oder Klassifikationsalgorithmen verbunden sind, können auch bei KI auftreten. Diese betreffen Gütekriterien wie Reliabilität, Sensitivität, Spezifität, Normierung, Overfitting und lassen sich gut quantifizieren und ggf. optimieren.[5] Dabei geht es im Wesentlichen darum, zu prüfen, ob eine KI zu ähnlich Vorhersagen, Klassifikationen oder – allgemein – Ergebnissen kommt wie Menschen. Beispielsweise sollte eine KI bei der Befundung eines Röntgenbildes nicht schlechter abschneiden als ein darin erfahrener Mensch. Neben diesen altbekannten, ergebnisbezogenen Gütekriterien ist in Bezug auf den Einsatz von KI aber auch mit neuartigen, methodenspezifischen Problemen zu rechnen. Diese betreffen die bereits

aufgeworfenen Fragen nach der Nachvollziehbarkeit der Mechanismen, die eine KI zu einer konkreten Entscheidungsleistung veranlassen. Es geht also um die Objektivität, Reliabilität, Validität, ethisch-moralische Integrität und Evidenzbasierung der Entscheidungspfade innerhalb der Black Box einer KI.

Autonomie

Descartes gilt als einer der zentralen Begründer des modernen wissenschaftlichen Denkens. In seiner „Methode"[6] beschreibt er Wissenschaft als umfassende Kritik: Alle Überzeugungen müssten hinterfragt werden, bis man auf einem festen Grund stehe, von dem aus ein schlüssiges und geprüftes Weltbild Baustein für Baustein errichtet werden könne. Diesen festen Grund sah er in dem berühmten „Cogito ergo sum".[7] Würde ein denkendes Wesen an der eigenen Existenz zweifeln, käme es zu einem Widerspruch (oder einer psychischen Erkrankung). Dabei versteht Descartes unter Denken, entgegen der landläufigen Meinung, auch Wahrnehmung und Emotionen. Alles, was wir heute als Bewusstseinsvorgänge bezeichnen würden, nennt er zusammenfassend „Denken".[8] Dieses denkende Ich steht für Descartes am Anfang jeder Weltbetrachtung. Im Gegensatz zu den Bewegungen des Körpers, die durch Ursachen ausgelöst werden, ist das denkende Ich der Startpunkt, der die Bewegung der Glieder in Gang setzt. Vor dem Ich steht aus seiner Sicht keine weitere erkennbare Ursache. Tatsächlich empfinden gesunde Menschen sich selbst und das „Ich" ihres Bewusstseins als verantwortlich für z. B. wichtige Entscheidungen und Handlungen. Menschliches Zusammenleben, Moral, Verantwortung, Rechtsprechung wären unmöglich, wenn nicht das Ich-denkende Ich der verantwortliche Ausgangspunkt wäre.

Das Problem ist so alt wie die Menschheit. Schon Aristoteles unterschied zwischen unbelebten Dingen, die sich nur bewegen, wenn sie durch eine von außen kommende kausale Ursache (causa causalis) dazu veranlasst werden, und lebenden Systemen, die sich aus sich selbst heraus auf Ziele hin bewegen (causa finalis).[9] Es ist hier nicht der Ort, eine Diskussion über die Freiheit des Willens zu führen,[10] aber für die oben gestellte Frage ist es relevant, dass lebende Systeme von sich aus Verhaltensoptionen wählen und sich deutlich autonomer verhalten als ein Stein, der einen Abhang hinunterrollt. Niemand würde in einem Stein einen Willen oder eine Intelligenz vermuten, die ihn bewegt, ganz

im Gegensatz zu lebenden Systemen. Die Neurobiologen Varela und Maturana sehen in der selbstorganisierten Autonomie das zentrale Definitionskriterium für Leben:[11] Leben ist die autonome, selbstorganisierte Aufrechterhaltung der eigenen Struktur (sie bezeichnen das Grundprinzip des Lebens als Autopoiese,[12] was Selbsterzeugung bedeutet). Niemand kommt von außen und programmiert Zellen, Organe oder Organismen. Die Unterscheidung zwischen der Selbstorganisation lebender Systeme einerseits und der einfachen kausalgesetzlichen Fremdorganisation von Maschinen andererseits war für Descartes zentral. Descartes soll selbst eine lebensecht wirkende Automatenfigur besessen haben.[13] Automatenfiguren ahmen menschliches oder tierisches Verhalten nach, sind aber an ihrer Unselbständigkeit leicht als von außen programmierte Maschinen zu erkennen. Nach Descartes fehlt ihnen die Fähigkeit, ein Ich-denkendes Ich hervorzubringen, das sie sinnvoll steuern könnte. Ein zentraler Trick des Lebens und des Denkens scheint in der autonomen Selbstorganisation zu liegen.

An dieser Stelle ist es sinnvoll, den Begriff der Intelligenz näher zu betrachten. Die Psychologie versteht darunter die Fähigkeit, potenzielle Hindernisse, die das Erreichen von Zielen beeinträchtigen könnten, eigenständig zu identifizieren und zu überwinden. Sie erfordert Wissen, schlussfolgerndes Denken, die Reflexion der eigenen Stellung in der Welt, die Fähigkeit zu abstrahieren und aus Erfahrungen zu lernen. Dabei geht es nicht um Wissen im Sinne von Bücherwissen, sondern um ein umfassenderes Verständnis und die Fähigkeit, mit den Veränderungen in der Umwelt Schritt zu halten, daraus einen Sinn abzuleiten und herauszufinden, was zu tun ist. [14] Intelligenz ist also ein innerpsychischer Prozess, der beim Gegenüber nicht direkt beobachtbar ist. Menschen erleben aber ihr eigenes Denken als einen bewussten Prozess, den sie daher auch bei anderen vermuten. Mit ähnlichen Beobachtungskategorien treten sie auch Maschinen gegenüber.[15] Das zentrale Kriterium für die Beurteilung von Intelligenz ist auch dort die zielgerichtete Autonomie im Umgang mit dem Unbekannten. Allerdings lassen sich Menschen in Bezug auf Technik gerne täuschen. Sie sind durchaus bereit, Computerprogramme wie Weizenbaums Eliza[16] aus den 1960er Jahren für intelligent zu halten, obwohl der Algorithmus nur eine geschickte Vortäuschung von Intelligenz darstellt.

Dies wirft erneut die Frage auf, wie KI im Detail funktioniert und ob es überhaupt möglich ist, einer Maschine Intelligenz oder andere Bewusstseinsprozesse einzuhauchen.

Descartes – und viel später Kant[17] – halten dies für unmöglich und lehnen die Maschinenmetapher menschlichen Verhaltens strikt ab. Das denkende Ich kann für sie nicht das Ergebnis einer Maschine sein. Denn Maschine und Ich sind einander in jeder Hinsicht wesensfremd. Ähnliche Fragen beschäftigen auch Erwin Schrödinger, der aus Sicht der Physik zum gleichen Schluss kommt,[18] aber auch ganz grundsätzlich fragt, wie aus Unordnung überhaupt (sinnvolle) Ordnung entstehen kann.[19] Als Beispiel führt er das menschliche Gehirn an, das aus ca. 100 Milliarden Neuronen besteht: Wie kann sich dieses gigantische System auf ein einziges, in sich konsistentes, Ich-Bewusstsein einigen?[20]

Selbstorganisierte Ordnungsbildung

Konkrete Antworten auf diese Fragen wurden erst in den 1960er und 1970er Jahren erarbeitet und basieren auf Erklärungsmodellen, die heute gerne als Theorien komplexer Systeme zusammengefasst werden. Komplexe Systeme, die unzählige Elemente enthalten können (z. B. Billionen Zellen im menschlichen Körper[21] oder mehrere Millionen Tierarten auf unserem Planeten[22]), bilden autonom (d. h. selbstorganisiert und nicht von außen erzwungen) Verhaltensmuster aus, die zu den inneren und äußeren Anforderungen dieser Systeme passen. Theorien wie die aus der Laserphysik stammende Synergetik[23] oder die mit dem Nobelpreis für Chemie ausgezeichnete Theorie dissipativer Strukturen[24] können zeigen, wie autonome Strukturbildung funktioniert. Dass dies möglich ist, war für die Naturwissenschaften eine Sensation. Denn bisher kannte man nur den Zerfall einer bereits bestehenden Ordnung. Die Entstehung einer neuen Ordnung – quasi aus dem Nichts der Unordnung – scheint auf den ersten Blick im Widerspruch zum zweiten Hauptsatz der Thermodynamik zu stehen.[25] Die Entdeckung der Selbstorganisation war daher der Schlüssel zu einem neuen Verständnis der Natur.

Auch in der unbelebten Welt gibt es Selbstorganisation, die zielgerichtet erscheint, weil sie eine autonome Anpassungsleistung an unvorhergesehene äußere und innere Einflüsse darstellt. Aus diesen Erkenntnissen erwuchs bald die Hoffnung, dass Selbstorganisation in lebenden und psychischen Systemen auf ähnlichen Prinzipien beruhen könnte.[26] Es wurde aber auch erkannt, dass das Verhalten eines sich selbst organisierenden, autonomen Systems nicht im Detail vorhergesagt werden kann, wie dies bei tri-

vialen Systemen der Fall ist. Komplexe Systeme bringen selbstorganisierte Ordnungsstrukturen hervor, die nicht im Detail vorhergesagt werden können.[27] Es handelt sich also um weitgehend neuartige Ordnungsstrukturen, die zudem eine adäquate Antwort des Systems auf interne und externe Rahmenbedingungen darstellen. Vereinfacht ausgedrückt: Komplexe Systeme können aus sich selbst heraus kreative, neue Lösungen für Probleme entwickeln.

Wenn man möchte, dass eine KI spannende Antworten auf Fragen gibt, dann sollten diese neu, kreativ, überraschend sein, aber dennoch sinnvoll und der Frage angemessen. Die genannten Theorien zeigen, wie komplexe Systeme solche Antworten selbstorganisiert hervorbringen können. In dem Maße, in dem Künstliche Intelligenzen dem Vorbild komplexer Systeme folgen (was z. B. bei der Verwendung neuronaler Netze der Fall wäre), können sie tatsächlich neuartige Antworten hervorbringen. Gleichzeitig entziehen sie sich damit zwangsläufig einer detaillierten Nachvollziehbarkeit. Denn diese kann es in komplexen Systemen nicht geben.

Um zu verstehen, wie selbstorganisierte Ordnungsbildung funktioniert, mussten die Naturwissenschaften viele Grundüberzeugungen über Bord werfen, die Galilei und Newton für unverzichtbar erklärt hatten. Galilei war davon überzeugt, dass die vielfältigen Wechselwirkungen in der Welt nicht verstanden werden können, wenn sie nicht in kleinere Einheiten zerlegt werden. Diese Zerlegung (griech. Analyse) gilt bis heute als Inbegriff des Verstehens. Empirische Experimente isolieren in einem System eine Variable, die gezielt verändert wird, während alles andere gleich bleibt (lat. ceteris paribus). So kann in einem größeren System nacheinander jede Variable einzeln untersucht werden. Die Idee war, die Welt Baustein für Baustein zu verstehen.[28] Heute wissen wir, dass die Zerlegung nur bei trivialen Systemen gelingt. Voraussetzung ist, dass sich die Bausteine isoliert genauso verhalten wie auch im zusammengesetzten System. In der Mechanik von Galilei und Newton ist dies in der Tat häufig der Fall. Komplizierte Uhren (grand complication), kunstvolle Automatenfiguren, mechanische Rechenmaschinen, die aus sehr vielen Teilen bestehen, können bis ins kleinste Detail verstanden werden, weil hier das Ganze die Summe der Teile ist.[29]

Es ist eine relativ neue Erkenntnis, dass es andererseits auch Systeme gibt, bei denen das Ganze etwas anderes ist, als aus der Kenntnis der Einzelteile zu erwarten wäre.[30] Die

bereits erwähnten Selbstorganisationstheorien zeigen mathematisch und empirisch zweifelsfrei, dass die Elemente der sog. Mikroebene eines Systems unter bestimmten Bedingungen auf der Ebene des Gesamtsystems (der sog. Makroebene) neuartige Eigenschaften hervorbringen, die auch als emergente Eigenschaften bezeichnet werden.[31] Die ca. 100 Milliarden Neuronen des menschlichen Gehirns bringen durch emergente Selbstorganisation das kohärente Ich-Erleben hervor. Kein Neuron ist ein Ich, hat Bewusstsein oder ist intelligent. Keine Zelle hat eine Ahnung von der Existenz anderer Zellen. Erst durch das Zusammenspiel der Elemente entsteht auf der Makroebene etwas, was auf der Mikroebene nicht existiert. Das Gesamtsystem erzeugt selbstorganisiert (autonom und nicht vorprogrammiert) das makroskopische Muster. Dieses kann daher auch nicht durch Analyse der Einzelteile verstanden werden. Entgegen Galileis Forderung kann man ein komplexes System nicht zerlegen, ohne genau die Eigenschaften zu zerstören, die man verstehen will, z. B. Denkprozesse, Intelligenz, Kommunikation. Kein Wunder, dass die Neurobiologie das Bewusstsein nicht finden kann. Es versteckt sich nicht in der Zelle, sondern ist eine emergente Eigenschaft des gesamten Systems.[32] Schon Descartes wusste: Denken ist etwas ganz anderes als ein mechanisches Ursache-Wirkungs-Prinzip. Was ihm noch nicht klar war: In komplexen Systemen kann durch Selbstorganisation ein emergentes makroskopisches Muster entstehen. Zwar geht das makroskopische Muster naturgesetzlich aus der Mikroebene hervor, aber es ist nicht möglich, auf der Mikroebene konkret und detailliert vorherzusagen, welches Verhaltensmuster sich auf der Makroebene einstellen wird.

Selbstorganisation kann auch in komplexen Systemen triviale Ordnungsmuster hervorbringen. Die Elemente auf der Mikroebene einigen sich dann auf etwas, das zumindest auf der Makroebene leicht verständlich ist und auf dieser Ebene auch im Detail vorhergesagt und umfassend verstanden werden kann. Es hat sich jedoch gezeigt, dass Systeme auch hochkomplexe Dynamiken erzeugen können, die aufgrund des Schmetterlingseffekts[33] nie vollständig vorhersagbar sind. Ein System mit Schmetterlingseffekt erzeugt zwar weiterhin gut angepasste Verhaltensmuster, reagiert aber auf mikroskopische Einflüsse exponenziell verstärkend, was eine Vorhersage schnell unmöglich macht. Tatsächlich ist es gerade das deterministische Chaos des Schmetterlingseffekts, das komplexen Systemen hilft, sich in ebenso komplexen Umwelten zu behaupten. Insofern ist der Schmetterlingseffekt trotz seiner begrenzten Vorhersagbarkeit ein Erfolgsgeheimnis der

Natur. Der Beherrschbarkeit, Prognostizierbarkeit und gezielten Beeinflussbarkeit komplexer Systeme sind somit durch zwei Prinzipien Grenzen gesetzt: Von der Mikroebene kann nicht auf das Verhalten der Makroebene geschlossen werden; und die Makroebene kann einen Schmetterlingseffekt enthalten.

Da aber komplexe makroskopische Muster passende Antworten auf komplexe Probleme liefern, stellt sich die Frage, wie sie im Computer erzeugt werden können. Die Lösung besteht darin, Selbstorganisationsprozesse im Computer anzuregen. Antworten werden nicht fertig formuliert vorgegeben, sondern eine Mikroebene z. B. eines neuronalen Netzes wird angeregt, selbst makroskopische Muster zu erzeugen. In dem Maße, in dem dies gelingt, sind diese Muster „intelligente" selbst generierte Antworten. Dieser Vorteil wird allerdings durch den Verlust an Beherrschbarkeit erkauft, die in solchen Systemen aus den genannten Gründen nicht gegeben sein kann.

Machine Learning

Nach Jiang et al.[34] beginnt das Google Machine Learning Manual mit dem Satz "If you can build a simple rule-based system that does not require machine learning, do that". Dies fasst das bisher Gesagte gut zusammen. Oben wurde gezeigt, dass bestimmte Systeme bis ins Detail verstanden werden können. Der Konstruktionsplan einer Uhr, so kompliziert er auch sein mag, kann Hebel für Hebel, Zahnrad für Zahnrad nachvollzogen und entsprechend programmiert werden. Für die klassischen Naturwissenschaften galt die Überzeugung, dass alles in der Welt nach klaren Regeln funktioniert. Für die Informatik folgte daraus, dass man erst die Regeln herausfinden musste, bevor man eine Maschine programmieren konnte. Eine Zeit lang versuchte man dies mit sog. Expertensystemen. Ausgangspunkt waren Interviews mit Fachleuten, um deren Vorgehen zu verstehen. Dieser Ansatz ist gescheitert,[35] u. a. weil vieles, was Menschen tun, nicht mit Regeln beschrieben werden kann. Insbesondere psychologische Studien konnten zeigen, dass Menschen, die gut darin sind, z. B. einen Tumor auf einem Röntgenbild zu erkennen, oft nicht in der Lage sind, die Regeln zu benennen, nach denen sie dies tun. Mit zunehmender Erfahrung automatisiert sich das komplexe Verhalten und ist dann verbal nicht mehr zugänglich.[36] Dies entspricht den bereits genannten Vorhersagbarkeitsproblemen, die in komplexen Systemen regelmäßig auftreten.

Machine Learning verlagert den Lernprozess in die Maschine. Interviews mit Fachleuten werden überflüssig. Die Maschine verknüpft selbst den Input (z. B. Röntgenbild) mit dem Output (Diagnose). Dabei kommt sie zu Lösungen, die im Idealfall funktionieren (gleiche Diagnose), aber in der Regel nicht dem entsprechen, was Fachleute tun. Einfache Methoden des maschinellen Lernens werden seit vielen Jahren in der Statistik eingesetzt. Medizinische Prognosemodelle wie SCORE2[37] verwenden Informationen über Alter, Geschlecht und Vorerkrankungen, um das Risiko für kardiovaskuläre Ereignisse vorherzusagen. Der maschinelle Lernansatz wählt die Variablen für die Vorhersage aus und gewichtet sie automatisch. Die Vorhersage ist dann die Summe dieser gewichteten Risikofaktoren. Das Modell ist daher leicht verständlich. Allerdings können die identifizierten Risikofaktoren inhaltlich, d. h. medizinisch, völlig falsch sein und das Modell kann trotzdem funktionieren, weil der selbstlernende Algorithmus so konstruiert ist, dass er sich optimal an die Daten anpasst. Daher ist es hilfreich, dass diese einfachen Verfahren des maschinellen Lernens transparent arbeiten. Dabei darf zudem nicht vergessen werden, dass das Grundmodell der gewichteten Summe nur ein Modell ist und in der Natur so nicht vorkommt.[38] Die bio-psycho-sozialen Prozesse, die zu einem kardiovaskulären Ereignis führen, sind keine gewichtete Summe von Risikofaktoren. Dennoch funktionieren solche Modelle aufgrund ihrer mathematischen Anpassungsfähigkeit recht gut, ihre Ergebnisse sind nachvollziehbar und wissenschaftlich diskutierbar. Mit anderen Worten: Einfache Algorithmen des maschinellen Lernens bilden Daten in einem Modell ab, das in der Regel nicht korrekt, aber verständlich und nützlich ist.

Solche einfachen Modelle reichen nicht aus, um komplexe Phänomene abzubilden. Komplexe Daten benötigen komplexe Modelle. Neuronale Netze und moderne KI-Systeme arbeiten daher in vielerlei Hinsicht anders als klassische Lernalgorithmen. Sie rekonstruieren ein komplexes System (neuronales Netz), in dem es durch Lernprozesse zu einer emergenten Strukturbildung auf der Makroebene kommt. Tritt Emergenz auf, so ist das makroskopische Muster aus den Algorithmen der Mikroebene nicht mehr nachvollziehbar. Zudem kann das makroskopische Muster selbst durch den Schmetterlingseffekt sehr komplex sein. Das ist gewollt, denn es liegt in der Natur der Sache, dass komplexe Strukturen nur durch komplexe Systeme abgebildet werden können.[39]

Um es mit Descartes zu sagen: Unser Ich-denkendes Ich begegnet den komplexen Anforderungen der Welt nicht mit einfachen Ursache-Wirkungs-Mechanismen, sondern ist

selbst ein komplexes System. Eine KI, die intelligent sein soll, muss zwangsläufig ein komplexes System sein. Schätzungen gehen davon aus, dass GPT-4[40] ca. 100 Billionen Parameter selbstlernend an die Datenstrukturen angepasst hat.[41] Diese bilden die Systemgleichungen, mit denen GPT-4 einen Text als makroskopisches Muster erzeugt.[42] Selbst wenn es sich um die Parameter einer einfachen additiven Gleichung handeln würde, sind dies so viele Parameter, dass niemand sie grob überschlagen kann, um zu verstehen, wie GPT-4 tut, was es tut.

Zusammenfassung

Wenn der Inhalt trivial ist, dann kann auch der Algorithmus trivial sein. Komplexe Inhalte erfordern jedoch zwingend komplexe Maschinen.[43] Das zentrale Kriterium für diese Form der Komplexität ist die Irreduzibilität.[44] Lässt sich das Verhalten der Maschine auf einen schrittweise nachvollziehbaren Algorithmus reduzieren, dann ist es nicht komplex. Entsteht jedoch ein emergentes makroskopisches Muster, so kann dieses nicht mehr aus den Algorithmen der Mikroebene abgeleitet werden. Das emergente Muster ist dann nicht auf die Elemente der Mikroebene reduzierbar. Denken, Intelligenz und Wahrnehmung wurden oben als Prozesse dargestellt, die den Neuronen fremd sind. Denken kann nicht auf der Ebene von Neuronen verstanden werden, da es algorithmisch nicht auf das reduziert werden kann, was einzelne Neuronen tun. Wenn also eine KI ein makroskopisches Muster bildet, kann der Fall eintreten, dass auch dieses makroskopische Muster nicht auf die einzelnen Parameter des neuronalen Netzes reduziert werden kann. Die ersten Berichte über irreduzible emergente Prozesse in großen Sprachmodellen wurden von Wei et al.[45] vorgelegt und seitdem ausführlich diskutiert.

Irreduzibilität ist ein zentrales Merkmal natürlicher Intelligenz, und in dem Maße, in dem Maschinen diesem Vorbild folgen, streben auch sie nach Irreduzibilität. Sie können dann nicht mehr auf der Ebene des Programmcodes verstanden werden. Je intelligenter eine KI wird, desto mehr müssen wir ihr mit einer Art Maschinen-Psychologie begegnen. Die Frage ist dann eher, ob man ihr vertrauen kann, und weniger, wie sie im Detail funktioniert.

Bei „echter“ Intelligenz und tatsächlicher Irreduzibilität kann man sie nicht wie ein Teleskop auseinandernehmen und Linse für Linse auf Zuverlässigkeit und Artefaktfreiheit

prüfen. Daraus folgt: Die Hoffnung, KI könne helfen, die Welt besser zu verstehen, beruht auf einem Missverständnis, denn Intelligenz ist eine emergente Eigenschaft komplexer Systeme, die zur Selbstorganisation fähig sind. In dem Maße, in dem Maschinen zur Selbstorganisation fähig werden, werden ihre Algorithmen zwangsläufig immer weniger durchschaubar. Dieser Aspekt der Unvereinbarkeit von KI mit einer vollständigen Überprüfbarkeit ihrer Algorithmen ist beim Einsatz der Technologie zu berücksichtigen und mahnt zur Vorsicht.

Endnoten, Literatur

[1] Bertolt Brecht, *Leben des Galilei: Schauspiel* (Suhrkamp, 1989/1939).

[2] Ibid., 48.

[3] Ibid., 48.

[4] René Descartes, *Discours de la Méthode. Bericht über die Methode. Französisch/Deutsch* (Reclam, 2001/1637).

[5] Gustav Lienert und Ulrich Raatz, *Testaufbau und Testanalyse* (Beltz, 1994); Christopher M Florkowski, "Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests," *The Clinical Biochemist Reviews* 29, no. Suppl 1 (2008); Frank E. Jr. Harrell, Kerry L. Lee, and Daniel B. Mark, "Tutorial in Biostatistics – Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," *Statistics in Medicine* 15 (1996).

[6] René Descartes, *Discours de la Méthode. Bericht über die Methode. Französisch/Deutsch* (Reclam, 2001/1637).

[7] René Descartes, *Prinzipien der Philosophie (Text nach der Übersetzung durch Julius Heinrich von Kirchmann von 1870)* (Verlag von L. Heimann, 1870/1644), 5f.

[8] Ibid.

[9] Aristoteles, *Physik (Übersetzt von Cristian Hermann Weiße, 1829)* (Hofenberg, 2016/4. Jhdt. v. Chr.).

[10] Guido Strunk, *Systemische Psychologie: Grundlagen einer allgemeinen Systemtheorie der Psychologie* (Complexity-Research, 2024).

[11] Francisco J. Varela, Humberto R. Maturana, and Ricardo B. Uribe, "Autopoiesis: The Organization of Living Systems, Its Characterization and a Model," *Biosystems* 5, no. 4 (1974); Humberto R. Maturana and Francisco J. Varela, *The Tree of Knowledge. The Biological Roots of Human Understanding* (Shambhala, 1987).

[12] Francisco J. Varela, Humberto R. Maturana, and Ricardo B. Uribe, "Autopoiesis: The Organization of Living Systems, Its Characterization and a Model," *Biosystems* 5, no. 4 (1974).

[13] Gaby Wood, *Edison's Eve. A Magical History of the Quest for Mechanical Life* (Anchor Books, 2002), 4.

[14] Richard E. Nisbett et al., "Intelligence: new findings and theoretical developments," *American Psychologist* 67, no. 2 (2012): 131; Linda S. Gottfredson, "Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography," *Intelligence* 24, no. 1 (1997/1994): 13.

[15] Alan M. Turing, "Computing machinery and intelligence," *Mind, a Quarterly Review* LIX, no. 236 (1950).

[16] Joseph Weizenbaum, “ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine,” *Communications of the ACM* 9, no. 1 (1966).

[17] Immanuel Kant, *Grundlegung zur Metaphysik der Sitten* (J.F. Hartknoch, 1786).

[18] Erwin Schrödinger, *Geist und Materie* (Diogenes, 1989/1958).

[19] Erwin Schrödinger, *Was ist Leben? Die lebende Zelle mit den Augen des Physikers betrachtet* (Piper, 1989/1944).

[20] Erwin Schrödinger, *Geist und Materie* (Diogenes, 1989/1958), Kap. 4.

[21] Steffen Schaal, Konrad Kunsch und Steffen Kunsch, *Der Mensch in Zahlen* (Springer, 2016).

[22] Camilo Mora et al., “How many species are there on Earth and in the ocean?,” *PLoS biology* 9, no. 8 (2011).

[23] Hermann Haken, *Synergetics. An Introduction. Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology* (Springer, 1977).

[24] Grégorire Nicolis and Ilya Prigogine, *Self-Organization in Nonequilibrium Systems* (John Wiley and Sons, 1977); Ilya Prigogine, “Time, structure, and fluctuations,” *Science* 201, no. 4358 (1978).

[25] Rudolf Clausius, *Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie: vorgetragen in der naturforsch. Gesellschaft den 24. April 1865* (éditeur inconnu, 1865); Rudolf Clausius, *Abhandlungen über die mechanische Wärmetheorie* (F. Vieweg und Sohn, 1864).

[26] Hans-Peter Dürr, *Das Netz des Physikers. Naturwissenschaftliche Erkenntnisse in der Verantwortung* (Deutscher Taschenbuch Verlag, 1990), 74.

[27] Ilya Prigogine, *Die Erforschung des Komplexen. Auf dem Weg zu einem neuen Verständnis der Naturwissenschaften* (Piper, 1987); David Ruelle and Floris Takens, “On the Nature of Turbulence,” *Communications in Mathematical Physics* 20 (1971); Edward N. Lorenz, “Predictability: Does the flap of a butterfly’s wings in Brazil set off a tornado in Texas?” (paper presented at the AAAS Conference, Section on Environmental Sciences. New Approaches to Global Weather: GARP (The Global Atmospheric Research Program), Washington, 29.12.1972); Hermann Haken, “Entwicklungslinien der Synergetik, I,” *Naturwissenschaften* 75, no. 4 (1988); Hermann Haken, “Entwicklungslinien der Synergetik, II,” *Naturwissenschaften* 75, no. 5 (1988).

[28] Guido Strunk, *Systemische Psychologie: Grundlagen einer allgemeinen Systemtheorie der Psychologie* (Complexity-Research, 2024).

[29] Ibid.

[30] Hans-Peter Dürr, *Das Netz des Physikers. Naturwissenschaftliche Erkenntnisse in der Verantwortung* (Deutscher Taschenbuch Verlag, 1990), 74.

[31] Vgl. das Grundmodell der Synergetik in Günter Schiepek und Guido Strunk, *Dynamische Systeme. Grundlagen und Analysemethoden für Psychologen und Psychiater* (Asanger, 1994).

[32] Guido Strunk, *Systemische Psychologie: Grundlagen einer allgemeinen Systemtheorie der Psychologie* (Complexity-Research, 2024).

[33] Edward N. Lorenz, "Predictability: Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?" (paper presented at the AAAS Conference, Section on Environmental Sciences. New Approaches to Global Weather: GARP (The Global Atmospheric Research Program), Washington, 29.12.1972)

[34] Yuchen Jiang et al., "Quo vadis artificial intelligence?," *Discover Artificial Intelligence* 2, no. 1 (2022).

[35] Ibid.

[36] Michael Polanyi, *Tacit Knowledge* (Doubleday Publishers, 1966/1983).

[37] ESC Cardiovasc Risk Collaboration and SCORE2 Working Group, "SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe," *European Heart Journal* 42, no. 25 (2021).

[38] Herbert Stachowiak, *Allgemeine Modelltheorie* (Springer, 1973).

[39] Z. B. in Bezug auf Strukturen in der Natur und in ökonomischen Prozessen: Benoît B. Mandelbrot and Richard L. Hudson, *The (Mis)Behavior of Markets: A Fractal View of Risk, Ruin, and Reward* (Basic Books, 2004); Benoît B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, 1977).

[40] Josh Achiam OpenAI et al., "Gpt-4 technical report, 2024," *URL https://arxiv.org/abs/2303.08774* (2024).

[41] Siegfried Handschuh, "Grosse Sprachmodelle," *Informationswissenschaft: Theorie, Methode und Praxis* 8, no. 1 (2024).

[42] Vgl. die umfassende Darstellung verschiedener medizinischer Modelle in: Elliot Bolton et al., "Biomedlm: A 2.7 b parameter language model trained on biomedical text," *arXiv preprint arXiv:2403.18421* (2024).

[43] Alexandre K. Zvonkin and Leonid A. Levin, "The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms," *Russian Mathematics Surveys* 25, no. 6 (1970); Andrei Nikolajewitsch Kolmogorov, "Three Approaches to the Definition of the Concept Quantity of Information," *IEEE Transactions on Information Theory IT14* (1965); Gregory J. Chaitin, "Information Theoretic Computational Complexity," *IEEE Transactions on Information Theory IT20* (1974). Vgl. eine Zusammenfassung dieser Argumente in: Guido Strunk, *Leben wir in einer immer komplexer werdenden Welt? Methoden der Komplexitätsmessung für die Wirtschaftswissenschaft.* (Complexity-Research, 2019), 461–476.

[44] Guido Strunk, *Systemische Psychologie: Grundlagen einer allgemeinen Systemtheorie der Psychologie* (Complexity-Research, 2024), 36.

[45] Jason Wei et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682* (2022).