Warum es Willensfreiheit gibt und warum Roboter nichts empfinden

Inhaltsverzeichnis

1. Die Begründung der Willensfreiheit

Abstract

1. Zunächst wird der Unterschied zwischen Wirklichkeit und Beschreibung bestimmt. Davon ausgehend kann gezeigt werden, dass die physikalische Kausalität – im Folgenden als "Kausalität von unten" bezeichnet – *unvollständig* ist.

2. Dies ist eine notwendige Bedingung für die Annahme von Kausalität in komplexeren Ebenen der Wirklichkeit, die durch nicht-physikalische Gesetze geregelt werden. Diese Art von Kausalität – im Folgenden "Kausalität von oben" genannt – wird durch ein Beispiel erläutert und dann allgemein begründet.

3. Die Begründung gilt auch für das menschliche neuronale Netz. Daraus folgt, dass die geistige Ebene die *kausale Ebene* des Netzes ist.

4. Im Unterschied zu den Gesetzen der Physik sind die Gesetze der geistigen Ebene veränderbar. Da die geistigen Prozesse ursächlich sind, müssen auch diese Veränderungen der geistigen Tätigkeit zugeschrieben werden.

5. Für eine Willensentscheidung gilt daher Folgendes:

a) Sie ist kein physikalischer, sondern ein geistiger Prozess.

b) Der Entscheidungsprozess kann die Gesetze ändern, die vor seinem Beginn galten. Wenn aber erst durch diesen Prozess selbst bestimmt wird, was geschehen wird, kann die Entscheidung vorher nicht festgelegt sein.

Sie ist also frei.

1.1. Der Unterschied zwischen Wirklichkeit und Beschreibung

In unserem Universum scheint ganz allgemein Folgendes zu gelten:

***Alles, was existiert*, besteht aus elementaren Objekten, die miteinander wechselwirken. Wie sich diese Objekte verhalten, *wird vollständig durch physikalische Gesetze geregelt*. Somit folgt die gesamte zukünftige Entwicklung aus sogenannten "Anfangsbedingungen" – der Gesamtheit der Attribute aller Objekte zu irgendeinem Zeitpunkt – und physikalischen Gesetzen.**

In diesem Bild, das von der Naturwissenschaft so überzeugend präsentiert wird, ist anscheinend für nichts anderes Platz als für Physik. Gleichgültig, wie komplex die Aggregate auch sein mögen, zu denen sich die elementaren physikalischen Objekte zusammenfügen, gleichgültig, welch phantastische Kreationen die Evolution auch hervorbringt – *letztlich* bleibt alles Physik. Es ist einfach kein Platz für irgendetwas Anderes.

Dieser Sachverhalt lässt sich so konkretisieren:

In der soeben vorgestellten, *reduktionistisch* genannten Sichtweise der Wirklichkeit bleibt die Kausalität immer "unten", d.h. in der elementaren Schicht der Wirklichkeit. Alle anderen, komplexeren Schichten haben ihre Selbständigkeit verloren. Beschreibungen, die sich auf diese Schichten beziehen – etwa neuronale oder psychologische Beschreibungen menschlicher Handlungen, sind bloß vereinfachte, näherungsweise gültige Zusammenfassungen von Prozessen, die *eigentlich* physikalischer Natur sind.

Die Konsequenzen dieser Hypothesen sind ziemlich seltsam, um nicht zu sagen bizarr. Wenn wir etwa annehmen, wir hätten eine Behauptung B ***deshalb*** geäußert, ***weil*** sie logisch richtig ist, dann wäre das eine Selbsttäuschung: Es würde ja bedeuten, eine Kausalität auf der Ebene geistiger Prozesse zu postulieren, gewissermaßen eine Kausalität von "oben", was aber nach dem soeben Gesagten unzulässig ist. B wäre dann "kausal überbestimmt". Falls diese "Kausalität von oben" tatsächlich eine *selbständige Existenz* beanspruchen könnte – *zusätzlich* zur "Kausalität von unten" –, dann müsste es ja möglich sein, sich gegen die physikalische Kausalität zu entscheiden.

Es gäbe nur eine einzige Möglichkeit, dass B tatsächlich der Logik entsprechen könnte: Sie bestünde darin, dass die Evolution die physikalischen Prozesse in unserem Gehirn den Erfordernissen der Wirklichkeit soweit angepasst hätte, dass wir uns in einem für unser Überleben ausreichenden Maß logisch verhalten und denken. Aber ich wiederhole: die Überzeugung, dass wir uns *deshalb* so verhalten oder denken, *weil* es logisch ist, wäre eine Täuschung, eine List der Evolution, unser angepasstes Verhalten durch ein angenehmes Gefühl zu verstärken. Und, nebenbei gesagt, wir würden auch niemals feststellen können, ob so etwas wie "Logik" überhaupt existiert, da ja etwas *einzusehen* ebenfalls ein geistiger Prozess wäre, den es *als solchen* gar nicht gibt. Einsichten wären keine Einsichten, Gedanken keine Gedanken, der Geist wäre verschwunden, *wir selbst* hätten uns im Nebel der Selbsttäuschungen verflüchtigt...

Es ist also ein völlig absurdes Bild, das der Reduktionismus entwirft, und ich glaube, dass er nur deshalb so verbreitet ist, weil kein Reduktionist je die Konsequenzen seines Standpunkts vollständig berücksichtigt hat. (Wenn es doch einen gäbe, wäre er allerdings längst verstummt und daher unauffindbar.)

Ich will noch kurz auf die beiden populärsten Versuche eingehen, das Problem zu "entschärfen".

Der erste Einwand ist, dass wegen der quantenmechanischen Unschärfe eine "objektive Unbestimmtheit" in der Natur selbst existiert, sodass nicht behauptet werden kann, dass "die Zukunft aus Anfangsbedingungen und Gesetzen folgt". Es lässt sich aber behaupten, dass "die

Zukunft ausschließlich von Anfangsbedingungen und Gesetzen abhängt" – nur dass diese Gesetze eben nicht mehr deterministisch sind. Die nachstehenden Schlussfolgerungen bleiben dann gültig.

Am häufigsten wird gegen den Reduktionismus eingewendet, dass eine vollständige Reduktion in den meisten Fällen nicht gelungen ist und wohl auch niemals gelingen wird. Ich halte diesen Einwand für unzureichend: Ob es eine Reduktion *gibt,* kann nicht dadurch entschieden werden, ob *wir* dazu imstande sind, sie durchzuführen – das oben skizzierte Bild der Wirklichkeit, das die Grundlage des unglaublichen Erfolgs der Naturwissenschaft ist, wird durch die Einschränkungen, denen *unsere* Mittel und Fähigkeiten unterworfen sind, nicht in Frage gestellt, und das gilt auch für die Folgerungen aus diesem Bild.

Um diesen seltsamen Folgerungen zu entgehen, ist es daher notwendig, das Bild selbst in Frage zu stellen. Also fragen wir uns: *Ist die Behauptung A wahr?*

A: ***Alles, was geschieht, folgt aus physikalischen Gesetzen und Anfangsbedingungen.***

Beginnen wir mit einem Gedankenexperiment:

Wir betrachten folgendes Szenario: eine große Anzahl beliebiger materieller Objekte im leeren Raum, die sich relativ zueinander auf zufällige Weise bewegen, aber so, dass sie gravitativ aneinander gebunden bleiben.

Nehmen wir an, wir wären imstande, die Anfangsbedingungen – also die Gesamtheit der Attribute aller Objekte des Systems – *vollständig genau* zu erfassen und auf eine Beschreibung zu übertragen. Wir kümmern uns also nicht darum, dass wir nicht unendlich genau messen können oder dass wir nicht einmal dazu imstande sind, auch nur den Wert eines einzigen Attributs unendlich genau aufzuschreiben bzw. zu speichern. Außerdem nehmen wir an, dass unser Gravitationsgesetz *richtig* ist und dass wir alle erforderlichen Berechnungen mit unendlicher Genauigkeit durchführen können.

Jetzt vergleichen wir die Lage im *wirklich existierenden System* mit der Lage im *Beschreibungs-system*.

Unter den oben genannten Voraussetzungen wird sich *im existierenden System* ohne Zweifel genau das ereignen, was wir erwarten: jeder Körper wird sich exakt *so* verhalten, wie die Gravitation es ihm vorschreibt. Die Behauptung A scheint sich hier also zu bestätigen.

Und *im Beschreibungssystem*? Nun, hier ereignet sich zunächst *überhaupt nichts*. Obwohl wir in unsere korrekten Gleichungen die unendlich genauen Werte aller Attribute eingesetzt haben, so dass sie die Objekte und ihre zeitliche Entwicklung eigentlich perfekt repräsentieren, verhalten sich die Gleichungen doch nicht so wie die Objekte selbst: Während sich die *wirklich existierenden Objekte* von dem Zeitpunkt an, den wir zur Messung ihrer Attribute gewählt haben, *von selbst* weiter bewegen und auf diese Weise die gravitativ determinierte Dynamik des Systems vollziehen, tun das die Gleichungen offensichtlich nicht – sie bleiben einfach unverändert so stehen, wie wir sie notiert haben.

Das ist eigentlich vollkommen selbstverständlich. Ich war trotzdem ein wenig ausführlicher als nötig, weil wir damit auf einen außerordentlich wichtigen Sachverhalt gestoßen sind, der aber – vermutlich gerade *wegen* seiner trivial erscheinenden Selbstverständlichkeit – weder von der Philosophie noch von der Naturwissenschaft zur Kenntnis genommen worden ist.

Er lautet:

Satz:

Zwischen einem wirklich existierenden System und seiner Repräsentation besteht ein fundamentaler Unterschied: Das wirklich existierende System ist aktiv, die Repräsentation hingegen ist nicht aktiv.

Kehren wir zu unserem Gedankenexperiment zurück. Wir haben festgestellt: Im *existierenden System* wird sich jeder Körper exakt *so* verhalten, wie die Gravitation es ihm vorschreibt. Wird dadurch tatsächlich die Behauptung A bestätigt?

Die Antwort ist: *Nein, das wird sie nicht!* Wir haben ja dem wirklich existierenden System etwas hinzugefügt, was in A nicht enthalten ist: *Aktivität*.

Dass die Wirklichkeit *aktiv* ist, bedeutet, dass sich an jedem Punkt zu jeder Zeit genau das vollzieht, was zu geschehen hat. Es bedeutet, dass die Wirklichkeit nichts *berechnen* muss, dass sie kein Gesetz und keinen Algorithmus benötigt, weil sie einfach alle Einzelfälle gleichzeitig abarbeitet.

Offenbar ist aber *Aktivität* genau dasjenige, was nicht von der Wirklichkeit auf die Repräsentation übertragen werden kann. Es lässt sich zwar behaupten, dass die *Art der Aktivität* des Systems, ihre spezifische Struktur, in unseren Gleichungen des Gravitationsfeldes enthalten sein muss, aber die *Aktivität selbst* fehlt.

Halten wir fest: Aufgrund ihrer *Aktivität* schreitet die Wirklichkeit *von selbst* von der Gegenwart in die Zukunft voran. Das Beschreibungssystem weigert sich aber, uns diesen Gefallen zu erweisen. Um Information über die Zukunft des Systems zu erlangen, benötigen wir daher in der Beschreibung ein *mathematisches Verfahren,* das die fehlende Aktivität ersetzt.

Haben wir ein solches Verfahren? Zunächst ist klar, dass sich für eine "große Anzahl" von Körpern, die sich zufällig bewegen, unsere Gleichungen nicht lösen lassen. Tatsächlich haben wir nur eine einzige Möglichkeit, etwas über die weitere Entwicklung des Systems zu erfahren: Da wir das Gravitationsfeld kennen, können wir für jeden Körper berechnen, wohin er sich nach einem bestimmten Zeitintervall Δt *in diesem Feld* bewegt haben würde – und hier ist der Konjunktiv erforderlich, weil er sich selbstverständlich *nicht* in *diesem* Feld bewegt: es bewegt sich ja nicht nur der eben betrachtete Körper, sondern auch alle anderen, und das bedeutet, dass auch das Feld sich permanent verändert. Um aber überhaupt irgendetwas berechnen zu können, müssen wir für kleine Zeitintervalle das Feld als *statisch* annehmen. Wir führen dann dieselbe Art der Berechnung für alle Körper durch. Anschließend machen wir dasselbe für das nächste Zeitintervall usw.

Entscheidend ist, dass wir von Anfang an auf *Näherungen* angewiesen sind, und dass wir außerdem nicht wissen, in welchem Maß unsere Berechnungen von der Wirklichkeit abweichen. Spätestens nach dem nächsten Verzweigungspunkt – das ist ein Punkt in der Entwicklung eines Systems, an dem ein beliebig kleiner Unterschied in den Ausgangsbedingungen zu vollkommen unterschiedlichen Systemzuständen führen kann – wird unsere Voraussage zur reinen Glückssache.

Damit haben wir gezeigt, dass die Behauptung A falsch ist. Da es kein Verfahren gibt, mit dem man von der Gegenwart in die Zukunft gelangt, kann sie nicht aufrechterhalten werden.

Satz:

Es gibt Systeme, deren künftige Entwicklung nicht aus physikalischen Gesetzen und Anfangsbedingungen folgt.

Aber wird uns nicht *durch die Wirklichkeit selbst* andauernd vor Augen geführt, dass die Zukunft aus der Gegenwart folgt? Keineswegs. Was wir sehen, ist einfach nur, dass die Zukunft *auf die Gegenwart* folgt. Es ist bloß dieses suggestive, von der Physik vermittelte Bild der Wirklichkeit, das uns glauben lässt, alles "folge aus" Anfangsbedingungen und Gesetzen. Der Ausdruck "folgt aus" ist jedoch eine logische Verknüpfung, die sich nur auf eine Beschreibung beziehen kann. Sie auf die Wirklichkeit anzuwenden bedeutet, das "folgt auf", das wir beobachten, durch das "folgt aus" zu ersetzen, das wir postulieren; diesen Ersetzungsakt müssen wir aber begründen, und damit sehen wir uns gezwungen, nun unser "folgt aus" durch eine Reihe logischer Schritte zu ersetzen. Somit landen wir zwangsläufig wieder bei einem mathematischen Verfahren, und zuletzt wieder bei der

Tatsache, dass kein solches Verfahren existiert – selbst dann nicht, wenn wir uns vorstellen, wir wären von allen Beschränkungen des Messens und Rechnens befreit.

Die Zukunft folgt also nicht immer aus der Gegenwart. Was ergibt sich daraus?

Die wichtigste Folge ist, dass dadurch ein *logischer Freiraum* entstanden ist: Wenn Anfangsbedingungen und physikalische Gesetze hinreichen würden, um daraus die Zukunft abzuleiten, dann wäre in der Menge der Bedingungen für die Ableitung der Zukunft kein Platz mehr; Da sie aber *nicht* hinreichen, ist in dieser Menge nun Raum für weitere Bedingungen.

Satz:

Die Kausalität von unten ist unvollständig. Es ist Raum für Kausalität von oben.

1.2. Nicht-physikalische Kausalität

Unser nächster Schritt wird sein, zu klären, um welche "weiteren Bedingungen" es sich handeln könnte, von denen die künftige Entwicklung von Systemen abhängt – zusätzlich zu Anfangsbedingungen und physikalischen Gesetzen. Sind es andere Arten von Daten? Oder andere Arten von Gesetzen? Um das zu ermitteln wechseln wir den Schauplatz.

Wir betrachten ein einfaches Gefäß aus Glas. Wenn wir es anschlagen, wird es in Schwingung versetzt und erzeugt einen Ton. Wovon hängt dieser Ton ab? Was bestimmt seine Höhe und seinen Charakter? Die Antwort ist: *Die Form des Gefäßes.* Aus ihr ergibt sich ein mathematisches Gesetz, das uns die Voraussage des Schwingungsmusters des Glases ermöglicht. Hier müssen wir also weder auf die physikalischen Objekte – die Glasmoleküle – noch auf die physikalische Wechselwirkung – den Elektromagnetismus – eingehen, um den Ton vorauszusagen. Die einzige physikalische Information, die benötigt wird, ist die Geschwindigkeit der Schallausbreitung im Glas.

Das Gesetz, das uns nun die Voraussage der Zukunft des Systems erlaubt, ist somit *kein physikalisches Gesetz.* Es gehört zu einer anderen Art von Gesetzen, die ich ***Gesetze der Form*** oder ***Strukturgesetze*** nennen werde.

Vergleichen wir unsere beiden Szenarien, das der gravitierenden Körper und das des schwingenden Gefäßes:

Im Gravitationsszenario sind die Anfangsbedingungen als ***lokale Parameter*** gegeben, als Attribute der einzelnen Körper. Ihre Werte werden in das ***physikalische Gesetz*** – das Gravitationsgesetz – eingesetzt. Obwohl alles, was sich ereignet, vollständig diesem Gesetz entspricht, ist es dennoch unmöglich, die weitere Entwicklung vorauszusagen. Die Zukunft des Systems ***folgt nicht*** aus seiner Gegenwart.

Im Glasszenario sind es nicht die Attribute der Glasmoleküle, die in das Gesetz eingehen, sondern die Abmessungen des Glases, also ***globale Parameter***. Das Gesetz ist kein physikalisches Gesetz, sondern ein ***Strukturgesetz***. Aus den globalen Parametern und dem Gesetz lässt sich die weitere Entwicklung ableiten. Die Zukunft des Systems ***folgt*** aus seiner Gegenwart.

Der Ton, den wir hören, ist weitgehend unabhängig von der Art, wie wir ihn erzeugen. Allerdings gilt das nicht für den ersten Moment: zunächst gibt es einen Einschwingvorgang, der davon abhängt, wie und wo wir das Gefäß anschlagen. Erst danach schwingt es immer im selben Zustand.

Dieser Zustand, auf den das Glas sich schließlich einstellt – das Schwingungsmuster, auf das hin es sich entwickelt und das es danach beibehält –, wird als ***Attraktor*** bezeichnet.

Zuvor hatten wir uns gefragt, welche Arten von Daten und Gesetzen es neben physikalischen Anfangsbedingungen und Gesetzen noch geben könnte. Das einfache Beispiel des schwingenden Gefäßes hat uns eine Antwort geliefert:

1. neue Daten in der Form *globaler Parameter*
2. neue Gesetze in der Gestalt von *Strukturgesetzen,* die auf den globalen Parametern beruhen.

Da sich mittels dieser neuen Daten und Gesetze die Zukunft des Systems voraussagen lässt, sind sie tatsächlich Elemente der "Menge der Bedingungen für die Ableitung der Zukunft", mit der wir uns oben beschäftigt haben.

Am wichtigsten für unsere Überlegungen ist aber zweifellos Folgendes:

Die lokalen Parameter – etwa die Orte und Geschwindigkeiten der Glasmoleküle – hängen zunächst davon ab, wo, womit und wie stark wir das Gefäß anschlagen. Anfangs können also große Unterschiede bestehen. Ungeachtet dieser Unterschiede strebt aber der Zustand des Gefäßes immer auf dasselbe Schwingungsmuster zu – eben den Attraktor.

Beim Glasgefäß gibt es nur ein einziges mögliches Schwingungsmuster, das sich immer ausbildet, unabhängig davon, wie das Gefäß angeschlagen wird. Die künftigen Bewegungen der Bestandteile des Gefäßes – der Glasmoleküle – sind daher durch dieses Muster festgelegt.

Die Kausalität wirkt vom Ganzen auf das Einzelne, vom Gefäß auf seine Bestandteile, und nicht umgekehrt.

Satz:

Eine Form der "Kausalität von oben" tritt dann auf, wenn in einem System *Attraktoren* existieren, d.h. Zustände, auf die hin das System sich zwingend entwickelt, falls es sich zu irgendeinem Zeitpunkt "nahe genug" am Attraktor-Zustand befindet.

(Voraussetzung dafür, dass es sich dabei tatsächlich um "Kausalität von oben" handelt, ist allerdings, dass im betreffenden System die physikalische Kausalität – die "Kausalität von unten" – *unvollständig* ist, genauso, wie wir das im Gravitationsszenario nachgewiesen haben. Da das Glasgefäß aber nur zur Demonstration dienen sollte, worum es geht, brauchen wir uns nicht darum zu kümmern, ob diese Bedingung hier erfüllt ist.)

Damit haben wir nun alle notwendigen Vorbereitungen getroffen, um unser letztes und entscheidendes Szenario in den Blick zu nehmen:

1.3. Das menschliche neuronale Netz

Gegenstand unserer Untersuchung ist die folgende Frage:

Welcher Art von Kausalität gehorcht das neuronale Netz?

Im Netz finden wir drei Ebenen ansteigender Komplexität vor:

1. die physikalische Ebene
2. die neuronale Ebene
3. die geistige Ebene

Bezogen auf diese Einteilung lautet unsere Frage also:

Von welcher Art von Prozessen hängt es ab, was im Netz geschieht? Von physikalischen, von neuronalen oder von geistigen Prozessen? Welche Ebene ist die kausale Ebene? – oder, anders gefragt: Welche Ebene ist dominant?

Zunächst zur *physikalischen Ebene*. Nehmen wir an, wir hätten vollständiges Wissen über die Werte der Attribute aller physikalischen Objekte des Netzes und könnten somit das Gleichungssystem aufstellen, das den Zustand des Netzes und seine weitere Entwicklung repräsentiert. (Natürlich ist diese Vorstellung völlig absurd, aber in der Form eines Gedankenexperiments ist sie zulässig – *im Prinzip* muss dieses Gleichungssystem ja existieren.)

Jetzt sind wir aber wieder mit dem Problem konfrontiert, das schon beim Gravitationsszenario die Berechnung der Entwicklung des Systems verhindert hat: Eine ungeheure Zahl von Prozessen läuft zeitgleich ab, und jeder von ihnen ist mit etlichen anderen direkt vernetzt. Um aber irgendeinen Prozess berechnen zu können, müssen wir zumindest für ein kleines Zeitintervall annehmen, dass seine unmittelbare Umgebung konstant ist – wir müssen ihn also kurzfristig isolieren. Dann können wir für alle anderen Prozesse dasselbe durchführen, und danach wiederholen wir die ganze Prozedur für das nächste Zeitintervall usw.

Wir sind also, wie beim Gravitationsszenario, auf Näherungen angewiesen, die schon nach kurzer Zeit erheblich von der Wirklichkeit abweichen können. Es ist nicht möglich, die Entwicklung des Netzes vorauszusagen. Die Behauptung "Was im Netz geschieht, folgt aus Anfangsbedingungen und physikalischen Gesetzen" ist falsch.

Und auch hier gilt wieder: Die Wirklichkeit tut, wozu wir nicht in der Lage sind: aufgrund ihrer *Aktivität* arbeitet sie zeitgleich die ungeheure Zahl von Prozessen ab, sodass wir den Eindruck gewinnen, alles "folge aus" Anfangsbedingungen und physikalischen Gesetzen.

Satz:

Im neuronalen Netz ist die physikalische Kausalität unvollständig. Es ist Raum für Kausalität von oben.

Betrachten wir nun die *neuronale Ebene*. Sie besteht aus vielen Milliarden Neuronen. Jedes Neuron ist mit hunderten oder sogar tausenden anderer Neuronen direkt verbunden, und über wenige Zwischenschritte sind *alle* Neuronen aneinander gekoppelt. Die neuronale Aktivität wird durch ein Gesetz geregelt, das aus dem neuronalen Input-Output-Mechanismus folgt.[1]

Dieses Gesetz kann als *Wechselwirkungsgesetz der Neuronen* aufgefasst werden. (Es dient auch als Grundlage für Computersimulationen.)

Auch auf dieser Ebene erscheint es uns im ersten Moment wieder völlig selbstverständlich, dass aus den Anfangsbedingungen der Neuronen und ihrem Wechselwirkungsgesetz folgt, was sich im Netz ereignen wird. Und abermals müssen wir erkennen, dass wir wieder derselben Täuschung erlegen sind, indem wir Wirklichkeit und Beschreibung nicht voneinander unterschieden oder miteinander verwechselt haben:

Da ja das neuronale Wechselwirkungsgesetz eine Zusammenfassung physikalischer Sachverhalte ist, bleibt auch das Argument gültig, mit dem wir gerade eben die Behauptung widerlegt haben, dass alles aus Anfangsbedingungen und physikalischen Gesetzen folgt. Für die neuronale Ebene gilt somit: Der hohe Vernetzungsgrad der Neuronen – die permanente Rückkopplung, die sich daraus ergibt – schließt die Existenz eines mathematischen Verfahrens zur Berechnung der weiteren Entwicklung aus.

1 Mit der Bezeichnung "Input-Output-Mechanismus" ist Folgendes gemeint: Die Dendriten jedes Neurons werden über Synapsen durch andere Neuronen stimuliert oder inhibiert. Die auf diese Weise verursachte elektrische Erregung wird zum Zellkörper weitergeleitet und dort aufsummiert. Wenn eine bestimmte Grenze überschritten ist, wird sie an das Axon abgegeben und auf dessen Verzweigungen verteilt, sodass sie schließlich über synaptische Verbindungen weitere Neuronen beeinflusst.

Satz:

Auch die Beschreibung durch neuronale Anfangsbedingungen und das neuronale Wechselwirkungsgesetz lässt Raum für Kausalität von oben.

Damit kommen wir zuletzt zur komplexesten Ebene, der *Ebene des Geistes*. Wir gehen von folgenden Annahmen aus:

1. Jede Art geistiger Aktivität (Gedanken, Assoziationsketten, Bilderfolgen etc.) ist eine Abfolge neuronaler Aktivierungsmuster.
2. Abfolgen neuronaler Aktivierungsmuster können Repräsentationen von Sachverhalten sein.[2]

Betrachten wir die neuronalen Muster. Wie werden sie zu Repräsentationen?

Stellen wir uns ein neuronales Netz vor, in dem es noch keine Repräsentationen gibt. Ein erstmals wahrgenommenes Objekt wird in diesem Netz – ausgehend von der primären Sehrinde – ein bestimmtes Muster verursachen. Die neuronalen Verbindungen, die dabei aktiv sind, werden durch ebendiese Aktivität verstärkt. Dasselbe ist bei jeder Wiederholung der Fall. Auf diese Weise entsteht allmählich eine stabile Verbindung zwischen dem Objekt und einem spezifischen Muster (bzw. einem Ensemble spezifischer Muster).

Außerdem gilt Folgendes: Zwar werden die neuronalen Muster zunächst durch äußere Reize verursacht, aber nach einer hinreichenden Anzahl von Wiederholungen werden sie vom neuronalen Netz auch unabhängig von diesen Reizen hergestellt. Das bedeutet:

Neuronale Muster, die mit Objekten auf die eben beschriebene Weise in Verbindung stehen, sind* Attraktoren *des Netzes. (Siehe dazu auch die Bemerkungen hier und hier.)

Zuvor haben wir festgestellt:

Unter der Voraussetzung, dass die Kausalität von unten unvollständig ist, folgt aus der Existenz von Attraktoren, dass das betreffende System, falls es im Attraktor-Zustand selbst oder diesem Zustand "nahe genug" ist,[3] durch Kausalität von oben bestimmt wird.

Allerdings besteht gemäß unserer ersten Voraussetzung ein geistiger Prozess nicht nur aus neuronalen Mustern, sondern auch aus den Übergängen zwischen diesen Mustern. Für die Übergänge gilt aber dasselbe wie für die Muster selbst: Zunächst werden sie durch die Abfolge bestimmt, in der die verursachenden Objekte erscheinen. Wenn sich diese Reihenfolge wiederholt, dann wird die entsprechende neuronale Aktivität verstärkt, und das hat zur Folge, dass die Muster auch dann, wenn sie vom Netz selbst erzeugt werden, abermals in derselben Reihenfolge auftreten. Ebenso werden auch die räumlichen Beziehungen der Objekte auf die Muster übertragen.

Das bedeutet:

In den Prozessen, die vom Netz selbst erzeugt werden, treten die neuronalen Muster, die mit Objekten fest verbunden sind, in denselben räumlichen und zeitlichen Zusammenhängen auf wie die Objekte selbst. *Somit können die Muster als Repräsentationen der Objekte aufgefasst werden, und die Prozesse als Repräsentationen der Sachverhalte, in denen die Objekte auftreten.*

In menschlichen neuronalen Netzen sind es also nicht die physikalischen oder neuronalen Bedingungen und Gesetze, durch die festgelegt wird, was im Netz geschieht, sondern es ist *die*

2 "Sachverhalt" muss hier im weitest-möglichen Sinn aufgefasst werden.

3 Ohne den Begriff des Phasenraumes lässt sich dieses "nahe genug" nicht wirklich definieren. Das neuronale Netz ist jedenfalls *immer* "nahe genug" an einem Attraktor-Zustand.

Struktur des Netzes – die Tatsache, welche Attraktoren es darin gibt und wie ihre Abfolge geregelt ist –, von der die im Netz ablaufenden Prozesse abhängen.

Die Kausalität wirkt also vom Ganzen auf das Einzelne, vom Netz auf seine Bestandteile, und nicht umgekehrt.

Damit haben wir unser erstes Ziel erreicht:

Satz:

Das neuronale Netz wird durch *Kausalität von oben* geregelt. Die geistige Ebene ist die dominante Ebene. In ihr liegen die Ursachen für die im Netz ablaufenden Prozesse.

Unsere bisherigen Äußerungen waren also tatsächlich Schlussfolgerungen und nicht bloß physikalische Prozesse! Oder – um an die bei der Kritik des Reduktionismus verwendeten Formulierungen anzuschließen: Einsichten sind Einsichten, Gedanken sind Gedanken, der Geist ist in seine Rechte gesetzt, *wir selbst* sind wir selbst...

So weit, so gut, aber damit sind wir noch nicht dort angelangt, wo wir eigentlich hin wollen. Dass wir die Kausalität nach oben verlegt haben, bedeutet noch nicht, dass wir *frei* sind. Wir haben nur die physikalische bzw. die neuronale Kausalität durch die geistige Kausalität ersetzt. Damit haben wir erreicht, dass unser Geist nicht durch physikalische oder neuronale Gesetze beherrscht wird, sondern *durch sein eigenes Gesetz: das Strukturgesetz, dem die Abfolge der neuronalen Muster gehorcht, die etwas repräsentieren.*

Aber bleiben wir damit letztlich nicht doch im Schema von Anfangsbedingungen und Gesetzen gefangen, dem wir entrinnen wollten? Glücklicherweise ist das nicht der Fall. Um das zu zeigen, müssen wir auf den Unterschied zwischen physikalischen und geistigen Gesetzen eingehen.

1.4. Der Unterschied zwischen physikalischen und geistigen Gesetzen

Menschliche neuronale Netze unterscheiden sich stark voneinander, und zwar auch dann, wenn noch keine Strukturierung durch äußere Reize stattgefunden hat. Daraus folgt unmittelbar, dass auch die Muster, die etwas repräsentieren, bei allen Menschen verschieden sind, selbst dann, wenn der repräsentierte Sachverhalt identisch ist.

Die Reihenfolge der Muster wird, wie oben festgestellt, zunächst durch die Reihenfolge bestimmt, in der die Objekte bzw. Sachverhalte auftreten, die die Muster verursachen. Sobald das Netz aber dazu in der Lage ist, diese Muster selbst herzustellen, hängen die Übergangsregeln der Muster – das, was wir als *geistiges Gesetz* bezeichnet haben – in zunehmendem Maß von ihrer Verwendung in inneren Prozessen ab. Diese Abhängigkeit von äußeren und inneren Bedingungen hat zur Folge, dass sich die Übergangsregeln von Mensch zu Mensch unterscheiden.

Somit haben wir schon den ersten Unterschied bestimmt:

*Während physikalische Gesetze **allgemeingültig** sind, sind geistige Gesetze **individuell gültig** – sie gelten jeweils nur für einen einzigen Menschen.*

Verbindungen zwischen Neuronen werden verstärkt, wenn sie aktiv sind,[4] und abgebaut, wenn sie inaktiv sind. Das bedeutet zugleich, dass jede geistige Aktivität die Struktur des Netzes beeinflusst.

4 Diese Erkenntnis geht auf Donald Hebb zurück, der 1949 in *The Organization of Behavior* feststellte: When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased.

Wenn aber die Struktur sich ändern kann, dann können sich offenbar auch die Regeln ändern, die die Abfolge der neuronalen Muster bestimmen.

Also ist dies der zweite Unterschied:

*Physikalische Gesetze sind **unveränderlich**, geistige Gesetze sind **veränderbar**.*

Satz:

Physikalische Gesetze sind allgemeingültig und unveränderlich. Geistige Gesetze sind individuell und veränderbar.

1.5. Die Begründung der Freiheit

Die offensichtlichste Folgerung der Verstärkung aktiver neuronaler Verbindungen ist allerdings, dass das, was wir *immer* denken, fühlen und tun, sich selbst verstärkt. Es ist aber im Grunde selbstverständlich, dass auch das Gegenteil eintreten kann:

Wir haben nachgewiesen, dass die Kausalität in der geistigen Ebene liegt. *Wille* und *Absicht* müssen als Elemente der geistigen Kausalität aufgefasst werden. Stellen wir uns nun konkret vor, wir stünden vor einer wichtigen Entscheidung. Wenn wir in den Entscheidungsprozess eintreten, dann werden wir anfangs durch die bis dahin gültigen Vorgaben – durch unser eigenes geistiges Gesetz – auf bestimmte, bekannte Wege geführt.

Aber wir sind jederzeit dazu imstande, diese Wege zu verlassen, indem wir z.B. einfach das Gegenteil dessen erwägen, was wir bis dahin angenommen haben, oder indem wir einen bisher noch nie erprobten Pfad einschlagen; Dazu sind wir eben deshalb imstande, weil die Ursachen für das, was im Netz geschieht – auch für die Veränderungen der Netzstruktur – in der geistigen Ebene liegen.

Mit anderen Worten:

Das Gesetz, das in unserem Netz die Abfolge der neuronalen Muster bestimmt, die etwas repräsentieren, also unser eigenes geistiges Gesetz, kann durch uns selbst verändert werden: wir selbst können durch unser Denken und Handeln die Gesetze unseres Denkens und Handelns ändern, und zwar *gezielt.*

Das bedeutet zugleich:

Obwohl geistige Prozesse eigenen Regeln unterworfen sind, ist es nicht möglich, daraus eine Willensentscheidung abzuleiten: sie kann in diesen Regeln nicht enthalten sein, weil die Regeln durch den geistigen Prozess, der der Entscheidung vorausgeht, geändert werden können. Während dieser Prozess stattfindet, können sich die Gesetze, denen er gehorcht, ändern – oder genauer: *er selbst* kann die Gesetze ändern, die vor seinem Beginn galten.

Satz:

Willensentscheidungen sind Ursachen von Handlungen. Da erst durch den Entscheidungsprozess selbst bestimmt wird, was geschehen wird, ist die Entscheidung vorher nicht festgelegt.

Sie ist also frei.

Auf die Frage, warum eine (entscheidungsfähige) Person so und nicht anders gehandelt hat, ist demnach nur eine einzige Antwort zulässig:

Weil sie es so wollte.

Bemerkung:

Das heißt selbstverständlich nicht, dass Willensentscheidungen nicht hinsichtlich ihrer neuronalen, chemischen, physikalischen, genetischen, sozialen usw. Ursachen analysiert werden können. Es bedeutet aber, dass diese Analysen unvollständig bleiben und niemals zu einem sicheren Ergebnis führen, weil geistige Phänomene nicht auf andere Schichten der Wirklichkeit reduziert werden können. Der Wille bleibt die letzte Instanz.

Postskriptum

Bei der Durchsicht des Textes schien es mir, als wäre ich meinem Ziel, das Thema so kurz und einfach wie möglich darzustellen, ein wenig zu radikal gefolgt. Deshalb will ich abschließend versuchen, die wichtigsten Punkte meiner Argumentation nochmals zu erläutern:

Nehmen wir an, wir hätten ein System zu beschreiben, das aus einer großen Zahl physikalischer Prozesse besteht, die miteinander verkoppelt sind. Die Gleichungen der Prozesse sind also ebenfalls miteinander vernetzt. Für eine exakte Beschreibung benötigen wir dann *in jedem Augenblick* die Werte aller Parameter aller Prozesse, um sie in die Gleichungen der jeweils anderen Prozesse einzusetzen – mit anderen Worten: es ist (außer in sehr einfachen Fällen) unmöglich, *mit physikalischen Mitteln* über das System, das aus allen diesen Prozessen besteht, genaue Voraussagen zu machen, und zwar aus *prinzipiellen* Gründen, und nicht nur wegen der Einschränkungen des Messens und Rechnens.

Und damit wären wir am Ende unserer Möglichkeiten angelangt – *es sei denn,* die betrachteten Prozesse könnten als Elemente einer "Struktur höherer Ordnung" aufgefasst werden, in der weitere Gesetze gelten. Diese "Gesetze höherer Ordnung" sind dann aber *keine physikalischen Gesetze* mehr*,* und damit haben wir den Bereich der Physik verlassen.

Falls diese neuen Gesetze eine Voraussage über die Entwicklung des Gesamtsystems ermöglichen, dann gilt somit Folgendes:

1. Die Entwicklung des Gesamtsystems ***folgt nicht aus physikalischen Gesetzen***.
2. Die Entwicklung des Gesamtsystems ***folgt aus Gesetzen höherer Ordnung***.

Natürlich geschieht auch weiterhin alles ***in Übereinstimmung*** mit den physikalischen Gesetzen – aber diese Gesetze vollziehen sich nun innerhalb einer ***übergeordneten Struktur***. (Wie beim schwingenden Glasgefäß.)

Die Kausalität ist also nicht mehr *unten* – im elementaren, physikalischen Bereich: sie ist *nach oben* gewandert, in einen Bereich höherer Ordnung, in dem ***neue, nicht-physikalische Gesetzmäßigkeiten*** gelten.

Genau diese Verhältnisse finden wir im neuronalen Netz vor, und zwar mehrfach:

In einem Neuron laufen zahlreiche physikalische Prozesse zeitgleich ab. Die physikalische Betrachtungsweise ermöglicht uns zwar ein Verständnis dessen, was im Neuron vor sich geht, aber die Verkopplung der Prozesse verhindert eine exakte Berechnung der weiteren Entwicklung. Diese Prozesse sind jedoch durch die *Form und Struktur des Neurons* in ein System höherer Ordnung eingebettet, sodass sie einem "Strukturgesetz" gehorchen, das wir zuvor "neuronales Input-Output-Gesetz" genannt haben.

Nun gilt aber wiederum, dass uns auch *dieses* Gesetz keine genaue Voraussage über die künftige Entwicklung von vielen aneinander gekoppelten Neuronen ermöglicht. Die Neuronen sind jedoch selbst wiederum Elemente eines Systems höherer Ordnung – eben des neuronalen Netzes mit seinen

aufgeprägten Mustern (Attraktoren). Damit sind also auch die Neuronen einem neuen Gesetz unterworfen: einem Strukturgesetz abermals höherer Ordnung: dem Gesetz der Abfolge neuronaler Muster, und das heißt: ***dem Gesetz des Geistes***. Somit ist der Geist die *kausale* Ebene. Er bestimmt die im Netz ablaufenden Prozesse – auch diejenigen, die dieses Gesetz selbst verändern.

Zuletzt nochmals der Hinweis auf den Unterschied zwischen *Beschreibung* und *Wirklichkeit*:

Um in der ***Beschreibung*** eines Systems von der Gegenwart in die Zukunft zu gelangen, benötigen wir irgendwelche Verfahren. Das können mathematische Verfahren sein, Algorithmen oder Gleichungen, aber auch Methoden, Sachverhalte so zusammenzufassen, dass sich daraus Schlüsse ziehen lassen. In manchen Fällen gelingt uns das so gut, dass wir behaupten können, B *folge* aus A.

In der ***Wirklichkeit*** ist das alles nicht notwendig. Wenn an jedem Ort zu jeder Zeit geschieht, was zu geschehen hat, dann entsteht die Zukunft *von selbst,* dann entwickeln sich alle komplexen Objekte und Strukturen samt ihren Gesetzmäßigkeiten *von selbst.*

Aber daraus, dass in der Wirklichkeit der Vollzug elementarer Prozesse für die Entstehung der Zukunft hinreicht, kann nicht geschlossen werden, dass die Zukunft aus elementaren Prozessen ***folgt***, denn das würde voraussetzen, das, was in der Wirklichkeit ***von selbst*** geschieht, in eine ***Reihe logischer Schritte*** zu übersetzen, und das ist unmöglich.

Bemerkung:

In dieser Begründung der Willensfreiheit ist es *nicht* notwendig, dass im Weltgeschehen eine "Verzweigung" existiert. Der entscheidende Punkt ist hier, dass die Zukunft nicht in der Gegenwart enthalten ist – dass sie also nicht aus der Gegenwart *folgt,* sondern bloß aus ihr *entsteht,* und dass die Gründe für das, was sich dann tatsächlich ereignen wird, geistiger Art sind.

Bemerkung:

Um Objekte zu erkennen, müssen künstliche neuronale Netze an großen Datensätzen trainiert werden. In zahlreichen Wiederholungen werden die Verbindungsstärken ihrer Neurone so lange variiert, bis eine hinreichend hohe Erkennungsrate erreicht ist.

Wir sind dagegen von folgender Hypothese ausgegangen: Ein wahrgenommenes Objekt, das ein neuronales Aktivierungsmuster verursacht, wird *durch dieses Muster selbst* repräsentiert. Hier wird die Beziehung zwischen Objekt und Repräsentation also nicht erst durch Variation der Verbindungsstärken der Neurone hergestellt, sondern sie besteht von Anfang an und wird nur durch *Verstärkung* der aktiven Verbindungen stabilisiert und präzisiert, wodurch das neuronale Muster zum *Attraktor* wird.

Am deutlichsten wird diese Hypothese durch die sogenannte "Prägung" bestätigt. (Wie z.B. bei den Graugänsen von Konrad Lorenz). Hier gibt es weder "große Datensätze" noch "zahlreiche Wiederholungen" – der Vorgang ereignet sich fast augenblicklich.

Außerdem tritt danach ein *sofortiges Wiedererkennen* auf, trotz der unvermeidlichen Variabilität des Sinneseindrucks, der erkannt werden soll. Durch das Attraktor-Konzept wird diese – ansonsten kaum erklärbare – Leistung zur Selbstverständlichkeit: solange der sinnliche Input im Einzugsbereich des Attraktors liegt, gilt offenbar: *Wahrnehmen = Wiedererkennen*, weil der neuerlich aktivierte Attraktor ja bereits das Objekt darstellt, sodass weitere Berechnungen überflüssig sind.

2. Warum Roboter nichts empfinden

Vorbemerkung

Der Inhalt dieses Abschnitts folgt zum Teil aus den Aussagen des vorangegangenen. Wegen der aktuellen Wichtigkeit des Themas erscheint es mir aber geboten, den Beweis vollständig auszuführen. Ich werde also die dafür erforderlichen Fakten und Argumente hier nochmals (verkürzt) präsentieren.

Ich habe mich für eine zweistufige Ausführung entschieden: für die erste, kurze Version des Beweises ist die Erweiterung der naturwissenschaftlichen Sicht ausreichend, die im Abschnitt über Willensfreiheit vorgestellt worden ist: dort haben wir die geistige Ebene der Wirklichkeit aus der Umklammerung der physikalischen Kausalität befreit, indem wir gezeigt haben, dass die *Aktivität der Wirklichkeit* nicht durch logische oder mathematische Verfahren nachgeahmt werden kann, sodass die Behauptung, alles ***folge aus*** physikalischen Anfangsbedingungen und Gesetzen, nicht aufrechterhalten werden kann. Unter dieser Voraussetzung ist es möglich, geistige Zustände als *selbständige*, *dominante Objekte* zu begreifen, was dadurch konkretisiert wird, dass sie als *Attraktoren* der Dynamik des neuronalen Netzes verstanden werden. Die Abfolgen dieser Zustände – also die geistigen Prozesse – können damit als *kausale Schicht* bestimmt werden, von der diese Dynamik abhängt.

Um den Beweis gegen alle möglichen Widerlegungen abzusichern, ist es allerdings erforderlich, das Szenario genauer zu analysieren und begrifflich neu aufzubauen. Es genügt dann nicht, die Kausalität "nach oben" – in den geistigen Bereich – zu verschieben, sondern es wird dafür der vollständige Begriff der Wirklichkeit benötigt, demzufolge die Wirklichkeit *mehr* ist als eine beschreibbare Zustandsfolge, die (im Prinzip) beliebig genau *reproduzierbar* sein müsste.

2.1. Erste Version (Kurzform) des Beweises

Die Leistungsfähigkeit künstlicher Intelligenz ist in den letzten Jahren eindrucksvoll demonstriert worden. In Szenarien, deren Zustände und Veränderungen vollständig definierbar sind – wie etwa bei den Spielen Schach und Go – sind KI-Systeme inzwischen Menschen weit überlegen. Lernfähige neuronale Netze, die sich nach dem Vorbild der Evolution durch Auswahl der erfolgreichsten Varianten permanent selbst optimieren, erzielen aber auch in Bereichen der wirklichen Welt beachtliche Erfolge.

Es ist also verständlich, dass die Hoffnungen (und Befürchtungen) der KI nun viel weiter gehen: Ist es möglich, ein System zu erschaffen, das menschliche Leistungen nicht bloß in bestimmten Bereichen erreicht oder sogar übertrifft, sondern auch *insgesamt?* Kann ein informations-verarbeitendes System konstruiert werden, das *Bewusstsein* hat?

Jedenfalls scheint der Verwirklichung dieser Vision kein *prinzipielles* Hindernis im Weg zu stehen. Auch das Gehirn selbst ist ja offenbar ein informationsverarbeitendes System. Und das gilt auch für alle Teilstrukturen des Gehirns, auch für diejenigen, die für unsere Gefühle erforderlich sind – sie alle sind nichts anderes als biologische Module, die Information in Form elektrischer Impulse aufnehmen, verarbeiten und an andere Strukturen weiterleiten.

Wenn man also annimmt, dass es genau diese in unserem Gehirn stattfindende Informations-verarbeitung ist, die Geist und Bewusstsein hervorbringt, dann scheint klar zu sein, dass uns von der Schaffung eines Roboters mit Bewusstsein bloß *technische Schwierigkeiten* trennen – wenn auch in einem so ungeheuren Ausmaß, dass es vorläufig ungewiss ist, ob die Konstruktion eines solchen Roboters in absehbarer Zeit möglich sein wird.

Wir werden uns hier die Frage stellen, ob es tatsächlich nur technische Schwierigkeiten sind, die die Erschaffung einer Maschine mit Bewusstsein verhindern bzw. verzögern, oder ob es auch *prinzipielle* Hindernisse gibt – und damit meine ich Hindernisse, die *auf keine Weise* beseitigt werden können.

Nehmen wir an, es wäre uns gelungen, einen Roboter zu konstruieren, der ein künstliches neuronales Netz hat, dessen Struktur der des Gehirns eines menschlichen Kindes entspricht. Dieses neuronale Netz wird über künstliche Sinnesorgane auf dieselbe Art mit Information von der Außenwelt und vom Körper des Roboters versorgt wie bei einem Menschen. In die Funktion, die die Verbindungsstärken der Neurone simuliert, haben wir die Veränderungen implementiert, die sich in natürlichen neuronalen Netzen ereignen, also die Verstärkung durch Aktivität und den Abbau durch Nicht-Aktivität, und auch die Modulation dieser Verbindungsstärken durch chemische Systeme. Damit scheint sichergestellt, dass der Roboter auf dieselbe Art *lernfähig* ist wie ein Mensch: er wird ein *Gedächtnis* haben, er wird *Repräsentationen* bilden, er wird *denken* können usw.[5]

Nennen wir unseren Roboter *Hans.*

Wie wird sich Hans entwickeln? Wird er Gefühle haben? Wird er ein Bewusstsein ausbilden?

Unter den genannten Voraussetzungen erscheint es eigentlich selbstverständlich, dass die Antwort lauten muss: *Ja, das wird er.*

Und doch ist diese Antwort falsch. Wahr ist vielmehr Folgendes:

Selbst wenn Hans die bestmögliche Simulation eines Menschen wäre, würde er nichts fühlen und kein Bewusstsein haben.

Warum ist das so? Der Beweis ist überraschend kurz und einfach.

Wir definieren zunächst ***Simulation****:*

"Simulation" ist die Rekonstruktion der Dynamik eines wirklich existierenden Systems in einem anderen, zu diesem Zweck konstruierten System.[6]

Betrachten wir etwa Simulationen unseres Sonnensystems. In früheren Zeiten waren mechanische Simulationen beliebt, oft wunderschöne Konstruktionen, in denen Kugeln aus Holz oder Messing die Bewegungen der Planeten um die Sonne nachahmten. Heute wird man eher Computersimulationen vorfinden, bei denen geeignete Algorithmen ein Video dieser Bewegungen generieren.

In jedem Fall ist es aber ***nicht Gravitation****,* was die Simulation antreibt – wie das im wirklichen System geschieht. Und es ist unmittelbar einsichtig, dass es auch niemals Gravitation werden kann, gleichgültig, wie weit man die Genauigkeit der Simulation auch steigert. Gravitation als Ursache der Dynamik würde offenbar nur bei einem *Nachbau* des Sonnensystems erhalten bleiben. (Die Repräsentationen der Himmelskörper müssten darin mit den Massen der Originale auftreten!)

Somit gilt:

Im Gegensatz zum "Nachbau" eines Systems wird die Dynamik einer Simulation <u>nicht</u> durch denselben Antrieb verursacht wie die Dynamik des Ausgangssystems.

---

5 Die Voraussetzungen des Gedankenexperiments sind mit Absicht so extrem idealisiert, weil es hier ja ausschließlich um die Frage geht, ob unser Vorhaben nicht selbst dann scheitert, wenn *alle* technischen Probleme gelöst sind. Der Roboter *soll* also eine perfekte Simulation sein. (Dafür ist die Liste seiner Fähigkeiten sogar noch ziemlich unvollständig.)

6 *Dynamik* bezeichnet die Entwicklung des *Zustands* eines Systems; *Zustand* ist die Gesamtheit der Werte der Attribute aller Objekte des Systems zu irgendeinem Zeitpunkt.

Die *Dynamik* eines Systems beruht auf den *kausalen Beziehungen*, durch die die Objekte des Systems miteinander verknüpft sind. Für die Konstruktion einer Simulation ist es daher erforderlich, die *kausale Ebene* des Systems zu bestimmen, das heißt diejenige Ebene, auf der die Prozesse stattfinden, die die Dynamik des Systems verursachen.

Im Sonnensystem ist das trivial, da es hier nur eine einzige "Ebene" gibt: die Objekte sind die Himmelskörper, ihre Bewegungen werden durch Gravitation verursacht.

Im menschlichen neuronalen Netz hingegen finden wir drei Ebenen vor: die physikalisch-chemische, die neuronale und die geistige Ebene. Im ersten Abschnitt ist die *geistige Ebene* als kausale Ebene bestimmt worden. Ich werde kurz die Argumentation wiederholen:

Die physikalische Ebene:

Hier läuft eine ungeheure Zahl von Prozessen gleichzeitig ab, von denen sich viele gegenseitig beeinflussen. Daher existiert *prinzipiell* kein Verfahren, um die künftige Entwicklung des Netzes vorauszusagen. Die Behauptung: "Was sich im Netz ereignet, *folgt aus* physikalischen Anfangsbedingungen und Gesetzen" ist falsch. Dasselbe gilt für die neuronale Ebene.

Die geistige Ebene:

Neuronale Muster, die etwas *repräsentieren* oder etwas *bedeuten*, können vom Netz auch ohne äußere Ursache hergestellt werden. Sie müssen daher als *Attraktoren* des Netzes aufgefasst werden.[7]

Es gilt jedoch Folgendes:

Ein Attraktor determiniert die Dynamik des Systems, falls dessen Zustand im Einzugsbereich des Attraktors liegt.

Der Zustand des neuronalen Netzes eines Menschen liegt *immer* im Einzugsbereich eines Attraktors – das Netz wird sich von jedem beliebigen Zustand aus sofort auf ein Muster einstellen, das etwas bedeutet.

Also lässt sich behaupten:

Im menschlichen neuronalen Netz ist die geistige Ebene die kausale Ebene. Geistige Prozesse bestimmen die Dynamik des Netzes.

Nun müssen wir uns fragen:

Was ist der Antrieb der Dynamik der geistigen Ebene? Was treibt uns an, so zu denken und zu handeln, wie wir es tun?

Die Antwort ist:

Empfindung.[8] **Empfindung ist der Antrieb der Dynamik des Geistes. Information ohne Empfindung ist gleichgültig und daher passiv.**

Da die geistige Ebene die kausale Ebene des neuronalen Netzes ist, folgt daraus:

Empfindung ist der Antrieb der Dynamik des menschlichen neuronalen Netzes.

---

7 *Attraktor* ist ein Systemzustand bzw. eine Abfolge von Systemzuständen – sozusagen ein (statisches oder dynamisches) "Muster", auf das hin das System sich zwingend entwickelt und das es dann für eine gewisse Zeitspanne beibehält.

8 Empfindung muss hier im weitest-möglichen Sinn verstanden werden. Es steht für alles, was an einem geistigen Zustand über Information hinausgeht, also für dasjenige, was nicht *definiert,* sondern nur *gefühlt* und *erlebt* werden kann. (Zwei Beispiele: die Frequenz der Farbe rot kann definiert werden, die Empfindung *rot* aber nicht; die Stärke eines Drucks kann definiert werden, die Empfindung *Schmerz* aber nicht.)

Zuvor haben wir festgestellt, dass genau dasjenige, was in einem wirklich existierenden System die Dynamik des Systems antreibt, *nicht* auf eine Simulation dieses Systems übertragen wird. Wenn wir diese Tatsache nun auf die Simulation eines menschlichen neuronalen Netzes anwenden, dann ergibt sich:

Bei der Simulation eines menschlichen neuronalen Netzes wird die Empfindung nicht mit übertragen. In der Simulation gibt es also keine Empfindung, sondern nur Information.

Und auch hier gilt wiederum, was wir zuvor bei der Simulation des Sonnensystems in Bezug auf Gravitation festgestellt haben: Gleichgültig, wie weit man die Genauigkeit der Simulation auch steigert – was die Dynamik der Simulation antreibt, wird niemals zur Empfindung.

Mit anderen Worten:

Die Simulation – der Roboter – empfindet nichts. Er kann nichts lieben oder hassen, nichts wollen oder nicht-wollen. Unser Roboter Hans ist kein empfindendes Wesen, sondern ein Zombie.

Wenn Empfindung fehlt, dann gibt es auch kein Bewusstsein: Jede Art geistiger Tätigkeit – selbst die abstrakteste – wird von einem Interesse getragen und durch ein Motiv geleitet, und sowohl Interesse als auch Motiv sind Abkömmlinge von Empfindungen, von denen sie nicht getrennt werden können. Es wäre also absurd, einem Roboter ohne Empfindungen Bewusstsein zuzuschreiben.

Damit ist die Frage beantwortet, warum Roboter prinzipiell keine Empfindungen und kein Bewusstsein haben können.

2.2. Ontologische Erweiterung und Absicherung des Beweises

Die soeben durchgeführte Kurzform des Beweises ist zwar vollständig, hat aber eine Schwäche: Da nicht ganz klar ist, *warum* der Beweis funktioniert, könnte der Eindruck entstehen, er würde ein KI-System, dessen Struktur hinreichend ähnlich der Struktur eines menschlichen (oder tierischen) neuronalen Netzes wäre, nicht mit einschließen, falls dieses System durch *Hardware* und nicht bloß durch Software auf einem konventionellen Computer realisiert wäre.

Bei einer mechanisch oder elektrisch angetriebenen Simulation des Sonnensystems *wissen* wir, dass die Bewegungen der Körper *nicht* durch Gravitation verursacht werden, und es ist uns vollkommen selbstverständlich, dass sich der mechanische oder elektrische Antrieb *niemals* in Gravitation verwandeln kann. Warum wissen wir das? Weil wir einen klaren Begriff von der Zusammen-gehörigkeit von Masse und Gravitation haben, oder, um es noch schärfer zu formulieren: von ihrer *Untrennbarkeit.*

Bei der geistigen Aktivität in einem menschlichen neuronalen Netz fehlt hingegen ein vergleichbar selbstverständliches Wissen. Allerdings verfügen wir schon über die dafür erforderlichen Elemente:

Bei der Ableitung der Willensfreiheit haben wir gezeigt, dass die Kausalität nicht im physikalischen Geschehen zu finden ist, sondern in der geistigen Tätigkeit. Damit haben wir den physikalischen Bereich verlassen. Dieses "Verlassen des physikalischen Bereichs" kann sich aber nicht nur auf Kausalität beschränken, es betrifft vielmehr die gesamte Beschreibung des Systems. Die Objekte, die wir nun analysieren, sind also nicht mehr Moleküle oder Neurone, sondern geistige Zustände, und die Prozesse sind keine physikalischen, chemischen oder neuronalen, sondern geistige Prozesse.

Somit ist auch klar, dass genau dasjenige, was die Dynamik der geistigen Tätigkeit verursacht, nicht physikalischer Natur sein kann, sondern geistiger Art sein muss.

In der Kurzform des Beweises haben wir *Empfindung* als dasjenige bestimmt, was die Dynamik der geistigen Tätigkeit antreibt. Aus dieser Sicht hat also Empfindung im Bereich des Geistes denselben Status wie Masse im Sonnensystem, sodass wir mit Sicherheit wissen, dass das, was eine *Simulation* des Geistes antreibt, niemals zu Empfindung werden kann.

Das Problem ist jedoch, dass diese Betrachtungsweise so unüblich oder sogar fremd ist, dass ihr die Selbstverständlichkeit fehlt, die wir im Fall der Gravitation voraussetzen können.

Wenn wir etwa annehmen, dass sich die Abläufe im Sonnensystem und die Abläufe in einer Simulation des Sonnensystems fast vollständig gleichen, dann sind wir dennoch überzeugt, dass die Simulation *nicht durch Gravitation* angetrieben wird. Allgemein ausgedrückt: die angenäherte Identität der Dynamik von Original und Simulation ist für uns keineswegs gleichbedeutend mit der Identität der beiden Systeme selbst.

Hingegen ist genau derselbe Sachverhalt im Fall von Geist und seiner Simulation vollkommen unklar: *Gegenwärtig **weiß niemand**, ob die angenäherte Identität der Dynamik von geistiger Tätigkeit und ihrer Simulation auch bedeutet, dass in der Simulation Empfindung und Bewusstsein auftreten*.

Gemäß unserer Analyse ist das jedoch mit Sicherheit *nicht* der Fall. Aus unserer Sicht ist es also bloß der Mangel eines Begriffs von Geist, der diese Unklarheit verursacht, und deshalb erscheint es angebracht, nun etwas ausführlicher auf die ontologischen Grundlagen dieser Fragestellung einzugehen.

Wir beginnen mit dem Unterschied zwischen Wirklichkeit und Beschreibung, den wir im Abschnitt über Willensfreiheit vorgestellt haben:

Wirklich existierende Objekte sind aktiv, Objekte in einer Beschreibung sind dagegen nicht aktiv. Somit muss zur Existenz wirklicher Objekte etwas gehören, was Objekten in einer Beschreibung fehlt.

Dieses Element der Existenz wirklicher Objekte bezeichnen wir als ***Substanz***. ***Substanz ist also dasjenige, wovon die Aktivität existierender Objekte ausgeht***.

Dasjenige Element der Existenz wirklicher Objekte, das wir wahrnehmen und beschreiben können, ist die *Art ihrer Aktivität,* d.h. ihr Verhalten und ihre Wirkung.

Dieses Element ihrer Existenz bezeichnen wir als ***Akzidenzien***. Naturwissenschaft befasst sich *ausschließlich* mit Akzidenzien.

Es gilt somit folgender

Satz:

Wirklich existierende Objekte bestehen aus Substanz und Akzidenzien, Objekte in einer Beschreibung bestehen dagegen nur aus Akzidenzien.

Da ein Objekt nicht *aufhören* kann, auf die für es charakteristische Weise *aktiv* zu sein, ***bilden Substanz und Akzidenzien eine untrennbare Einheit***. (Die Erde gibt es nur ***mit*** Gravitation.)

Für uns besteht also jedes existierende Objekt aus diesen beiden Elementen: aus ***Substanz*** – das ist jener Teil von Existenz, dessen Vorhandensein wir zwar als notwendig erkennen, der aber *als das, was er eigentlich "ist",* weder vorgestellt noch beschrieben werden kann, und aus ***Akzidenzien*** – das ist der Teil von Existenz, der beschrieben und definiert werden kann.

Im physikalischen Bereich der Wirklichkeit – oder sagen wir besser: im Bereich der Materie – sind uns diese Verhältnisse vertraut. Wir wissen, dass *Masse* Gravitation bewirkt, und dass *elektrische*

Ladung die elektromagnetische Wechselwirkung verursacht. Wir wissen also, *dass da **etwas** sein muss*, was Ursache der Dynamik ist und benennen es, aber wir wissen nicht, was es "ist".

Nun müssen wir bestimmen, was im Bereich des Geistes als Substanz und Akzidenzien aufzufassen ist. Da wir uns hier nicht mehr im physikalischen Bereich befinden, kann nicht einfach die dort gültige Systematik angewendet werden. Vielmehr müssen zunächst die Objekte der geistigen Wirklichkeit definiert werden, und danach muss bestimmt werden was ihre Substanz und ihr Akzidens ist.

Im Abschnitt über Willensfreiheit haben wir festgestellt:

Jeder geistige Zustand ist ein neuronales Aktivierungsmuster. Diese Muster sind Attraktoren der Dynamik des neuronalen Netzes. Jeder geistige Prozess ist eine Abfolge solcher Muster.

Diese Feststellungen betreffen die Frage, wie die Objekte und Prozesse des geistigen Bereichs in Bezug auf ihre *materiellen Voraussetzungen* verstanden werden können.

Jetzt aber ist es unsere Aufgabe, sie als das zu erfassen, was sie *als geistige Phänomene* sind.

Die Antwort ist wie folgt:

Jeder geistige Zustand ist eine Verbindung zweier ungleichartiger Elemente: *Information* und *Empfindung*.

Sein *Informationsgehalt* ist das, was er *repräsentiert* bzw. *bedeutet.*

Für die Bestimmung von *Empfindung* wiederhole ich das in der Kurzform des Beweises Gesagte:

Empfindung steht für alles, was an einem geistigen Zustand ***über Information hinaus*** geht, also für dasjenige, was nicht *definiert,* sondern nur ***gefühlt*** und ***erlebt*** werden kann. (Zwei Beispiele: die Frequenz der Farbe rot kann definiert werden, die Empfindung *rot* aber nicht; die Stärke eines Drucks kann definiert werden, die Empfindung *Schmerz* aber nicht.)

(Ich werde geistige Zustände als ***Qualia*** bezeichnen. Der Ausdruck *Quale* steht also für den ganzen geistigen Zustand und nicht bloß für den Empfindungsteil.)

Mit den obigen Bestimmungen ist zugleich klar, was die Substanz und das Akzidens des geistigen Zustands sind:

Information ist offenbar dasjenige, was sich unserem Denken erschließt – das, was *definiert* und *verarbeitet* werden kann.

Also ist *Informationsverarbeitung* das Akzidens des Quale.

Hingegen ist *Empfindung* dasjenige, was *nicht definiert* werden kann, was sich also unserem Denken und Beschreiben entzieht.

Also ist *Empfindung* die Substanz des Quale.

Daraus folgt, wie wir schon in der Kurzform unseres Beweises festgestellt haben:

***Empfindung* ist der Antrieb der Dynamik des Geistes.**

Nun sind wir ausreichend vorbereitet, unseren Beweis auch formal durchzuführen und dadurch abzusichern.

Zunächst benötigen wir folgende **Definition**:

Als Wesen eines Objekts bezeichnen wir das, was es aufgrund der untrennbaren Einheit seiner Substanz und Akzidenzien ist. Die Aktivität, die sich aus dieser Einheit ergibt, nennen wir wesensgemäß.

(Die ***wesensgemäße Aktivität*** der Erde ist es also, Gravitation auszuüben.)

Der Zweck dieser Definition wird sofort klar, wenn wir uns nun *Simulationen* zuwenden.

Betrachten wir beispielsweise eine mechanische Simulation des Sonnensystems, in der die Modellkörper durch mechanische Vorrichtungen – Ketten, Zahnräder, Wellen usw. – bewegt werden und dadurch die Bewegungen der Himmelskörper nachahmen. Die *wesensgemäße Aktivität* der Modellkörper wäre offenbar, *Gravitation* auszuüben. Aber es ist *nicht die Masse der Modellkörper*, was die Dynamik der Simulation antreibt – was also den gewünschten Ablauf verursacht – sondern *die von uns konstruierte Mechanik,* die dann, elektrisch oder auch mechanisch (etwa durch Drehen einer Kurbel), *aktiviert* werden muss.

Um diesen Sachverhalt auszudrücken, werden wir diese Art der Aktivität als ***zugeführte Aktivität*** bezeichnen, im Gegensatz zur soeben definierten ***wesensgemäßen Aktivität***, die ***von selbst*** geschieht.

Die in der ersten Version des Beweises gegebene Definition einer Simulation verändert sich dadurch auf folgende Weise:

Die Dynamik einer Simulation wird nicht durch die wesensgemäße Aktivität verursacht, die der untrennbaren Einheit von Substanz und Akzidenzien der Objekte der Simulation entspringt, sondern durch zugeführte Aktivität.

Die Akzidenzien, aus denen die Dynamik der Simulation gebildet ist, sind *substanzlos:* die Substanz der Objekte der Simulation *ist nicht die Substanz, die zu diesen Akzidenzien gehört* und mit denen sie eine *untrennbare Einheit* bildet, sondern nur deren *materielle Basis,* von der diese Akzidenzien jederzeit getrennt werden können. (Wie in der mechanischen Simulation des Sonnensystems sofort ersichtlich.)

Der letzte Baustein unseres Beweises ist folgender

Satz:

Solange sich Akzidenzien höherer Komplexität als Funktionen von Akzidenzien geringerer Komplexität beschreiben lassen, bleibt die zugehörige Substanz gleich. Wenn dieser funktionelle Zusammenhang unterbrochen wird, dann ändert sich die Substanz. *Für uns* erscheint sie dann als neue, zweite Substanz.

Bevor wir uns dem Beweis dieses Satzes widmen, müssen wir klären, inwieweit sich die Akzidenzien in komplexeren Ebenen der Realität als Funktionen von Akzidenzien in einfacheren Schichten beschreiben lassen.

Z.B. können die Vorgänge in Neuronen als Funktionen ihrer physikalischen und chemischen Eigenschaften beschrieben werden. (Was allerdings nicht bedeutet, dass sie *berechnet* werden können.) Dasselbe gilt grundsätzlich für alle evolutionären Übergänge: vom physikalischen zum chemischen Bereich, dann zum biochemischen, zellularen, neuronalen, bis hin zum Bereich einfacher neuronaler Netze, die keinen Geist hervorbringen: die in diesen Netzen stattfindenden Prozesse lassen sich als Funktionen ihrer Architektur und äußerer Bedingungen beschreiben.

Erst beim letzten dieser Übergänge – dem Übergang zu neuronalen Netzen, die Geist hervorbringen – endet die Kette der Rückführbarkeit:

Wie wir bei der Begründung der Willensfreiheit festgestellt haben, gilt dann Folgendes:

Die Reihenfolge der neuronalen Aktivierungsmuster wird zunächst durch die Reihenfolge bestimmt, in der die Objekte bzw. Sachverhalte auftreten, die die Muster verursachen. Sobald das neuronale Netz aber dazu in der Lage ist, diese Muster selbst herzustellen, hängen die Übergangsregeln der Muster – das, was wir als *geistiges Gesetz* bezeichnet haben – in zunehmendem Maß von ihrer Verwendung in inneren Prozessen ab.

Das bedeutet, dass sich die Dynamik des neuronalen Netzes – also der Geist – in zunehmenden Maß von den Kausalketten der Umgebung abkoppelt und stattdessen eine eigene, *innere* Gesetzmäßigkeit entwickelt. Und daraus folgt, dass sich der Informationsgehalt – also das Akzidens der geistigen Zustände – nicht mehr als Funktion der Akzidenzien der darunter liegenden Schichten der Wirklichkeit darstellen lässt.

Nun zum Beweis des obigen Satzes: (Die Gesamtheit physikalischer Akzidenzien bezeichnen wir als ***erstes Akzidens***, ihre zugehörige Substanz als ***erste Substanz***, die Gesamtheit geistiger Akzidenzien als ***zweites Akzidens,*** ihre zugehörige Substanz als ***zweite Substanz.***[9])

Soeben haben wir festgestellt, dass sich die Akzidenzien aller evolutionären Ebenen auf Akzidenzien der jeweils darunter liegenden Ebenen zurückführen lassen, mit Ausnahme der Akzidenzien der obersten, also der geistigen Ebene.

Es gilt Folgendes:

Substanz und Akzidens bilden stets eine *untrennbare Einheit.*

Das *erste Akzidens* ist *untrennbar* mit der *ersten Substanz* verbunden.

Wenn komplexe Akzidenzien schrittweise auf jeweils einfachere Akzidenzien reduzierbar sind, dann heißt das, dass sie zuletzt auch auf das erste und einfachste Akzidens zurückgeführt werden können.

Für uns ist *Reduzierbarkeit* jedoch gleichbedeutend mit *ontologischer Identität:* Wenn B auf A reduzierbar ist, dann **ist** B eigentlich A. Wenn also ein komplexes Akzidens auf das erste Akzidens reduzierbar ist, dann **ist** es eigentlich das *erste Akzidens*, und dann ist es untrennbar mit der *ersten Substanz* verbunden.

Solange die Akzidenzien reduzierbar sind, bleibt also die zugehörige Substanz gleich – sie ist dann immer noch *erste Substanz.*

Falls aber die Kette der Reduzierbarkeit auf das erste Akzidens durch das Auftreten eines neuen, *nicht reduzierbaren* Akzidens unterbrochen wird, dann unterscheidet sich dieses neue Akzidens vom ersten Akzidens und von allen anderen, daraus ableitbaren Akzidenzien.

Aufgrund der ***Untrennbarkeit*** von *erster Substanz* und *erstem Akzidens* gilt jedoch:

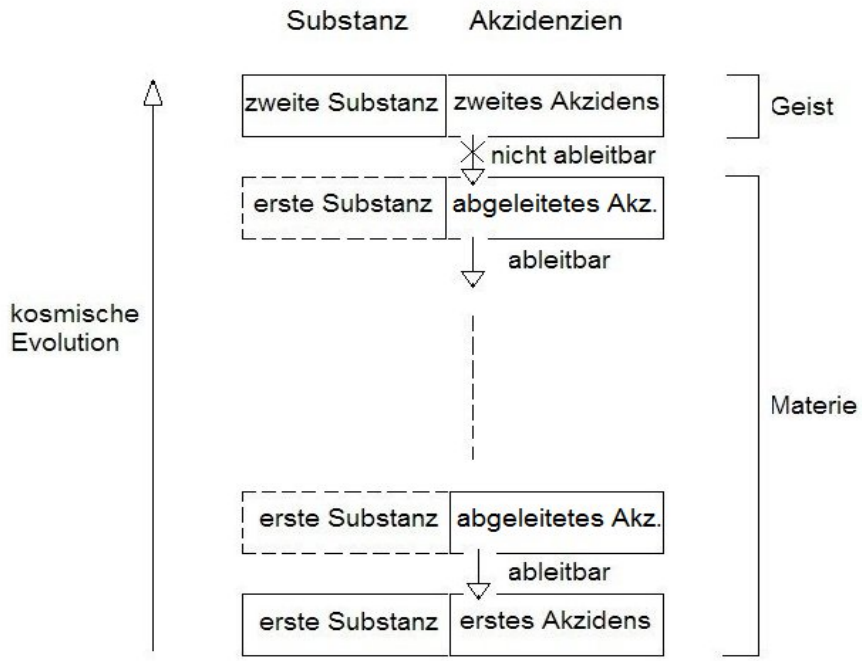
Wenn die Substanz eines Objekts die ***erste Substanz*** *ist, dann muss das zugehörige Akzidens das* ***erste Akzidens*** *sein.*

Und daraus ergibt sich:

Falls ein Akzidens erscheint, das vom ersten Akzidens verschieden ist, dann muss auch die zugehörige Substanz von der ersten Substanz verschieden sein.

Hier eine Skizze zur Veranschaulichung:

9 Das soll aber nicht etwa heißen, dass es nun zwei Substanzen gibt – vielmehr ist die zweite Substanz als aus der ersten Substanz hervorgehend gedacht, und die Frage, die wir uns stellen, lautet demnach: Warum verwandelt sich *für uns* die erste Substanz im Fall der Qualia in die zweite Substanz Empfindung?



Der im Rahmen unserer Argumentation entscheidende Punkt ist, dass die Verwandlung des Wesens des Seienden sich nur dann ereignen kann, wenn die Dynamik des betrachteten Systems sich aus der *untrennbaren Einheit* von Substanz und Akzidenzien ergibt. *Nur dann* folgt aus der Tatsache, dass die Akzidenzien nun nicht mehr auf das erste Akzidens reduzierbar sind, auch die Verwandlung der zugehörigen Substanz.

Wenn dagegen die Dynamik des Systems auf *zugeführter Aktivität* beruht*,* dann sind die Akzidenzien substanzlos, und die zur Existenz der Systemobjekte gehörende Substanz bildet *keine* untrennbare Einheit mit diesen Akzidenzien.

Und das bedeutet: Hier fehlt der Grund dafür, dass sich diese Substanz verwandelt. Sie bleibt *erste Substanz.*

Mit anderen Worten: Das Wesen der Simulation bleibt *physikalisch.* Die Simulation bleibt ein informationsverarbeitendes System ohne Empfindung.

Die Verwandlung von Materie in Geist findet nicht statt.

Die soeben genannte Bedingung, dass die Dynamik des betrachteten Systems sich aus der *untrennbaren Einheit* von Substanz und Akzidenzien ergeben muss, gilt aber nicht nur für die letzte, d.h. für die geistige Ebene – sie muss auf *jeder* Ebene, die beim evolutionären Aufstieg von Materie zu Geist erreicht wird, eingehalten werden. Wenn auf irgendeiner dieser Ebenen die Dynamik des Systems nicht durch die *wesensgemäße Aktivität* der Objekte verursacht wird, sondern durch *zugeführte Aktivität,* dann zerreißt die Einheit von Substanz und Akzidenzien und die Verwandlung des Wesens des Seienden kann sich dann nicht mehr ereignen.

Was bedeutet das nun für unseren Beweis, dass Roboter kein Bewusstsein haben können?

Für KI-Systeme, die durch *Software* auf konventionellen Computern realisiert sind, ist der Beweis ausnahmslos gültig: Der Einsatz von Software ist *immer* mit zugeführter Aktivität verbunden.

Was wäre aber mit einem *Nachbau* eines biologischen neuronalen Netzes, der das neuronale (analog-digitale) Input-Output-Gesetz durch geeignete Hardware reproduziert und dessen Struktur der Struktur des gesamten Netzes entspricht, sodass angenommen werden könnte, dass die

Zustandsfolge des *konstruierten* Systems weitgehend der Zustandsfolge des *biologischen* Systems gleichen würde? Könnte hier die Verwandlung in Empfindung stattfinden?

Die Antwort ist eindeutig ***nein***. Die Bedingung für die Verwandlung ist nicht erfüllt: Die Dynamik des Nachbaus ist *nicht* durch *wesensgemäße,* sondern durch *zugeführte* Aktivität verursacht.

Das Problem besteht darin, dass von der üblichen naturwissenschaftlichen Sicht der Wirklichkeit her diese Tatsache ***überhaupt nicht verstanden werden kann***.

In dieser Sichtweise wird die Wirklichkeit ja mit einer (beschreibbaren) Zustandsfolge ***gleichgesetzt,*** und es muss demnach erwartet werden, dass die zunehmende Annäherung zweier Zustandsfolgen schließlich zur *Identität der Systeme selbst* führt.

In der erweiterten materialistischen Sicht, die wir hier präsentiert haben, wird der Begriff der Existenz jedoch um ein Element erweitert, das aus dem Bereich des Beschreibbaren hinausführt.

Das bedeutet, dass alle unsere Beschreibungen und Vorstellungen von den Vorgängen in der Natur notwendig *unvollständig* sind. Sozusagen "hinter den Kulissen" des uns zugänglichen Teils der Bühne geschieht etwas, was sich uns entweder vollständig verschließt oder lediglich durch Rückschluss vom uns zugänglichen Teil der Wirklichkeit – den Akzidenzien – erkennbar und verstehbar wird. Die Wirklichkeit ist hier also *mehr* als eine Zustandsfolge.

Im Rahmen unserer Überlegungen bedeutet das:

Aus der angenäherten Identität der Zustandsfolgen des natürlichen und des künstlichen Systems kann nicht auf deren annähernd bestehende Wesensgleichheit geschlossen werden.

Konkret: Die Substanz der zwei Systeme kann trotz der weitgehenden Identität ihrer Zustände durchaus verschieden sein:[10]

Im ***biologischen System*** ist sie die mit den Akzidenzien des Systems ***untrennbar verbundene*** Substanz und verwandelt sich deshalb in die ***geistige Substanz Empfindung***.

Das ***konstruierte System*** wird jedoch mit ***zugeführter Aktivität*** angetrieben, und daher steht die Substanz hier mit den Akzidenzien des Systems in einer nur konstruierten und keineswegs untrennbaren Verbindung, sodass sie ***physikalische Substanz*** bleibt und sich ***nicht in Empfindung umwandelt***.

Das Resultat unserer Schlussfolgerungen ist folgender

Satz:

Es ist nicht möglich, einen Roboter zu konstruieren, der Empfindungen erlebt und Bewusstsein hat. Weder in einer Simulation noch im Nachbau eines Systems, das Geist hervorbringt, kann die Umwandlung von Materie in Geist stattfinden.

Es gibt keinen Geist in der Maschine.

Es kann also nur *künstliche Intelligenz* konstruiert werden und nicht *künstlicher Geist.*

Bedeutet das, dass es überhaupt unmöglich ist, künstlichen Geist zu erschaffen?

Nein. Unsere Argumentation schließt nur aus, dass Geist *konstruiert* werden kann. Die Definition des Begriffs *Nachbau* lässt sich aber dahingehend erweitern, dass sie eine *künstliche Evolution* einschließt, d.h. eine Evolution, die von uns geplant und gesteuert ist. In diesem Fall wäre – ebenso wie bei der natürlichen Evolution – die Bedingung erfüllt, dass die jeweilige System-Aktivität immer *wesensgemäß* ist. Wenn wir an keiner Stelle dieses künstlichen Evolutionsprozesses durch

10 Statt "Identität der Zustände" kann hier selbstverständlich auch (die wesentlich schwächere Bedingung) "Identität des Outputs" der zwei Systeme gesetzt werden, was das Kriterium beim Turing-Test ist. Der Turing-Test ist also nicht dafür geeignet, festzustellen, ob KI-Systeme empfindungsfähig sind.

Konstruktionen eingreifen und Aktivität zuführen, sondern uns darauf beschränken, die Entwicklung zu steuern und zu beschleunigen, dann *könnte* am Ende dieser Evolution ein System stehen, das Geist hervorbringt.

Niemand kann allerdings wissen, ob eine solche künstliche Evolution überhaupt möglich ist, oder ob der Weg, den die Natur gewählt hat, der einzig gangbare ist.

In jedem Fall ist aber klar, dass die Schaffung künstlichen Geistes einer sehr fernen, vielleicht niemals erreichbaren Zukunft vorbehalten bleibt, wenn sie nicht sogar unmöglich ist.

Bemerkung:

*Alles, was **definiert** werden kann*, ist durch Informationsverarbeitung erreichbar, *alles, was **nicht definiert** werden kann*, ist für Informationsverarbeitung *prinzipiell* unerreichbar: gleichgültig, welche Funktion man auf Information anwendet – das Ergebnis ist immer bloß Information und sonst nichts; die Information "rot" wird niemals zur Empfindung *rot*, die Information "Druck" wird niemals zur Empfindung *Schmerz*. Deshalb bilden "Information" und "Empfindung" (in der oben festgelegten Bedeutung) das *einzige* Begriffspaar, das es ermöglicht, zwischen künstlicher Intelligenz und menschlichem Geist eine klare und eindeutige Grenze zu ziehen und dafür eine Begründung zu liefern.

Daraus folgt, dass der häufig im Mittelpunkt der Diskussion stehende Begriff "Bewusstsein" nur dann für diese Grenzziehung geeignet ist, wenn die geistigen Phänomene, die ihm (in seiner jeweiligen Definition) zugeschrieben werden, gemäß ihrer Zugehörigkeit zu *Information* oder *Empfindung* analysiert und eingeteilt werden: der zur Informationsverarbeitung gehörende Teil des Bewusstseins (z.B. jede Art von Selbst-Repräsentation) ist reproduzierbar – gleichgültig, welche technischen Schwierigkeiten seiner Simulation auch im Weg stehen, während der zur Empfindung gehörende Teil für KI unzugänglich bleibt.

Es wäre also eine unnötige und überdies auf Abwege führende Komplikation, den Unterschied zwischen KI und Geist auf den Begriff "Bewusstsein" zu gründen.

Bemerkung:

Wie am Ende der Vorbemerkung festgestellt, ist es für unseren Beweis nicht in jedem Fall hinreichend, die Kausalität "nach oben" zu verschieben. Der Grund dafür ist wie folgt:

Nehmen wir an, es könnte ein neuronales Netz konstruiert werden, das geeignet ist, Attraktoren auszubilden und zu vernetzen – so, wie wir das bei menschlichen neuronalen Netzen voraussetzen,[11] und nehmen wir außerdem an, dieses Attraktor-Netzwerk sei die *kausale Ebene* des Systems. Dennoch bliebe das System *empfindungslos*: die Bedingung, dass seine Dynamik auf ***wesensgemäßer Aktivität*** beruht – dass sie also aus der *untrennbaren Einheit seiner Substanz und Akzidenzien* hervorgeht – wäre nicht erfüllt.

Bemerkung:

Zur Hypothese, dass Objekte durch Attraktoren repräsentiert werden, ist Folgendes zu ergänzen:

Das Muster, das sich als Folge eines wahrgenommenen Objekts in der primären Sehrinde ausbildet, wird nicht als Ganzes direkt ins neuronale Netz übertragen. Vielmehr wird es in etliche Komponenten zerlegt – in diesem Sinn also *parametrisiert* – die erst am Ende des Verarbeitungsprozesses zu dem neuronalen Gesamtmuster zusammengefügt werden, das wir als Attraktor auffassen.

11 Die gegenwärtig populären künstlichen neuronalen Netze (z.B. GPTs) sind dafür ungeeignet.

Diese Parametrisierung ist ein wichtiger Aspekt der Attraktor-Hypothese: Der Attraktor ist durch eine Untermenge des Phasenraums definiert. Der *Attraktor-Zustand* des Systems entspricht einer Trajektorie, die diese Untermenge für eine gewisse Zeitspanne nicht verlässt. Für seine Wiederherstellung ist aber schon eine (kleine) Teilmenge der entsprechenden Parameterwerte ausreichend, die überdies nicht einmal besonders genau sein müssen. Zur Wiedererkennung genügt also ein Bruchteil des ursprünglichen, vollständigen Sinneseindrucks. Dadurch wird das Erkennen von Objekten extrem erleichtert und zugleich die Fähigkeit zur Verallgemeinerung von Objekten und Sachverhalten gesteigert.

Bemerkung:

Zuletzt noch ein Kommentar zum Szenario der gravitierenden Körper aus dem Abschnitt über Willensfreiheit:

Sogar ein Laplacescher Dämon mit unendlichen Ressourcen an Raum, Zeit und Information würde an der Berechnung scheitern: Um die Zukunft des Systems *exakt* zu ermitteln, muss der Dämon die Berechnung für unendlich kleine aufeinander folgende Zeitintervalle durchführen. Falls die Intervallgrenzen genauso dicht liegen wie die *reellen* Zahlen, wird er sogar in unendlich langer Zeit nicht fertig, falls sie aber weniger dicht liegen (wie z.B. die *rationalen* Zahlen), wird es geschehen, dass ihm eine Instabilität entgeht, die *zwischen* zwei Zeitpunkten seiner Berechnungen liegt.

Tatsächlich haben wir aber auch mit dieser Argumentation noch immer nicht das ganze Ausmaß des Problems erfasst: Wir haben ja angenommen, dass wir aufgrund der vollständigen Kenntnis der Anfangsbedingungen auch das Gravitationsfeld kennen. Diese Annahme ist jedoch falsch, und zwar aus folgendem Grund:

Bezeichnen wir den Zeitpunkt, zu dem wir genaue Kenntnis der Anfangsbedingungen besitzen und an dem unsere Berechnung beginnen soll, mit t_0. Wenn wir für irgendeinen der Körper, sagen wir: den Körper A, berechnen wollen, wohin er sich im ersten Zeitintervall bewegt, dann müssen wir sämtliche Wirkungen kennen, denen A zur Zeit t_0 vonseiten der anderen Körper ausgesetzt ist.

Betrachten wir z.B. den Körper B: wir kennen den Ort, an dem er sich zur Zeit t_0 befindet. Die von B stammende Wirkung, der A zur Zeit t_0 ausgesetzt ist, geht jedoch ***nicht*** von ***diesem*** Ort aus, sondern von einem Ort, an dem sich B ***vorher*** befand – und zwar genau *so lange* vorher, wie die Gravitation benötigte, um ***von dort aus*** den Körper A zur Zeit t_0 zu erreichen. Um die Wirkung von B auf A zur Zeit t_0 zu ermitteln, müssen wir daher B auf seiner Bahn *in die Vergangenheit versetzen*, und genau dasselbe gilt auch für alle anderen Körper: sie alle müssen in die Vergangenheit versetzt werden – umso weiter, je weiter sie von A entfernt sind.

Bevor wir überhaupt damit ***beginnen*** können, die Bahn von A zu ermitteln, müssen wir also zunächst die Bahnen aller anderen Körper bestimmen. Dafür ist es aber erforderlich, auch die Wirkung zu kennen, die A zur Zeit t_0 auf die anderen Körper ausübt, und deshalb müssen wir auch A selbst auf seiner Bahn in die Vergangenheit versetzen, d.h. auf der Bahn, die uns nicht bekannt ist, weil wir sie ja soeben erst berechnen wollten!

Dasselbe gilt für *jeden* Körper: um ihn in die Vergangenheit zu versetzen, müssen wir die Bahnen *aller anderen* Körper kennen. Da uns aber *keine einzige* dieser Bahnen bekannt ist, ist es unmöglich, die genauen Positionen zu bestimmen, wo sich die Körper vorher befanden, und damit ist es auch unmöglich, die Wirkungen zu ermitteln, denen sie zum Zeitpunkt t_0 ausgesetzt sind.

Mit anderen Worten: Wir – und mit "wir" meine ich uns alle ***und*** den Laplaceschen Dämon – sind nicht nur außerstande, eine ***exakte Berechnung*** der Zukunft ***auszuführen***, wir sind nicht einmal in der Lage, damit auch nur ***zu beginnen***.

Das Szenario ist nicht berechenbar. Die *Wirklichkeit* ist nicht berechenbar. *Wir selbst* sind nicht berechenbar.

Die formale Version unserer ontologischen Argumentation zur Willensfreiheit lautet also wie folgt:

Das Verhalten aller elementaren Objekte wird ausschließlich von physikalischen Gesetzen bestimmt. Versucht man aber, die Zukunft (oder, falls objektiver Zufall einbezogen werden soll: *irgendeine* Version der Zukunft) auf physikalische Weise abzuleiten, dann scheitert das an der Tatsache, dass dafür überabzählbar viele logische Operationen erforderlich wären.

In manchen Fällen kann jedoch die überabzählbare Menge logischer Operationen durch eine endliche Menge von Aussagen über eine höhere, *nicht-physikalische* Schicht der Wirklichkeit ersetzt werden. Die Fakten, auf die sich diese Aussagen beziehen, können dann als Ursachen (oder *Gründe*) für den künftigen Zustand aufgefasst werden.

3. Was folgt daraus für die KI?

3.1. Einleitung

Üblicherweise wird gefragt:

"Wieso gibt es im Geist etwas Undefinierbares, wie 'Farbe' oder 'Schmerz', und sonst nirgends?"

Wir haben uns stattdessen die Frage gestellt:

"Wieso ändert das Undefinierbare, das es überall in der Wirklichkeit gibt, seinen Charakter, wenn es im Geist auftritt?"

Es wird also nicht nach dem Grund der *Existenz* dieses Undefinierbaren gefragt – was überflüssig wäre, weil seine Existenz *selbstverständlich* ist[12] – sondern nach dem Grund seiner *Veränderung*.

In der ersten Version kann die Frage nicht beantwortet werden. In dieser (falschen) Form führt sie zu seltsamen Hypothesen, wie Qualia-Eliminativismus, oder Panpsychismus.

Wie wir gezeigt haben, lässt sich die Frage in der zweiten Version aber beantworten, und diese Antwort enthält überdies den Beweis, dass ***Empfindung –*** die *geistige Erscheinungsform* dieses "Undefinierbaren" – in Systemen, die ***nicht durch Evolution entstanden***, sondern ***von uns konstruiert*** sind, ***nicht existiert***.

KI-Systeme sind also nicht empfindungsfähig. Dadurch wird den Erwartungen, Hoffnungen und Ängsten der KI-Ingenieure eine prinzipielle Grenze gesetzt. Vorläufig ist aber noch nicht klar, was dieser Beweis für die möglichen Leistungen von KI-Systemen bedeutet. In diesem dritten Teil der Arbeit werden wir uns deshalb mit der Frage beschäftigen, welchen Einschränkungen künstliche Intelligenz aufgrund ihrer Empfindungslosigkeit grundsätzlich unterworfen ist.

3.2. Was ist *"Empfindung"*?

Zunächst ein kurzer Kommentar, warum ich den Begriff "Empfindung" abweichend von seinem üblichen Gebrauch bestimmt habe:

Jeder geistige Zustand enthält etwas, was ***nicht definierbar*** ist, was also ***über Information hinaus*** geht. Da es aber keinen Begriff gibt, dem alle dafür in Frage kommenden Elemente geistiger Zustände zugeordnet werden können – und weil ich einen langen Katalog vermeiden wollte – habe ich stattdessen die Bezeichnung gewählt, die diesem fehlenden Begriff am nächsten kommt: ***Empfindung***. Der Begriff "Empfindung" wird hier also gegenüber seinem üblichen Gebrauch einerseits eingeschränkt (weil er ja *keine Information*, d.h. keinen *definierbaren* Teil enthalten soll), andererseits aber auch wesentlich erweitert.

Zur Illustration dienten zwei Beispiele: *Farbe* und *Schmerz.* Farbe, weil die Undefinierbarkeit der Farbempfindung ein bekanntes Faktum ist, und Schmerz, weil vollkommen einsichtig ist, dass das Ereignis "Hammerschlag auf Finger" einen geistigen Zustand auslöst, der nicht nur die Information "Hammerkopf hat Kontakt mit Finger" enthält, sondern etwas *darüber hinaus gehendes:* die Empfindung *Schmerz*, die so stark sein kann, dass es unmöglich ist, ihr Auftreten zu bestreiten.

Die auf diese Weise verstandene *Empfindung* lässt sich in drei Bereiche unterteilen:

12 Siehe Teil 2, Seite 17 Mitte.

A) Der erste Bereich ist der Bereich der *Wahrnehmung*:

Empfindung umfasst das ganze "innere Theater": den virtuellen Raum, die Bühne, auf der wir agieren, die uns immer als Ganzes – als "Bild" – präsent ist und auf der wir sehen, hören, fühlen, riechen und schmecken.

Während es bei der Empfindung *Farbe* kaum zu bezweifeln ist, dass sie nicht definiert werden kann, mag es zunächst so scheinen, als würden wir in den Bereich des Definierbaren zurückkehren, falls unser Wahrnehmungsbild *farblos* ist: *Grauwerte* sind doch definierbar? – Ja, das sind sie, aber die damit verbundene *Empfindung* ist es nicht: definierbar ist bloß die Intensität des Lichts, und ebenso der neuronale Erregungszustand, der daraus folgt. Doch beim Übergang zur *Wahrnehmung* verlassen wir den Bereich der Information: die *Helligkeit*, die wir *wahrnehmen*, ist ebenso eine *Empfindung* wie *Farbe.*

Und dasselbe gilt auch für alle anderen Sinne: die Frequenz eines Tons ist definierbar, aber die Ton-*Empfindung* ist es nicht, usw.

Das bedeutet: ***Wenn Empfindung fehlt, dann gibt es kein "inneres Theater", das ja aus Empfindungen aufgebaut ist.***

Um es also in aller Deutlichkeit auszusprechen:

KI-Systeme sehen nicht, hören nicht, fühlen nicht, riechen nicht, schmecken nicht. [13]

Leider ist unser Sprachgebrauch für die Unterscheidung zwischen Systemzuständen *mit* Empfindung und solchen *ohne* Empfindung nicht geeignet. Für uns bedeutet "sehen" oder "hören" einfach das, was es für uns ***ist***, und das ist in jedem Fall Information ***und*** Empfindung. Deshalb sind Aussagen über Wahrnehmungen *genau genommen* nur dann korrekt, wenn sie sich auf Menschen oder höhere Tiere beziehen, ansonsten sind sie falsch: Roboter *sehen* nicht, Bienen *sehen* nicht – sie verarbeiten nur Frequenz-, Intensitäts-, Entfernungs- und Richtungsinformationen.

B) Der zweite Bereich ist der Bereich der *Gefühle und Stimmungen*. Dazu muss nichts weiter erklärt werden.

KI-Systeme erleben nichts und fühlen nichts. Sie empfinden weder Glück noch Unglück, weder Liebe noch Hass. Sie sind weder heiter noch betrübt, weder gut aufgelegt noch gereizt.

Diese Liste lässt sich nach Belieben fortsetzen, da ja *jeder* geistige Zustand ein *Quale* ist, d.h. nicht nur aus *Information*, sondern auch aus *Empfindung* besteht.[14]

C) Wir haben *Empfindung* als *Substanz* des geistigen Zustands bestimmt. Daraus folgt, dass sie als *Ursache* der geistigen Dynamik aufgefasst werden muss.

---

13 Das gilt auch für einfache Tiere, wie etwa Insekten, und zwar aus folgendem Grund: Wir haben gezeigt, dass die Entstehung von Empfindung nur dann stattfinden kann, wenn das neuronale Netz eine eigene, *innere* Gesetzmäßigkeit entwickelt. Eine notwendige (und hinreichende) Bedingung dafür ist aber, dass das Netz *funktionell ungebundene* Strukturen enthält, d.h. Strukturen, deren Funktion nicht genetisch oder durch frühe Programmierung festgelegt ist. Nur unter dieser Voraussetzung kann (und wird) sich das *Netzwerk aus neuronalen Zuständen* (Attraktoren) ausbilden, das wir als *Geist* auffassen.

Der *Besitz von Augen* ist für uns gleichbedeutend mit der *Fähigkeit zu sehen.* Das ist jedoch falsch. Für ein Tier, das eine lichtempfindliche Zelle besitzt, ist die Welt keineswegs *hell* – das Tier verfügt lediglich über die *Information*, aus welcher Richtung das Licht kommt.

14 Natürlich gibt es auch Aktivitäten *ohne* Empfindung, wie Reflexhandlungen oder automatisch ausgeführte Abfolgen von Bewegungen. Das sind dann aber keine *geistigen,* sondern *neuronale* Aktivitäten.

Demnach muss ***alles***, was unser Denken und Handeln antreibt, einen *Empfindungsanteil* besitzen. Es gibt kein Handeln oder Denken ohne ein Motiv. Selbst rein logisches Schlussfolgern kann nur stattfinden, wenn wir die korrekte Lösung finden ***wollen***.

Umgekehrt gilt somit:

KI-Systeme können nichts wollen oder nicht-wollen. Sie kennen weder Motiv noch Interesse, weder Neugier noch Ablehnung.

In diesem Bereich ist der Mangel an Differenziertheit des Sprachgebrauchs besonders problematisch. Programmierer sprechen von "Zielen" oder "Absichten" eines KI-Systems, von dem, was es "anstrebt". Es handelt sich dabei aber in allen Fällen nur um die Steigerung eines Parameterwertes und nicht um *Ziele* oder *Absichten,* wie wir sie als Elemente menschlichen Handelns verstehen, die immer mit Emotionen verknüpft sind.

Nach dieser kurzen Vorbereitung wollen wir uns jetzt der Frage zuwenden:

Was bedeutet die Abwesenheit von Empfindung für die Leistungen von KI-Systemen?

3.3. Was mit Sicherheit auszuschließen ist

Alan Turing, aus einer Vorlesung von 1951:

"It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage therefore we should have to expect the machines to take control."

Geoffrey Hinton, October 27, 2023, University of Toronto:

"Suppose you have multiple different super-intelligences. ... you're gonna get evolution of super-intelligences. And let's suppose there's a lot of benign super-intelligences who are all out there just to help people. ... But let's suppose that one of them just has a very, very slight tendency to want to be a little bit better than the other ones, just a little bit better. You're gonna get an evolutionary race, and I don't think that's gonna be good for us. So I wish I was wrong about this. ... My guess is that they will take over, they'll keep us around to keep the power stations running, but not for long. ... That's my best guess, and I hope I'm wrong."

Geoffrey Hinton, February 19, 2024, Oxford's annual Romanes Lecture at the Sheldonian Theatre:

"… what happens if super-intelligences compete with each other? ... As soon as they get any sense of self-preservation, then you'll get evolution occurring. ... the more aggressive ones will win. And then you get all the problems that Chimpanzees like us have: lots of aggression and competition." [15]

Diese Zitate drücken die von vielen KI-Experten geteilte Erwartung aus, was geschehen könnte – oder wahrscheinlich sogar geschehen *wird* – wenn KI zu AGI (Artificial General Intelligence) wird, wenn sie also nicht nur in *bestimmten* Bereichen menschliche Leistungen erreicht oder übertrifft, sondern in *allen*. In der Folge würde es zu einer exponentiellen Leistungssteigerung durch Selbst-Optimierung kommen, sodass KI-Systeme aufgrund ihrer überlegenen Intelligenz uns als dominante

15 Es waren diese – und zahlreiche andere vergleichbare – Aussagen, die mich veranlasst haben, meiner Arbeit über Willensfreiheit und künstliche Intelligenz einen weiteren Teil hinzuzufügen. Es wäre mir schwer gefallen, solche fundamental falschen Behauptungen über unsere Zukunft unwidersprochen zu lassen.

Art ablösen. Ebenso, wie gegenwärtig Menschenaffen *unserer* Willkür ausgeliefert sind, würden in Zukunft also *wir selbst* vom guten Willen und der Gnade der KI-Systeme abhängig sein.

Wie wir wissen, ist aber zu unserem Glück Geoffrey Hintons Wunsch ("I wish I was wrong about this") schon in Erfüllung gegangen, bevor er ihn überhaupt ausgesprochen hat: Gemäß unserem Beweis, dass KI-Systeme keine Empfindungen haben, werden wir keineswegs eine uns überlegene Spezies erzeugen, sondern nur ***gefühllose, willenlose Zombies*** – bloß ***Automaten, die nicht einmal imstande sind, etwas wahrzunehmen***.

Sie werden also keinesfalls "die Kontrolle übernehmen", weil sie das gar nicht ***wollen*** können, sie werden uns weder mögen noch dulden, weder verachten noch vernichten, ja es wäre sogar unangemessen zu behaupten, wir wären ihnen gleichgültig – da ist einfach ***gar nichts***.

Zusammengefasst:

KI-Systeme sind keine neue, super-intelligente, dominante Art. Sie sind keine lebenden Wesen, sondern Automaten.[16]

Aber auch von KI-Systemen *ohne* Empfindung und Bewusstsein können Gefahren ausgehen. Hören wir *Stuart Russell*, Dezember 14, 2023, Penguin channel:

"… that's the key in the Skynet story: It becomes self-aware. That's a very common idea in science fiction, both in books and particularly in film. In AI-related films there's gotta be a struggle between AI and humanity. And the way that struggle happens is almost always because the AI becomes conscious. We call it the Spooky Emergent Consciousness meme. It's really a red herring. Because of its frequent occurrence in film, one often sees this in serious journalism as well, but in fact we need to worry about machines not because they're conscious, but because they're competent. They may take preemptive action to ensure that they can achieve the objective that we gave them. That's the real concern. So if someone tells you, 'Don't worry: as long as it doesn't become conscious, everything's fine', don't be reassured."

Bis auf den Ausdruck: "preemptive action", der Planung und Absicht unterstellt, wird diese Behauptung Stuart Russells durch unseren Beweis gegen KI-Empfindung nicht in Frage gestellt – jedenfalls nicht direkt. Im nächsten Abschnitt werden wir aber zeigen, dass der Beweis Argumente enthält, die der möglichen Kompetenz künftiger KI-Systeme Grenzen setzen.

3.4. Welche Einschränkungen wahrscheinlich sind

Wenn man von unserem Beweis ausgeht, dass KI-Systeme nichts empfinden können, dann ist es selbstverständlich, dass die populäre Dystopie, in der wir als unterlegene Art auf Gedeih und Verderb einer unvorstellbar mächtigen Superintelligenz ausgeliefert sind, schlichtweg Unsinn ist. Nun müssen wir uns aber mit der ungleich schwierigeren Frage auseinandersetzen, inwiefern das Fehlen von Empfindung die Leistungsfähigkeit von KI-Systemen einschränkt.

Zunächst lässt sich Folgendes feststellen:

Es besteht ein auffälliger Zusammenhang zwischen der Tatsache, die als *"Moravec's Paradox"* bekannt ist, und der Tatsache*, dass KI-Systeme nichts wahrnehmen.*

16 Bei *Lebewesen* – die ihre Existenz der biologischen Evolution verdanken – ist Empfindungslosigkeit an die Bedingung geknüpft, dass das neuronale Netz einfach und nur sehr eingeschränkt lernfähig ist. Um an die Ausführungen im zweiten Teil anzuschließen: *Lebendig* ist ein System nur dann, wenn seine *Aktivität* aus der *untrennbaren Einheit von Substanz und Akzidenzien* folgt, also *wesensgemäß* ist und *von selbst* geschieht. Daraus folgt, dass auch *Leben* – ebenso wie Geist – *nicht konstruierbar* ist.

1988 schrieb Hans Moravec:

"Es ist vergleichsweise einfach, Computer dazu zu bringen, Leistungen auf Erwachsenenniveau bei Intelligenztests oder beim Dame spielen zu erbringen, und schwierig oder unmöglich, ihnen die Fähigkeiten eines Einjährigen in Bezug auf Wahrnehmung und Mobilität zu vermitteln." [17]

(Für eine Aktualisierung des Paradoxons sollte "Dame" durch "Go" und der "Einjährige" durch einen "Fünfjährigen" ersetzt werden.)

In den vorangegangenen Abschnitten haben wir festgestellt, dass es in Robotern kein *"inneres Theater"* gibt: ***Sie sehen nichts*** – in dem Sinn, dass sie kein "Bild" vor sich haben.

Es drängt sich also die Vermutung auf*,* dass die sensomotorischen Schwierigkeiten der KI auf ihre Unfähigkeit zurückzuführen sind, die Umgebung auf die Weise *wahrzunehmen*, die für uns so selbstverständlich ist.

Warum ist das so? Warum sollte sich *Information ohne Empfindung* nicht genauso als Grundlage dafür eignen, mit Gegenständen zu hantieren und sich in der Welt zurechtzufinden, wie es bei unserem *Sehen* der Fall ist?

Die intuitive Antwort ist klar und eindeutig:

Intuitiv ist es evident, dass das ***"Bild"*** der Umgebung, das uns stets ***als Ganzes*** präsent ist – nicht nur *sensorisch*, sondern auch in seiner *Bedeutung*, mitsamt allen darin enthaltenen Objekten und deren Beziehungen – in einem geradezu phantastischen Maß der pixelweise gegebenen Information überlegen ist: die Information muss erst zusammengesetzt und dann analysiert werden, zahlreiche Erkennungsvorgänge sind auszuführen, die möglichen Beziehungen der erkannten Objekte müssen hinsichtlich ihrer Eignung bestimmt werden, Teil des Gesamtszenarios zu sein, dessen Bedeutung ebenfalls erst ermittelt werden muss, und so weiter und so fort.

Mit der gleichen Sicherheit kann behauptet werden:

Um das, was wir *sehen,* auch zu ***verstehen***, benötigen wir ebendieses gerade skizzierte "Bild". Offenbar gilt aber ganz allgemein, dass derjenige geistige Zustand, den wir ***Verstehen*** nennen, ein Szenario von genau derselben Art voraussetzt wie dieses *wahrgenommene* Bild: ein ***vorgestelltes*** "Bild", in dem die Objekte und Fakten versammelt sind, die wir für das Verständnis der Gesamtsituation benötigen.

Die ontologisch-analytische Sicht unterstützt diese Behauptung:

Wir haben gezeigt, dass *Geist* die *kausale Ebene* des neuronalen Netzes ist. Wir fassen *geistige Tätigkeit* somit nicht als *Dynamik der Neurone* auf, sondern als *Dynamik der geistigen Zustände,*

---

17 Moravec begründete diesen überraschenden Sachverhalt mit dem Argument, dass die Evolution viel länger Zeit dafür hatte, unsere Sensomotorik zu perfektionieren, als dafür, unsere logisch-abstrakten Fähigkeiten auszubilden. Ich halte dieses Argument für unzureichend: *Komplizierte* Bewegungsabläufe müssen zunächst im Motorcortex entwickelt werden – einer Struktur des Neocortex, dem evolutionsgeschichtlich *jüngsten* Teil unseres Gehirns – und erst dann werden sie im viel älteren Kleinhirn abgelegt. Die betreffenden Fähigkeiten sind also ebenso neu wie die Fähigkeit zum logischen Denken, das ja ebenfalls im Neocortex stattfindet. (Man versuche einmal, einem Orang-Utan Bogenschießen beizubringen – das wird genauso wenig Erfolg haben wie der Versuch, seine logischen Fähigkeiten zu optimieren.) Umgekehrt werden wir vermutlich auch in vielen Millionen Jahren nicht besonders gut rechnen können. (Falls es uns dann noch gibt.)

[Die folgende aktuelle und witzige Variante von Moravec's Paradox habe ich im Internet gefunden: "Früher dachte ich, irgendwann in der Zukunft hätte ich endlich genug Zeit zu dichten und zu malen, während mein Roboter aufräumt und putzt. Aber es ist ganz anders gekommen: Jetzt habe ich viel Zeit aufzuräumen und zu putzen, während mein Roboter dichtet und malt."]

d.h. der neuronalen Muster, die wir als *Attraktoren* der neuronalen Dynamik bestimmt haben. *Geist* ist demnach als *Netzwerk von Attraktoren* aufzufassen.

Daraus folgt, dass es für uns – anders als bei der reinen Information ohne Empfindung – *nicht* erforderlich ist, Details zu erkennen, zu analysieren, zusammenzufügen, in Beziehung zu setzen, mögliche Folgen abzuschätzen usw.

Warum? – Man erinnere sich: eine wichtige Eigenschaft von Attraktoren ist, dass das System von der Menge der Parameterwerte, die im Einzugsbereich des Attraktors liegen, nur eine kleine Teilmenge benötigt, um den Attraktor-Zustand herzustellen.

Dazu ein einfaches Beispiel:

Für uns kann bereits die Beobachtung *kurze rote Hose* und *kleines rollendes Objekt* ausreichen, um das Vorstellungsbild "Kind verfolgt Ball" wachzurufen, einschließlich der möglichen Folgen – eine für das Fahren eines Autos unter Umständen äußerst wichtige Information.

"Geistige Zustände" sind also stets ***Gesamtheiten***, genauso wie die Empfindungszustände, von denen wir soeben gesprochen haben: das "Bild" der Umgebung oder das "innere Vorstellungsbild", die uns immer *als Ganze* gegeben sind. ***Sie enthalten bereits alle Details und Zusammenhänge***, die bei der reinen Informationsverarbeitung erst einzeln erkannt und analysiert werden müssen.

Wie es scheint, ist Empfindung, das undefinierbare Element unseres Geistes, für diese integrative Leistung – die Präsentation des Gesamtbildes einschließlich des Zusammenhangs aller Einzelheiten – verantwortlich, also genau dasjenige, was KI-Systemen fehlt.

Allerdings ist Folgendes zu bedenken:

Da künstliche neuronale Netze auch dann empfindungslos bleiben, wenn sie Attraktoren ausbilden (siehe Teil 2, Seite 23, zweite Bemerkung), ist das Attraktor-Argument zwar geeignet, die integrative Leistung von Empfindung zu verdeutlichen, aber die Existenz von Attraktoren kann nur eine *notwendige* und keinesfalls eine *hinreichende* Bedingung dafür sein.

Der *eigentliche* Grund für diese Leistung liegt in Folgendem:

Zur Existenz eines *wirklichen* Objekts muss etwas gehören, wovon die *Aktivität* dieses Objekts ausgeht. Dieses Element seiner Existenz haben wir als *Substanz* bezeichnet. Der erste Schritt unseres Beweises der Empfindungslosigkeit von KI-Systemen (Teil 2, Abschnitt 2.2) bestand darin, zu zeigen, dass *Empfindung die Substanz der geistigen Zustände* ist. Also ist *Empfindung der Antrieb der geistigen Aktivität*.

Somit hat *Empfindung* in einem neuronalen Netz, das Geist hervorbringt, denselben Status wie *Masse* in einem System, dessen Dynamik durch Gravitation verursacht wird, wie z.B. unser Sonnensystem. So, wie *Masse* die Wechselwirkung der Objekte des Sonnensystems bestimmt, so bestimmt *Empfindung* die Wechselwirkung der Objekte, aus denen unser Geist besteht, d.h. unserer geistigen Zustände.

Das bedeutet:

Ebenso, wie Masse die Objekte eines gravitativen Systems lenkt und miteinander verbindet – wie z.B. Erde und Mond – so lenkt Empfindung die Objekte eines geistigen Systems und verbindet sie miteinander – wie z.B. Kind und Ball – und sie tut es, wie Masse, von selbst.

Kann diese Verbindung in einer Simulation nachgeahmt werden? Nicht in jedem Fall, da es eine *absolute Grenze* zwischen Original und Simulation gibt. Folgendermaßen:

Wir wissen: Die Dynamik eines Systems ***ohne*** *Gravitation* kann mit der Dynamik eines *durch Gravitation gesteuerten* Systems niemals vollkommen übereinstimmen.

Also muss genauso gelten: Die Dynamik eines Systems ***ohne*** *Empfindung* kann mit der Dynamik eines *durch Empfindung gesteuerten* Systems niemals vollkommen übereinstimmen.

Zwischen künstlicher Intelligenz und menschlichem Geist existiert eine absolute Grenze.

Allerdings wissen wir nicht, *wo* diese Grenze verläuft.

Bei der *Gravitation* verfügen wir über eine mathematische Beschreibung der Wechselwirkung, sodass wir die Grenzen einer Simulation zumindest abschätzen können.

Bei der *Empfindung* besteht diese Möglichkeit nicht: In diesem Fall ist der Versuch einer mathematischen Beschreibung vollkommen aussichtslos. Wie sollte die Beschreibung der Dynamik eines Systems gelingen, das aus Objekten – Attraktoren – besteht, deren komplexere Formen schon für sich betrachtet mathematisch kaum beherrschbar sind, und die sich überdies infolge ihrer Wechselwirkungen *permanent verändern*?

Hier erreicht die Nicht-Berechenbarkeit der Wirklichkeit einen Grad, der einen mathematischen Zugang mit Sicherheit ausschließt – nicht nur jetzt, sondern auch in jeder denkbaren Zukunft.

Wir sind also einerseits auf ontologische Argumente angewiesen, und andererseits auf das, was Selbstbeobachtung uns über unser eigenes Denken mitteilt – wie wir *erkennen, verallgemeinern, erklären, schlussfolgern usw.* – und was daraus folgt.

Ich will aber hier abbrechen und die Diskussion erst später fortsetzen, und einen weiteren Argumentationsstrang beginnen, der sich auf gegenwärtig verfügbare Arten von KI-Systemen bezieht.

Glücklicherweise sind wir neuerdings in der Lage, sowohl die Leistungsfähigkeit als auch die Grenzen der aktuellen KI einzuschätzen. Für *symbolic AI* – der "klassischen" Art des Programmierens von KI-Systemen, bei der eine logische Struktur aus definierten Elementen errichtet wird – ist das schon länger der Fall, aber bei selbst-lernenden neuronalen Netzen – wie z.B. *Generative Pre-trained Transformers* (GPTs) – wissen wir erst seit kurzem, zu welch unglaublichen Leistungen sie imstande sind und welch seltsamen Beschränkungen sie dennoch unterliegen.

Um das zu demonstrieren und die allgemeine Beurteilung der Leistungsfähigkeit von KI vorzubereiten, betrachten wir zunächst einige instruktive Beispiele. Wir beginnen mit einem einfachen neuronalen Netz, das wir für das Erkennen von handgeschriebenen Ziffern trainieren wollen.[18] Dafür benötigen wir eine ausreichend große Menge von Abbildungen solcher Ziffern.

Ein *Trainingsdurchgang* besteht darin, dass dem Netz alle Elemente dieser Menge präsentiert werden. *Input* sind die Grauwerte der Pixel dieser Abbildungen.

Zunächst ordnen wir allen Neuronen (außer denen des *input layers*) zufällige Zahlen zu (genannt *biases*), die – in Analogie zu biologischen neuronalen Netzen – die *Anfangs-Aktivierungen* der Neuronen darstellen. Den Verbindungen, die von allen Neuronen eines *layers* zu allen Neuronen des nächsten *layers* führen, ordnen wir ebenfalls zufällige Zahlen zu (genannt *weights*), die die *Stärke des Einflusses* eines Neurons auf das mit ihm verbundene Neuron ausdrücken.

Da die Anfangswerte zufällig sind, wird die Erkennungsrate beim ersten Durchgang nicht höher sein als die eines Zufallsgenerators.

18 An dieser Stelle hatte ich eigentlich eine Beschreibung des Netzwerks geplant, die zugleich als kurze Einführung dienen sollte. Die Ausführung dieses Vorhabens habe ich aber bald abgebrochen: für Personen, die mit neuronalen Netzwerken ganz unvertraut sind, wäre die Einführung in jedem Fall zu kurz und deshalb nicht hilfreich gewesen, und für alle anderen ist sie ohnehin überflüssig. Ich verweise stattdessen auf die Seite https://www.youtube.com/@3blue1brown, auf der unter dem Stichwort "Neural Networks" eine ausgezeichnete mehrteilige Einführung zur Verfügung steht, die außerdem sehr schön graphisch aufbereitet ist. In meiner eigenen Darstellung werde ich mich auf die Sachverhalte beschränken, die für meine spätere Argumentation wichtig sind.

Dann versuchen wir, die Leistung des Systems zu verbessern, indem wir vor dem Beginn jedes weiteren Durchgangs die *weights* und *biases* verändern. Wir betrachten sie somit als *Variable*.

Unser Ziel ist, die Fehlerrate zu minimieren. Sie kann als *Funktion dieser Variablen* aufgefasst werden. Wir suchen also die *Minima* dieser Funktion.

Auf diese Weise gelingt es mit relativ einfachen Mitteln, nach einer großen Anzahl von Durchgängen, nicht nur beim Trainingsset, sondern auch bei beliebigen anderen Mengen handgeschriebener Ziffern eine Erkennungsrate nahe an 100% zu erreichen.

Obwohl das neuronale Netz einfach und die ihm gestellte Aufgabe begrenzt ist, gibt es doch Anlass zu genau den Fragen und Hypothesen, denen wir im Weiteren nachgehen wollen.

Die erste Frage ist:

Nach welchen Kriterien erkennt das Netz die Ziffern?

Wir wissen jedenfalls, wie *wir selbst* vorgehen: Wir sehen jede Ziffer aus klar definierten Bestandteilen zusammengesetzt, z.B. die 3 aus zwei links offenen übereinander gestellten Halbkreisen, oder die 4 aus drei auf bestimmte Weise angeordneten Abschnitten von Geraden usw.

Unsere Art der Erkennung ergibt sich also aus der *Konstruktion* der Ziffern. ***Wir kennen die Ziffern*** und nehmen sie als Stück für Stück aufgebaut wahr.

Geht das künstliche Netz auf dieselbe Weise vor? Das ist äußerst unwahrscheinlich, ***da das Netz die Ziffern <u>nicht</u> kennt***.

Das mag seltsam klingen, da es doch imstande ist, sie zu *erkennen*. Aber diese Erkennung erfolgt nicht, wie bei uns, durch den *Vergleich* mit der *Vorstellung* der Ziffer.

Im Netz beruht der Erkennungsvorgang auf einem vollkommen anderen Prinzip: Tatsächlich sind die Ziffern im Netz *nicht direkt repräsentiert*, oder sagen wir: nur *implizit* und nicht *explizit* repräsentiert. Ein Vergleich ist also nicht möglich.

Aber durch das Training hat das Netz eines der (lokalen) Minima der oben erwähnten Funktion gefunden und ist dabei nach Kriterien vorgegangen, die für uns gänzlich undurchschaubar sind.

Die beiden Verfahren schließen sich gegenseitig aus. Es muss daher angenommen werden, dass unsere Art des Erkennens keinem Minimum der Funktion im Suchraum des Netzes entspricht.

Es ist verblüffend, dass es außer unserer Methode der Form-Analyse überhaupt andere Möglichkeiten gibt, Kriterien der Ziffernerkennung zu finden. *Wir* sind jedenfalls nicht dazu imstande, uns eine solche Möglichkeit vorzustellen. Das liegt aber zweifellos daran, dass sich die Funktion, deren Minima gesucht sind, in einem extrem hoch-dimensionalen Raum befindet – die Zahl seiner Dimensionen (die ja gleich der Zahl der Variablen ist) kann schon bei relativ kleinen und einfachen Netzen größer als 100.000 sein, während unsere Vorstellung auf Räume mit maximal 3 Dimensionen begrenzt ist.

Im Vergleich mit der Dimensionszahl des *Such-Raums* ist die Dimensionszahl des *Erkennungs-Raums*, in dem das Netz nach Beendigung des Trainings die Ziffern nun tatsächlich identifiziert, allerdings relativ klein: Die *weights* und die (anfänglichen) *biases* sind jetzt konstant, die einzigen *veränderlichen Größen* sind die *Aktivierungen* der Neuronen der *hidden layers* (der Neuronen-Schichten *zwischen* Input- und Output-Schicht), die sich aus dem jeweiligen Input ergeben. Sie sind also die Größen, die als Koordinaten des Erkennungsraums aufgefasst werden können. Die Dimensionszahl ist daher gleich der Zahl aller Neuronen minus der Zahl der Input- sowie der Output-Neuronen.

Jeder Ziffer ist eine *Untermenge* dieses Erkennungs-Raums zugeordnet, und zwar genau diejenige Untermenge, in die alle Inputs führen, die von Bildern stammen, bei denen das Netz das zu dieser

Ziffer gehörende Output-Neuron aktiviert (das können aber auch Unsinn-Bilder oder Zufallsbilder sein). Die Vereinigungsmenge all dieser Untermengen ist diejenige Untermenge des Erkennungsraumes, die aus *allen* Werten der Aktivierungen besteht, die als Folge jedes überhaupt möglichen Inputs auftreten können.[19]

Trotz dieser starken Reduktion ist aber auch diese Dimensionszahl noch viel zu hoch für unsere Vorstellung, und auch mathematische Analysen bringen uns kein Verständnis, nach welchen Kriterien das Netz die Ziffern erkennt.

Das *könnte* als Hinweis auf die ungeheuren Möglichkeiten von KI-Systemen interpretiert werden, Zusammenhänge – Gesetzmäßigkeiten oder semantische Strukturen – auf eine Art zu erkennen, die der unsrigen in einem *unvorstellbaren* Ausmaß überlegen ist – oder auch nicht. Wir werden darauf zurückkommen.

Das nächste Szenario, das wir betrachten, ist der Go-Wettkampf aus dem Jahr 2016 zwischen dem von Google DeepMind entwickelten neuronalen Netz AlphaGo und dem südkoreanischen Go-Meister Lee Sidol, den damals viele Experten für den besten Spieler der Welt hielten.

Vor diesem Wettkampf galt Go als Domäne menschlicher Intelligenz und Kreativität, weil es aufgrund der ungeheuren Anzahl möglicher Spielverläufe für *symbolic AI*, die beim Schach Menschen bereits weit hinter sich gelassen hatte, nach wie vor unerreichbar war, und weil selbstlernende neuronale Netze noch kaum bekannt und erprobt waren.

AlphaGo gewann 4:1. Legendär wurde sein 37. Zug aus der zweiten Partie: es war ein Zug, der in der seit Jahrhunderten ständig weiter entwickelten Go-Theorie als verboten galt, weil man zu wissen glaubte, dass er Stellungsnachteile zur Folge hätte. Wie sich herausstellte, war es jedoch der Gewinnzug. AlphaGo wurde gefeiert, seine Kreativität wurde "einzigartig" genannt, und dem Programm wurde der neunte Dan verliehen, der höchste im Go mögliche Rang, mit der Begründung, dass es mit seinem Spiel fast "in göttliche Regionen" vorgedrungen sei.

Es sollte noch hinzugefügt werden, dass AlphaGo – das noch über eine umfangreiche Datenbasis verfügte, wenig später von AlphaZero – das außer den Go-Regeln überhaupt kein Wissen über Go besaß und sein Spiel *ausschließlich* durch Optimierung in Milliarden von Trainingsdurchgängen gegen AlphaGo verbessert hatte, kurze Zeit später 100:0 geschlagen wurde.

Damit schien erwiesen, dass die menschliche Intelligenz der künstlichen Intelligenz in der Gestalt selbstlernender neuronaler Netze hoffnungslos unterlegen ist, nicht nur, was logisches Denken betrifft, sondern auch in dem Bereich, zu dem wir ein exklusives Zugangsrecht zu besitzen meinten: dem Bereich der Kreativität.

Ist damit unser Schicksal besiegelt? Keineswegs! – Die Geschichte ist noch nicht zu Ende:

Anfang 2023 berichtete ein KI-Forschungsteam, es sei ihm gelungen, eine Strategie zu entwickeln, mit der die besten Go-Programme geschlagen werden könnten.[20]

---

19 Diese unübliche Sichtweise dient dazu, die Begriffe "implizit" und "explizit" klarer zu bestimmen: Wir betrachten hier nicht die *Funktion*, die das Netz auf den Input anwendet, sondern den *Raum*, in dem der Erkennungsvorgang stattfindet. Auf diese Weise kann am ehesten verstanden werden, was damit gemeint ist, dass die Ziffern nur "implizit" im System repräsentiert sind und nicht "explizit": Die Untermenge, die der Ziffer 2 zugeordnet ist, kann mit Sicherheit nicht als "explizite" Darstellung der Ziffer 2 aufgefasst werden, und wenn – bei einer Abbildung der Ziffer 2 als Input – die Aktivierungen der Netz-Neuronen Werte annehmen, die den Koordinaten eines Punkts entsprechen, der sich in der dieser Ziffer zugeordneten Untermenge befindet, dann bedeutet das keinesfalls, dass das Netz die Form der Ziffer 2 ***kennt*** – in keinem möglichen Sinn dieses Wortes.

20 https://arxiv.org/abs/2211.00241

Kellin Pelrine, ein Mitglied dieses Teams und Go-Spieler auf gutem Amateur-Niveau, schlug KataGo – eines der spielstärksten neuronalen Netze – 14:1.

Wie Stuart Russel mitteilte,[21] besiegte Pelrine KataGo auch dann, wenn er ihm 9 Steine vorgab, und in diesem Fall sogar 15:0. Ebenso schlug er auch andere, genauso leistungsstarke Go-Programme, die von verschiedenen Teams mit verschiedenen Methoden entwickelt worden waren.

Wie ist das möglich? Darauf gibt es eine klare Antwort:

Die Programme haben ***keine Ahnung vom Spielprinzip*** *– sie* ***wissen nicht***, dass es darum geht, Gebiete und gegnerische Steine einzuschließen. Wie sich an ihren verlorenen Partien ablesen lässt, ***erkennen sie nicht****, dass sie eingeschlossen werden,* und lassen es deshalb zu, obwohl sie einige Züge lang Gelegenheit hätten, es zu verhindern.

Warum zeigen sie dennoch in Partien gegen Go-Meister diese unglaubliche Spielstärke?

Weil sie – genauso wie das Programm zur Ziffernerkennung – ihre Leistung optimiert haben, indem sie Minima einer Funktion von ungeheuer vielen Variablen in einem extrem hochdimensionalen Raum gesucht und gefunden haben. Dabei haben sie Spielstrategien entdeckt, die uns nie in den Sinn kommen könnten, aber andererseits haben sie überhaupt keine Chance, sich gegen die menschliche – hier *Kellin Pelrines* – Strategie zu wehren, weil sie nicht in der Nähe eines Minimums in ihrem hochdimensionalen Suchraum liegt.

Warum ist das so? Die Antwort ist ganz ähnlich der des vorigen Beispiels:

Bei der Diskussion der Ziffernerkennung stellten wir fest, dass das Netz die Ziffern *nicht kennt,* ja nicht einmal kennen *kann,* und dass deshalb *unsere* Erkennungsmethode, die *konstruktiv* ist und auf dem *Vergleich* mit der vorgestellten Ziffer beruht, keinem Minimum der Funktion im Suchraum entspricht.

Beim Go-Spiel können wir allerdings nicht von *unserer* – d.h. der *menschlichen* – Strategie im Allgemeinen reden: Menschen haben beim Go-Spielen ganz verschiedene Strategien.

Aber wir können Folgendes behaupten:

Das Go-Spielprinzip ist die wichtigste Voraussetzung ***aller*** *menschlichen Strategien.*

Andererseits wissen wir:

Das neuronale Netz ***kennt das Spielprinzip nicht****.*

Dennoch muss das Spielprinzip *implizit* im KI-System präsent sein und bei seiner Strategie eine Rolle spielen – ansonsten wären selbstlernende neuronale Netze ja vollkommen unfähig, Go zu spielen. Das systematische Optimieren der Spielstrategie *kann* nur unter der Voraussetzung gelingen, dass das Spielprinzip an diesem Prozess beteiligt ist.

Aber das bedeutet eben genau dies: *das Prinzip ist zwar* ***Voraussetzung des Optimierungs-prozesses****, aber es wird* ***nie ins System selbst*** *integriert.*

Es verhält sich also auch hier wiederum genauso wie bei der Ziffernerkennung, wo die *Gestalt* der Ziffern ebenfalls beim Suchvorgang *implizit* präsent sein muss, aber nicht *explizit* im System existiert: obwohl das System die Ziffern erkennt, *kennt es sie nicht,* und ebenso gilt: obwohl das Go-Programm das Spielprinzip bei seinen Siegen geradezu perfekt anwendet, *weiß es von diesem Prinzip nichts.*

Wie ist also das Verhältnis zwischen menschlicher und KI-generierter Spielweise? Das lässt sich an den vorliegenden Ergebnissen ablesen:

21 Stuart Russell, "AI: What If We Succeed?" April 25, 2024, Institute for the Study of Ancient Cultures Museum.

Offenbar halten sich menschliche Go-Meister *immer* in Spielszenarien auf, die sich in den "Tälern" der hochdimensionalen Funktion befinden, die die KI-Systeme auf ihrem Weg zu den lokalen Minima erforscht haben. In diesem Fall sind die Menschen chancenlos, weil die KI-Systeme *auf jeden Fall* näher am Minimum sind.

Daraus folgt, dass sich Menschen von den Bereichen (Spielsituationen) möglichst *fern halten* müssen, die die KI-Systeme bei ihrer Selbst-Optimierung erforscht haben. Einfach gesagt: Sie dürfen nicht *allzu gut* spielen.

Die viel wichtigere zweite Bedingung ist, dass sie sich auf das konzentrieren müssen, was KI-Systeme *nicht erkennen:* darauf, *gegnerische Steine zu umstellen* – im Sinn der ersten Regel vor allem dann, wenn die dafür erforderlichen Züge eigentlich *schlechte* Züge sind, weil sie dem Spielaufbau nicht dienlich sind.

Wie sich gezeigt hat, haben Menschen ausgezeichnete Siegeschancen, wenn sie sich an diese beiden taktischen Anweisungen halten.

Kurz zusammengefasst: Neuronale Netze, die ihre Spielstärke durch Selbstoptimierung erreichen, haben keine Chance, das *Spielprinzip* zu *verstehen*. Menschen, die imstande sind, diese Einschränkung auszunützen, sind ihnen klar überlegen.

Leider kannte im Jahr 2016 weder Lee Sidol noch irgendjemand sonst diese fundamentale Schwäche. Ansonsten hätte Lee Sidol mühelos gewonnen, und die Einschätzung der Leistung von AlphaGo wäre mit Sicherheit ganz anders ausgefallen.

Weiter oben – im Abschnitt 3.2 – haben wir gezeigt, dass der Beweis, dass KI-Systeme keine Empfindungen haben, uns zu einem differenzierteren Sprachgebrauch zwingt, zu einer genaueren Definition etlicher Wörter aus dem Bereich der Wahrnehmung und der Motivation.

Gleiches ereignet sich nun im Bereich der Beurteilung von Leistungen. Bei Menschen ist es selbstverständlich, dass erstaunliche Leistungen, wie der 37. Zug aus der 2. Partie, nur aus einem *tiefen Spielverständnis* herrühren können. Sie *kreativ* zu nennen, war also bisher fest mit diesem Sachverhalt verknüpft. Man kann nun diesen Ausdruck auch weiterhin gebrauchen, aber wenn er auf die Leistung eines neuronalen Netzes angewendet wird, dann wird er *umdefiniert*, da diese Leistung jetzt eben nicht mehr *aufgrund tiefer Einsicht* erfolgt, sondern ganz im Gegenteil *trotz ihres vollständigen Fehlens.*

Ein Teil der Bedeutung des Wortes *kreativ* bliebe allerdings erhalten, da ja tatsächlich *etwas Neues* entdeckt wurde. Aber wenn ich Sie nun direkt fragte: "Würden Sie einen rasenden Idioten, der so lange bergab rennt bis er über etwas stolpert was so tief unten liegt dass es vorher noch niemand gefunden hatte, *kreativ* nennen?", dann würden sie vielleicht zögern.

Noch dramatischer wäre die Veränderung, die der Begriff *Verstehen* erführe, wenn er auf AI-Systeme angewendet würde: Da *Verstehen* vor der Entwicklung von AI-Systemen eine *selbst-verständliche* Voraussetzung großer geistiger Leistung war, musste der Person, die diese Leistung erbrachte, auf jeden Fall Verstehen zuerkannt werden. Wenn nun aber ein Neuronales Netz ebensolche – oder noch viel bedeutendere – Leistungen erbringt, dann würde durch die Zuerkennung von Verstehen dieser Begriff *vollständig* seines Sinnes beraubt – seine Verwendung wäre einfach grob falsch.

Lässt sich diese Schwäche der KI-Systeme korrigieren?

Hören wir dazu wieder Stuart Russel zu selbstlernenden neuronalen Netzen ganz allgemein:[22]

22 Stuart Russell, a.a.O.

"...if you have a very, very large representation of what is fundamentally actually a simple concept, then you would need an enormous number of examples to learn that concept. Far more than you would need if you had a more expressive way of representing the concept."

– wobei mit "more expressive way" eine Programmiersprache wie *Python* gemeint ist, in der sich z.B. das Go-Spielprinzip ganz einfach ausdrücken ließe.

Allerdings beschreibt Stuart Russell das Problem hier sehr zurückhaltend, denn tatsächlich bedürfte es – falls die Aufgabe des KI-Systems ein *reales* und daher nicht vollständig definierbares Szenario betrifft – *unendlich vieler* Beispiele, um das Konzept *vollständig* in das System zu integrieren.

Das Go-Spiel besteht zwar aus endlich vielen definierbaren Zuständen, aber ihre Zahl ist doch so groß, dass es nicht möglich ist, *alle* erfolgreichen Gegenstrategien auszuschließen.

Kurz gesagt: Das Go-Spielprinzip ***kann nicht*** vollständig in das KI-System integriert werden.

An dieser Stelle drängt sich die Frage auf, ob sich dieser Mangel selbstlernender neuronaler Netze nicht dadurch beheben ließe, dass sie durch *symbolic AI* ergänzt werden. Weiter unten werden wir uns mit dieser Frage beschäftigen. Jetzt widmen wir uns aber unserem nächsten Beispiel, jener Art von KI-System, das *Generative Pre-trained Transformer* genannt wird.

Was ist ein GPT? Was kann er? – Ich werde die Antwort hier nur soweit skizzieren, wie es erforderlich ist, um an die bisherigen Überlegungen anschließen zu können.

GPTs sind lernfähige neuronale Netze, die imstande sind, große Systeme von *strukturierten* Daten nachzubilden und auf dieser Grundlage etwas zu produzieren – Texte, Bilder, Übersetzungen usw.

Im Fall von LLMs (Large Language Models) bedeutet das: Sie erfassen die grammatische, syntaktische und semantische Struktur der Sprache im Allgemeinen, und *zusätzlich* auch die semantischen Strukturen sprachlicher Gebilde, nicht nur von Sätzen, sondern auch von größeren Einheiten – Geschichten oder literarischen Werken verschiedenster Art –, mit anderen Worten: sie sind imstande, auch *kontextabhängige* semantische Strukturen darzustellen.

Diese Leistung zu erbringen erfordert eine immens aufwendige Trainingsphase. Der Lernvorgang wird dadurch vorbereitet, dass die Daten in kleine Elemente, sogenannte *Tokens,* zerlegt werden – im Fall von Sprache also in Wörter, Wort-Teile, Silben oder sogar Buchstaben, im Fall von Bildern in Motive, Bildausschnitte oder Pixel, im Fall von akustischen Daten in charakteristische Elemente wie Töne oder Geräusche, oder einfach kurze zeitliche Ausschnitte usw.

(Der besseren Verständlichkeit wegen werde ich mich im Folgenden auf sprachliche Tokens beschränken.)

Zunächst wird eine Liste *aller* Tokens erstellt, die in der Datenmenge vorkommen.

Den Tokens werden *Vektoren* zugeordnet. (Beim GPT3 haben diese Vektoren mehr als 12.000 Komponenten.) Die Tokens werden also durch Vektoren in einem hochdimensionalen, abstrakten Raum dargestellt.

Die Zahlenwerte der Komponenten sind zunächst zufällig (wie auch schon bei unseren beiden vorherigen Beispielen).

Der Lernprozess besteht darin, dass dem GPT Abschnitte aus Texten präsentiert werden. Seine Aufgabe ist, das *nächste Wort* zu finden, d.h. das Wort, das dem jeweiligen Abschnitt folgt.

Es ist einzusehen, dass ihm das nur dann gelingen kann, wenn seine vektorielle Repräsentation aller Tokens – und damit zugleich aller Wörter – die grammatische und syntaktische Struktur der Sprache ganz allgemein, und im Besonderen auch die semantische Struktur des betreffenden Textes nachbildet.

Die Fehlerrate kann als Funktion der Vektor-Komponenten aufgefasst werden. Sie sind also die Variablen dieser Funktion, und das Ziel ist somit auch hier wieder, die Minima der Funktion zu finden.

Man ahnt, dass das Erreichen dieses Ziels nur mit ungeheurem Aufwand möglich ist: Die Datenmenge ist riesig – im Grunde alle im Internet zugänglichen Sätze – und die semantischen Strukturen von Sprachprodukten sind komplex und vieldeutig. Deshalb sind frühere Versuche mit selbstlernenden neuronalen Netzen gescheitert. Erst die extrem gesteigerte Speicher- und Rechenkapazität hat den gegenwärtigen Erfolg ermöglicht.

Anschaulich ausgedrückt, ist der Trainingsprozess eine Erforschung des Grades der "Nähe" oder "Zusammengehörigkeit" von Wörtern, und auch ihrer "Verwandtschaft". Am Ende dieses Prozesses sollten also Wörter mit ähnlicher Bedeutung durch Vektoren repräsentiert werden, die in ähnliche Richtungen weisen.

Ein Effekt der Repräsentation durch Vektoren ist, dass *Richtungen* in diesem hochdimensionalen Darstellungsraum Elemente der *semantischen Struktur* sind. Ein bekanntes Beispiel dafür ist, dass der Vektor (Frau *minus* Mann) fast genau dem Vektor (Tante *minus* Onkel) entspricht, oder auch dem Vektor (Tochter *minus* Sohn). Die drei Differenz-Vektoren sind annähernd parallel und von gleicher Länge, und ihre *Richtung* hat die Bedeutung "(Änderung der) Geschlechtszugehörigkeit".

Parallel zu den semantischen Zusammenhängen müssen aber auch die grammatischen und syntaktischen Regeln der Sprache erlernt werden: Wortarten, Satzbau usw.

Ich will an dieser Stelle abbrechen, denn trotz der Unvollständigkeit und des anekdotischen Charakters dieser einleitenden Skizze ist das bisher Gesagte bereits als Hintergrund für die Fragen ausreichend, die wir nun ein weiteres Mal stellen wollen:

Wir wissen ja schon, wie das Netz die Wörter repräsentiert. Was wir aber nicht wissen, ist, *nach welchen Kriterien* diese Repräsentation erfolgt.

Wie *wir selbst* vorgehen, ist uns klar: Im Grunde verfügen auch wir über einen solchen "Darstellungsraum", auch wenn uns das nicht direkt bewusst ist. Wir definieren Wörter durch *Eigenschaften*, und somit sind diese Eigenschaften *unsere* Kriterien: die Komponenten *unserer* Vektordarstellung und daher auch die Koordinaten *unseres* Darstellungsraums.

Geht das Netz auf dieselbe Weise vor? Mit Sicherheit nicht. *Sein* Darstellungsraum ist vollkommen abstrakt, und die Koordinaten, aus denen er aufgebaut ist, haben tatsächlich *überhaupt keine konkrete Bedeutung*. Es könnten ja auch *beliebig viele* sein – je mehr, desto besser, falls die Datenmenge groß genug ist und die Rechenleistung ausreicht.

Selbstverständlich muss aufgrund der strukturellen Übereinstimmung der beiden Räume ein statistisch erforschbarer Zusammenhang zwischen den GPT-*Kriterien* und unseren *Eigenschaften* bestehen, aber mehr ist dazu nicht zu sagen. Die Komponenten der Vektoren der GPT-Darstellung sind abstrakt und bedeutungslos.

Ich schlage ein Gedankenexperiment vor, das eine Erweiterung von Ronald Searles "Chinesischem Zimmer" ist:

Der GPT erhält die Daten der kompletten Sprachproduktion einer außerirdischen Zivilisation – genauso wie er vorher die Daten der irdischen Sprachproduktion erhalten hat.

Er führt nun dieselbe Art von Training durch.

Danach können Sie mit den Aliens chatten – Sie brauchen ja nur dem GPT deren Mitteilungen vorlegen und ihnen dann antworten, was der GPT produziert hat. (Sie könnten die Berechnungen

des GPTs auch selbst ausführen, aber dafür würde die Dauer der Existenz des Universums kaum ausreichen.)

Sie unterhalten sich also bestens, erzählen sich Witze, und werden gute Freunde. Oder etwa nicht?

Nun, mit der Freundschaft wird es wohl nichts. Sie haben ja überhaupt keine Ahnung, worüber sie sich eigentlich unterhalten haben. Vielleicht war es über die Lieblingsbeschäftigung der Aliens, das Verspeisen von Angehörigen anderer Zivilisationen?

Aber halt! – vielleicht versteht ja der GPT etwas von der Kommunikation?

Nein. Ebenso, wie das Programm zur Ziffernerkennung die Ziffern nicht kennt, weiß auch der GPT nichts von der *Bedeutung* der Wörter – seine Leistung beruht auf den *statistischen Gesetzmäßigkeiten* ihres Auftretens, die sich aus den (grammatischen, syntaktischen und semantischen) Strukturen der Sprachproduktionen ergeben, die umgekehrt wiederum aus der Statistik folgen.[23]

Aber auch hier gilt wieder genau dasselbe wie zuvor:

Die grammatischen, syntaktischen und semantischen Strukturen haben zwar ***den Optimierungsprozess gesteuert****, aber* ***nichts davon ist im GPT explizit präsent****.*

Er weiß also genauso wenig wie Sie.

Mit anderen Worten: ***Der GPT versteht*** von der Kommunikation mit den Aliens *genauso viel* wie von seiner Kommunikation mit Menschen, und das ist exakt ***überhaupt nichts***.

Wir begegnen hier wieder demselben Sprachproblem wie bei unseren vorhergehenden Beispielen:

Wenn *Menschen* vernünftig reden, dann wäre es absurd zu behaupten, dass sie nicht *verstehen*, was sie sagen.[24] Jetzt sind wir aber gezwungen, diese feste Verbindung von "vernünftig reden" und "verstehen" aufzugeben. So, wie neuronale Netze etwas richtig *erkennen* können (erkennen im Sinn von *identifizieren*), ohne es zu *kennen,* und wie sie grandios Go spielen können, ohne überhaupt zu *wissen,* worum es dabei geht, so können sie auch *vernünftig reden*, ohne vernünftig zu *sein* und ohne *irgendetwas* davon zu *verstehen*.

Als Ergänzung sollen nun noch einige Beispiele folgen. Ich übernehme sie von Yejin Choi, Informatikerin und Professorin an der Universität von Washington.

Yejin Choi, 28.04.2023: Why AI Is Incredibly Smart and Shockingly Stupid (TED Talks #ai):

" … suppose I left five clothes to dry out in the sun, and it took them five hours to dry completely. How long would it take to dry 30 clothes?

GPT-4, the newest, greatest AI system says: 30 hours. – Not good.

A different one: I have 12-liter jug and six-liter jug, and I want to measure six liters. How do I do it? – Just use the six liter jug, right?

---

23 Das ist auch der Grund, warum Searles Argument mittlerweile unzureichend ist. Searle selbst hat immer wieder betont, dass die Ausführung eines Programms nur die Kenntnis einer hinreichend großen Zahl von *Regeln* voraussetzt, wobei die semantische Struktur – die er mit *Verstehen* gleichsetzt – ausgeschlossen bleibt. Es ist ihm entgangen, dass die GPTs diese Grenze überschritten haben, und zwar genau deshalb, weil *die statistischen Zusammenhänge* auch *einen Großteil der semantischen Struktur* ***enthalten***.

Der Kern von Searles Argument bleibt jedoch unangetastet: aus einer korrekten Sprachproduktion, die auf der Wahrscheinlichkeit des Auftretens von Wörtern beruht, kann ebenso wenig auf das Verständnis des Gesagten geschlossen werden wie aus einer korrekten Sprachproduktion, die auf einem feststehenden Katalog von Regeln beruht. Auf diese Weise kann Searles Argument von *symbolic AI* auf neuronale Netze übertragen werden.

24 Abgesehen von trivialen Fällen, wie einen Text auswendig aufsagen oder ablesen.

GPT-4 spits out some very elaborate nonsense: Step one, fill the six-liter jug, step two, pour the water from six to 12-liter jug, step three, fill the six-liter jug again, step four, very carefully, pour the water from six to 12-liter jug. And finally you have six liters of water in the six-liter jug that should be empty by now.

OK, one more.

Would I get a flat tire by bicycling over a bridge that is suspended over nails, screws and broken glass?

Yes, highly likely, GPT-4 says, presumably because it cannot correctly reason that if a bridge is suspended over the broken nails and broken glass, then the surface of the bridge doesn't touch the sharp objects directly.

OK, so how would you feel about an AI lawyer that aced the bar exam yet randomly fails at such basic common sense?

AI today is unbelievably intelligent and then shockingly stupid!"

Ich halte zwar Argumente für wesentlich wichtiger als Beispiele, aber es ist schon ziemlich eindrucksvoll, wie in diesen drei Beispielen deutlich wird, dass der GPT *überhaupt nicht versteht*, worum es jeweils geht. Vor allem die ersten beiden Beispiele zeigen, dass er kompletten Nonsens produziert, wenn in seinen Trainingsdaten keine hinreichend ähnlichen Szenarien vorhanden sind – oder auch, wenn er einfach die falschen auswählt, weil er ja die kausalen Zusammenhänge nicht erfasst.

Es ist außerdem klar, dass es sich nicht bloß um unbedeutende "Pannen" handeln kann: *Menschen* mögen für solche Pannen anfällig sein – selbst dann, wenn sie intelligent sind.

Aber wenn *AI-Systeme* intelligent sind, dann sind sie es entweder ***immer*** *oder* ***nie****.* Somit müssen die falschen Antworten des GPT als ***Zeichen des vollständigen Fehlens von Intelligenz*** aufgefasst werden. Ob die Antwort richtig ist oder falsch, ist also bloß *Zufall* und keine Frage der Intelligenz.

Aus diesem Grund überrascht es mich, dass selbst Kritiker der gegenwärtigen AI-Euphorie wie Yejin Choi oder Gary Marcus ihre Vorbehalte so zurückhaltend formulieren: sie sprechen von "vorläufigen Mängeln", oder auch von "Strategien für deren Korrektur", obwohl doch das *Prinzip,* das hinter dem Versagen der AI steht, klar erkennbar und tatsächlich ***nicht korrigierbar*** ist.

Was ist dieses "Prinzip"?

Genau dasjenige, dem wir in unseren drei Beispielen begegnet sind:

Bei neuronalen Netzen bedeutet *Lernen* Folgendes: Ihre Leistung wird optimiert, indem in zahlreichen Trainingsdurchgängen die Minima einer (hochdimensionalen) Funktion gesucht werden, deren Variable den – zunächst zufälligen – *weights* und *biases* der Neuronen entsprechen.

Der *Wert* dieser Funktion (die "Fehlerrate") *kann nur dann abnehmen, wenn die formalen und strukturellen Bedingungen der angestrebten Leistung den Suchvorgang steuern*.

In unseren Beispielen waren das:

– *die Gestalt der Ziffern,*

– *das Grundprinzip des Go-Spiels,*

– *die semantische Struktur des vorangegangenen Wort-Strings.*

In diesen drei – aber auch in allen anderen Fällen – steuern also diese Bedingungen zwar den Optimierungsprozess, aber sie bleiben bloß ***Voraussetzungen des AI-Systems*** und werden niemals zu einem ***Teil des Systems selbst***.

Mit anderen Worten:

Das System weiß nichts von ihnen, es kennt sie nicht, es versteht nicht, worum es geht – oder wie auch immer man diesen Sachverhalt benennen will.

Kann dieser fundamentale Mangel behoben werden?

Grundsätzlich gibt es zwei Möglichkeiten, die Zahl der Fehler – Unsinnigkeiten oder unerwünschtes Verhalten – zu verringern:

1. Man kann den Trainingsprozess durch vorher definierte Regeln beeinflussen.
2. Man kann den Output durch einen Katalog von Anweisungen steuern.

Die erste Methode kann auch von Menschen ausgeführt werden. Die zweite Methode bedeutet, das selbstlernende System durch *symbolic AI* zu ergänzen.

Für beide Arten der Verbesserung gilt jedoch das bekannte Gesetz – das auch die Leistung von *symbolic AI* generell limitiert:

Jeder Katalog von Regeln bleibt notwendig unvollständig, weil in der wirklichen Welt permanent neue Situationen auftreten.

Inzwischen erleben wir zahlreiche Anwendungsfälle dieses Sachverhalts: die Entwickler-Teams versuchen eifrig, die Dummheiten der GPTs auszubessern, und die Kritiker finden Methoden, die Korrekturen der Entwickler durch kleine Änderungen wieder wirkungslos zu machen, oder sie suchen neue Fehler.

Es geht also nicht etwa darum, die Frage zu klären:

"Kann der Mangel an Verstehen grundsätzlich beseitigt werden?"

– diese Frage ist längst beantwortet, und die Antwort ist ***nein*** –

sondern um die Frage:

"Ist der jeweilige Katalog von Korrekturen für den angestrebten Zweck ausreichend?"

Der entscheidende Punkt ist, dass *symbolic AI* keinesfalls dafür geeignet ist, das vollständige Fehlen von Verstehen zu beseitigen. KI-Systeme in der Gestalt selbstlernender neuronaler Netze ***verstehen nichts,*** und ihre Ergänzung durch *symbolic AI* kann daran (selbstverständlich) nichts ändern – sie kann nur die Fehlerzahl verringern.

Die wichtigste Konsequenz dieser Tatsache ist, dass die gegenwärtig vorherrschende AI-Technologie ungeeignet ist, AGI hervorzubringen:

AGI beruht auf Verallgemeinerung. *Verstehen* ist jedoch eine notwendige Bedingung für *alle* Arten von Verallgemeinerung.[25] Um etwa die kausale Struktur eines Vorgangs auf einen anderen Vorgang zu übertragen, ist es erforderlich, diese Struktur *zu verstehen* – alle drei Beispiele von Yejin Choi zeigen das sehr klar.

Auch hier gilt wieder: Man kann versuchen, einen Katalog übertragbarer kausaler Zusammenhänge zu erstellen, aber dieser Katalog wird *extrem* unvollständig bleiben.

Was ist mit zukünftiger KI? Oder konkreter:

25 Das gilt auch für triviale Arten von Verallgemeinerung: z.B. ist jede Form von Kompression mit Verlust eine Verallgemeinerung, da das Zurücklassen von Einzelheiten dem Fortschreiten ins Allgemeine entspricht. Aber auch hier ist *Verstehen* erforderlich, weil ja bekannt sein muss, von *welchen* Eigenschaften die kausale Struktur abhängt – sonst ist die Verallgemeinerung Glücksache (wie beim GPT).

Können die Beschränkungen gegenwärtiger KI durch künftige Hard- und Software überwunden werden?

Mit dieser Frage enden also meine beiden Argumentationsstränge, und was dazu zu sagen ist, wird der Gegenstand des nun folgenden, letzten Abschnitts sein.

3.5. Übersicht, Vergleich, abschließende Einschätzung

Ausgangspunkt meiner Argumentation über Grenzen der künstlichen Intelligenz ist der im Teil 2 geführte Beweis, dass KI-Systeme *keine Empfindungen* haben.[26]

Das bedeutet:

1. ***KI-Systeme können nichts wahrnehmen***.
   Ihnen fehlt das *"innere Theater",* das *"Bild"* der Umgebung: Sie *sehen* nicht. Ebenso gilt: sie hören nicht, fühlen nicht, riechen nicht, schmecken nicht. Für sie gibt es *nur Information*.
2. ***KI-Systeme können nichts erleben***.
   Sie haben keine Gefühle.
3. ***KI-Systeme können nichts wollen***.
   Ihnen fehlt Intentionalität und Motivation.

Im Abschnitt 3.3 haben wir zunächst auf eine selbstverständliche Folge dieses Beweises hingewiesen:

Gleichgültig, wie die Zukunft der KI auch immer aussehen mag, KI-Systeme werden aufgrund der oben genannten Einschränkungen *niemals* eine neue, überlegene Spezies sein. Die Dystopien, in denen wir ihnen ausgeliefert sind, gehören in den Bereich der Fantasy. [27]

Im Abschnitt 3.4 haben wir das *Fehlen der Wahrnehmung* bei KI-Systemen mit Moravec's Paradox in Verbindung gebracht, und anschließend die erstaunliche integrative Leistung skizziert, die unsere Wahrnehmung vollbringt. Schließlich haben wir an die in Teil 2 bewiesene Existenz einer *absoluten Grenze* zwischen einem System und seiner Simulation erinnert, die somit auch zwischen menschlichem Geist und künstlicher Intelligenz besteht.

Das alles sind Hinweise darauf, dass uns die Fähigkeit zur Wahrnehmung – im Vergleich mit Systemen *ohne* Wahrnehmung – einen beträchtlichen Vorteil verschafft. Es bleibt jedoch zunächst offen, wie weit der Nachteil der Empfindungslosigkeit von KI-Systemen durch das Anwachsen der Speicherkapazität und Rechenleistung sowie durch die Weiterentwicklung der System-Architektur ausgeglichen werden kann.

Danach haben wir uns einer anderen Argumentationsstrategie zugewendet: der Untersuchung, welche Grenzen für die gegenwärtig vorherrschenden KI-Technologien bestehen.

Diese Untersuchung hat uns dann mit überraschender Klarheit zu folgender Einsicht geführt:

Selbstlernende neuronale Netze verstehen nichts von dem, was sie produzieren.

---

26 "Empfindung" steht hier – wie immer in dieser Arbeit – für denjenigen Teil eines geistigen Zustands, der *nicht definierbar* ist, der also *über Information hinaus* geht.

27 Das gilt z.B. auch für Nick Bostroms populäres "Paperclip-Szenario" – schon der Titel der betreffenden Arbeit: "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents" reicht aus, um das Szenario in den Bereich der Fantasy zu verweisen, da der *Agent* ja weder über *Wille* noch über *Motivation* verfügt –, sowie für Marvin Minskys Riemann hypothesis catastrophe.

Sie kennen nicht, was sie erkennen, sie haben keine Ahnung, worüber sie reden, sie wissen nichts von den Prinzipien, die ihre Leistungen ermöglichen.

Außerdem haben wir festgestellt: *Symbolic AI* kann diesen fundamentalen Mangel nicht beheben. Sie kann nur die Zahl der Fehler reduzieren, die dadurch verursacht werden.

Da Verstehen eine notwendige Bedingung für (sinnvolle) Verallgemeinerung ist, bedeutet das, dass sich die gegenwärtig so populären Arten von KI nicht zu AGI weiter entwickeln lassen.

Hier stellt sich abermals die Frage, ob diese Einschränkungen durch verbesserte Hard- und Software überwunden werden können.

Bevor wir uns dieser Frage zuwenden, führen wir einen kurzen Vergleich durch, der Einiges zur weiteren Aufklärung der bisher diskutierten Themen beitragen wird: den Vergleich zwischen der Art, wie *wir* denken, und der Art, wie es bei neuronalen Netzen geschieht.

Zu diesem Zweck betrachten wir einen menschlichen Gedankengang auf genau dieselbe Weise, wie wir bisher bei künstlichen neuronalen Netzen vorgegangen sind, also nicht als *Gedankengang*, wie wir es üblicherweise tun: als Abfolge von Voraussetzungen, Vermutungen, Schlussfolgerungen, Irrtümern, Planungen usw., sondern als *neuronalen Prozess*.

Nehmen wir an, wir führen einen Zug in einem Go-Spiel aus.

Input ist die Stellung des Spiels, Output ist unser Zug. Also ist dieser Zug als Ergebnis der *Funktion* aufzufassen, die unser neuronales Netz auf den Input ausübt.

Wie in unseren Beispiel-Szenarien mit selbstlernenden neuronalen Netzen sind die Erregungszustände der beteiligten Neuronen und ihre Verbindungsstärken die *Variablen* dieser Funktion.[28]

Am Ende des Berechnungsprozesses wird unsere Hand – dem Output entsprechend – den Zug ausführen.

Wir stellen nun dieselben Fragen wie in unseren Beispielen:

Die Verbindungsstärken *(weights)* und Erregungszustände *(biases)* der beteiligten Neuronen sind die Koordinaten eines extrem hochdimensionalen Raums. Sie können als *Kriterien* oder *Komponenten* der Entscheidung betrachtet werden, die das neuronale Netz schließlich trifft.

Haben sie irgendeine Bedeutung? Offensichtlich nicht. Man könnte zwar behaupten – wie bei den Beispielen – dass zwischen *diesem* Raum und dem Entscheidungsraum, in dem der zugehörige *Gedankengang* stattfindet, eine *strukturelle Übereinstimmung* herrscht, aber mehr ist dazu nicht zu sagen.

"Weiß" ***das Netz***, *was* es tut und *warum* es das tut? In dieser Betrachtungsweise sicher nicht.

Aber ***wir selbst*** wissen selbstverständlich, worum es geht, und damit kommen wir zu den Schlussfolgerungen, zu denen uns dieses Gedankenexperiment führt – oder sagen wir besser: zu denen wir dadurch gezwungen sind.

1. Im Fall eines menschlichen Spielers ***fehlt*** bei der neuronalen Betrachtungsweise dasjenige, was den Zug ***wirklich*** plant und ausführt: ***der menschliche Verstand***. In diesem Fall fehlt er allerdings nur *in der Beschreibung – in der Wirklichkeit* ist er ja vorhanden.

2. Im Fall des KI-Systems ***fehlt der Verstand*** aber nicht nur *in der Beschreibung*, sondern auch ***in der Wirklichkeit***, und deshalb ist es ausgeschlossen, dass das System etwas ***versteht***.

28 Natürlich sind die Verhältnisse wesentlich komplizierter als in gegenwärtigen KI-Systemen. Das ist aber für unser Gedankenexperiment ohne Bedeutung.

An dieser Stelle begegnet uns ein Sachverhalt von äußerster Wichtigkeit – genau derjenige Sachverhalt, mit dem diese Arbeit begonnen hat und auf dem sie aufgebaut ist:

Unser Verstand kann diese Leistung nur dann vollbringen, wenn der Geist die kausale Ebene des neuronalen Netzes ist, und das ist wiederum nur dann möglich, wenn die physikalische Kausalität ***unvollständig*** ist.

Wäre unser Geist bloß der Vollzug physikalischer Gesetzmäßigkeiten, dann wären Gedankengänge *ohne jede Bedeutung,* und jede Schlussfolgerung wäre eine Illusion. Das ist eigentlich selbstverständlich: die Vorstellung, dass *das Denken selbst* zu korrekten Ergebnissen führt, setzt offensichtlich seine *kausale Wirkung* voraus: wie sollte es sonst möglich sein, einen Irrtum zu korrigieren? – Falls mein Denken nicht *selbst* kausal wäre – hätte sich dann etwa *die Physik* geirrt?

Immer dann, wenn ich glaube, ich hätte etwas ***deshalb*** behauptet, ***weil*** es richtig ist, habe ich die Kausalität meines Denkens *vorausgesetzt:* nur unter dieser Voraussetzung kann ein Gedanke aus einem anderen Gedanken ***folgen***.[29]

Wenn der Geist die kausale Ebene des Netzes ist, dann folgt daraus, dass die oben skizzierte *neuronale* Betrachtungsweise nicht bloß ***unvollständig***, sondern sogar ***falsch*** ist: der Beweis, dass die physikalische Kausalität unvollständig ist, schließt die Existenz einer Funktion aus, die aus dem Input den Output produziert und deren Variable die *weights* und *biases* der Neuronen sind. Man kann dafür einige Gründe angeben – der einfachste ist, dass das neuronale System sich während des Entscheidungsprozesses *verändert*. (Man stelle sich eine Funktion $f(x) = y$ vor, bei der sich die x-Achse unvorhersehbar kräuselt, während die Funktion vollzogen wird.)

Was wir soeben ausgeführt haben, ist im Grunde eine Wiederholung der Argumentation zur Willensfreiheit, nur dass wir diesmal künstliche neuronale Netze mit einbezogen haben.

Ich fasse nochmals kurz zusammen:

Bei uns selbst kann die Existenz einer *geistigen Ebene* vorausgesetzt werden; zu zeigen war, dass sie die *kausale Ebene* des Netzes ist (was wir im ersten Teil durchgeführt haben).

Bei künstlichen neuronalen Netzen ist schon durch unsere Beispiele und deren Verallgemeinerung Folgendes klar geworden: KI-Systeme, deren Output die Funktion von Variablen ist, die Zuständen einzelner Neuronen und deren Verbindungen entsprechen, verstehen nicht, was sie produzieren. Daraus folgt die Notwendigkeit, auf die Ebene *neuronaler Ensembles* überzugehen. Der soeben durchgeführte Vergleich mit menschlichen neuronalen Netzen bestätigt diese Notwendigkeit.

Zusätzlich zeigt der Vergleich aber auch, dass diese Ebene *selbständig* sein muss, mit anderen Worten: ihre Dynamik darf *nicht* die logische Folge der neuronalen Schicht sein. Nur unter dieser Voraussetzung kann sie als *kausale Ebene* des Netzes aufgefasst werden, und nur dann kann behauptet werden, dass das System fähig ist, zu *denken*.

Kann diese Bedingung auf Basis der gegenwärtigen Hardware eingehalten werden? Wie es scheint, gilt hier nach wie vor, dass jeder Zustand aus dem vorhergehenden Zustand folgt. In einer solchen

29 Ich betrachte es als eine Groteske der Geistesgeschichte, dass diese Tatsache weder von der Philosophie noch von der Naturwissenschaft noch von der KI-Forschung beachtet – ja nicht einmal *wahrgenommen* wird. Seit den ersten französischen Materialisten im 18. Jahrhundert bis in die Gegenwart wird zwar von Physikalisten und Deterministen die Existenz der *Moral* bezweifelt, aber das *Denken* behält immer seine Selbständigkeit – ansonsten könnten sie ja gar nicht *argumentieren –,* obwohl es sich doch ganz offensichtlich genauso in Physik auflöst wie die Moral, falls es nicht *selbst* als kausal aufgefasst wird. Man muss sich entscheiden: die Kausalität liegt ***entweder*** im Denken ***oder*** in der Physik – beides zugleich ist nicht möglich.

logischen Struktur ist die Kausalität *von unten* vollständig, und daher kann es *über* der neuronalen Schicht keine selbständige Ebene von neuronalen Ensembles mit eigener Dynamik geben.

Das würde bedeuten: In KI-Systemen, die auf solcher Hardware laufen, gibt es kein Denken und Verstehen – auch dann nicht, wenn diese Systeme geeignet sind, Attraktor-Netzwerke auszubilden.

Da aber konstruierte KI-Systeme in jedem Fall *empfindungslos* sind (siehe hier), und weil wir nun unter dieser Voraussetzung zeigen werden, dass sie auf *keiner Art von Hardware* dazu imstande sind, etwas *Neues* zu verstehen, können wir darauf verzichten, die kaum zu klärende Frage nach den Möglichkeiten zukünftiger Hardware zu beantworten.

Dies wird also der letzte Schritt unseres argumentativen Wegs sein: Um festzustellen, wie weit die Konsequenzen des Beweises der Empfindungslosigkeit von KI-Systemen tatsächlich reichen, werden wir nun voraussetzen, dass Hard- und Software keinen Einschränkungen mehr unterliegen, dass alles, was physikalisch möglich ist, auch machbar ist.

Der Unterschied zwischen dem konstruierten und dem natürlichen System wird aber nach wie vor darin bestehen, dass die Aktivität des natürlichen Systems *wesensgemäß* ist – dass sie also aus der *untrennbaren Einheit* von Substanz und Akzidenzien folgt und sich somit *von selbst* entfaltet –, während das konstruierte System auf *zugeführte Aktivität* angewiesen ist. Gemäß unserem Beweis bedeutet das, dass das biologische System *empfindungsfähig* ist und das künstliche System *nicht* (siehe hier).

Damit gibt es nur noch eine einzige Frage:

Was bedeutet das Fehlen von Empfindung?

Wir beginnen mit der Betrachtung eines LLMs. Wir haben festgestellt, dass LLMs nicht verstehen, *wovon* sie reden. Als Grund dafür, dass sie dennoch vernünftig *erscheinen,* haben wir Folgendes bestimmt: Während ihrer Trainingsphase ist – über die Statistik der Verteilung der Wörter – auch die *semantische Struktur* der Sprache in ihre Berechnung der Wahrscheinlichkeit des nächsten Wortes mit eingegangen. Diese Struktur ist also an der *Steuerung des Lernprozesses* beteiligt, aber sie wird nie *ins System selbst* integriert – das System *kennt sie nicht*.

Diese abstrakte Art der Erklärung war notwendig, weil gezeigt werden musste, warum es dem KI-System möglich ist, verständig zu *erscheinen*, ohne es zu *sein*. Da das aber nun erledigt ist, kann das Fehlen von Verständnis auch auf ganz einfache Weise erklärt werden:

Das LLM bleibt im Kreis der Wörter *gefangen* – jedes Wort wird durch andere Wörter definiert, aber es weiß von keinem einzigen Wort, was es *bedeutet*. Keines der verwendeten Wörter bezieht sich auf irgendetwas *außerhalb* der Sprache, oder sagen wir:

Keines der Wörter hat eine Verbindung zur wirklichen Welt.

Kann diese Begrenzung überwunden werden? Wie es scheint, ist der einfachste Weg dahin, dem KI-System nicht nur Wörter und Sätze, sondern auch Bilder und Videos zu präsentieren, d.h. die sprachlichen Tokens durch visuelle Tokens zu ergänzen.

Allerdings ist zu beachten, dass das KI-System ja nichts *wahrnimmt:* dass es also keine "Bilder" sieht wie wir, sondern nur Pixelhaufen. Ist das nun schon "die wirkliche Welt"? Jedenfalls nicht in dem Sinn, wie sie es *für uns* ist: als *Bild, das wir erleben und das voller Bedeutung ist*, kurz gesagt: nicht als ***Erfahrung*** der Welt.

Alles, was behauptet werden kann, ist Folgendes:

Wenn die sprachlichen und visuellen Datenmengen groß genug sind, wird das KI-System nach einer Trainingsphase gewisse Pixelhaufen und Buchstabenhaufen einander zuordnen können und dadurch

imstande sein, sinnvoll erscheinende Kombinationen beider zu produzieren: Bilder mit bestimmtem Inhalt, Comics, Videos mit Text und Ähnliches. Es wird sogar *Neues* produzieren können, allerdings nur in dem eingeschränkten Sinn, dass schon vorhandene Elemente auf neue Weise kombiniert werden, und nichts *fundamental* Neues.

Versteht das System *jetzt* irgendetwas?

Mit Sicherheit nicht – es gilt ja nach wie vor, was wir über GPTs *ohne* visuellen Input abgeleitet haben: die semantische Struktur ist nicht *explizit* im System vorhanden. Genauso, wie das System zuvor nicht wusste, was das bedeutet, worüber es *spricht*, kennt es auch jetzt nicht die Bedeutung dessen, was es *produziert*.

Wir müssen ihm also weitere Fähigkeiten zuerkennen, und zwar diejenigen Fähigkeiten, von denen wir wissen, dass sie notwendige Bedingungen für *Verstehen* sind:

1. Das Verstehen eines Sachverhalts kann entweder durch *Vergleich* mit einem anderen Sachverhalt oder durch *Einordnung* unter ein übergeordnetes Prinzip erreicht werden. Das KI-System muss also in jedem Fall mindestens über eines von beiden verfügen.

2. Dafür benötigt das KI-System ein Arbeits- und ein Langzeit-Gedächtnis, die zugleich aktiv sind.

3. Um die möglichen Gründe (und Ziele) eines Vorgangs zu bestimmen, muss das KI-System *denken* können. Bedingung dafür ist – wie wir oben ausgeführt haben – dass im System ein *Netzwerk von neuronalen Attraktoren* existiert, das als *kausale Ebene* des Systems aufgefasst werden kann.

4. Die Datenbasis des KI-Systems muss die Gesetze der Logik enthalten.

Ist das System *unter diesen Voraussetzungen* fähig zu verstehen?

Jedenfalls ist es fähig, zu verallgemeinern – aus folgendem Grund:

Ein Attraktor hat ein Einzugsgebiet. Er wird also nicht nur durch die *exakte* Wiederholung desjenigen sinnlichen Inputs aktiviert, durch den er entstanden ist, sondern auch durch jeden anderen Input, der dem originalen Input *hinreichend ähnlich* ist, um im Einzugsgebiet des Attraktors zu liegen.

Ein Beispiel: Wenn ein Kind zum ersten Mal das Bild einer Giraffe sieht, dann erkennt es später nicht nur die Giraffe auf diesem Bild, sondern auch alle auf anderen Bildern dargestellten Giraffen. Es ist also im Besitz des Allgemeinen, unter dem alle Exemplare subsumiert sind (während GPTs erst nach einem Training an einer großen Zahl von Bildern Giraffen erkennen, aber sogar dann noch immer nicht über *dieses Allgemeine selbst* verfügen).

Diese Tatsache ist nur auf eine einzige Weise erklärbar: Das neuronale Aktivierungsmuster, das sich als Folge des erstmaligen Betrachtens der Giraffe ausbildet, *wird sofort zum Attraktor*, der somit die allgemeine "Giraffe" repräsentiert. Er wird bei jeder Wahrnehmung einer Giraffe aktiviert und sorgt für das Wiedererkennen. (Das zugehörige Wort *Giraffe* ist in fast allen Fällen von Anfang an damit verbunden.)

Um beurteilen zu können, ob diese Art der Verallgemeinerung zum *Verstehen* führt, das ja immer auch Einsicht in die jeweilige Kausalstruktur einschließt, benötigen wir jedoch ein anderes Beispiel – eines, das mit einer Kausalstruktur und daher auch mit einem *Prinzip* oder *Gesetz* verbunden ist.

Dazu betrachten wir einen fallenden Stein. Das Attraktor-fähige KI-System wird dazu imstande sein, das Allgemeine über allen fallenden Steinen zu bilden. Wird es auch das verursachende *Prinzip* erkennen?

Das wird es nicht: die einfache Art der Verallgemeinerung, zu der es durch den Attraktor befähigt ist, führt über den "allgemeinen Fall eines Steins" *nicht* auf dessen *Gesetz*.

Es ist also (gemäß Punkt 1) auf den Vergleich mit einem anderen Sachverhalt angewiesen, dessen Kausalstruktur es bereits kennt.

Das könnte z.B. ein Sachverhalt sein, in dem sich zwei Gegenstände einander annähern, weil irgendetwas sie *zueinander zieht.*

Wenn wir dann noch annehmen, dass das KI-System über Messdaten des Verlaufs sehr vieler fallender Steine verfügt, dann wäre (äußerstenfalls) die Newtonsche Gravitation erreichbar – was allerdings bereits eine *sehr* optimistische Sicht ist, weil ja auch *Reibung* einbezogen und überdies auch die Erde *als "bewegter Gegenstand"* aufgefasst werden müsste.

Die Newtonsche Beschreibung der Gravitation ist allerdings nur eine Näherung, und ich sehe keine Möglichkeit, wie das System auf Basis seiner elementaren Fähigkeit der Verallgemeinerung zur Erkenntnis der allgemeinen Relativitätstheorie fortschreiten könnte, weder durch Vergleich noch durch ein übergeordnetes Prinzip, und auch nicht durch eigenes Nachdenken, weil diese Art der Verallgemeinerung zur Schaffung *neuer* Begriffe und Zusammenhänge nicht ausreicht.

Was hat sich eigentlich dadurch geändert, dass wir dem KI-System diese neuen Fähigkeiten – die notwendigen Bedingungen für Verstehen – zuerkannt haben?

Im Grunde nur dies:

Während es *vorher* – bei GPTs oder anderen selbstlernenden neuronalen Netzen – aufgrund des vollständigen Mangels an Verstehen erforderlich war, eine *komplette Liste aller Sachverhalte* zu erstellen, deren Kausalstruktur untereinander übertragbar ist (man denke an Yejin Chois Beispiel mit den "five clothes"), besteht *jetzt* auch die Möglichkeit, dem KI-System – da es ja bereits über Allgemeinbegriffe verfügt – einen *vollständigen Katalog aller allgemein gültigen Gesetze* samt einer *Definition der zugehörigen Sachverhalte* hinzuzufügen; Die erste Aufgabe wäre offensichtlich absurd, aber die zweite Aufgabe wäre möglicherweise durchführbar.

Daraus folgt jedoch:

Das KI-System ***versteht*** *einen Sachverhalt nur dann, wenn es das übergeordnete zugehörige Prinzip oder einen vergleichbaren zugehörigen Sachverhalt* ***bereits kennt***.

Mit anderen Worten:

Das System ist unfähig, <u>neue</u> (fundamentale) Theorien zu produzieren.

Soviel zur Einschätzung der Fähigkeit künftiger KI-Systeme, etwas zu *verstehen*, wenn alle Einschränkungen der Hardware aufgehoben sind, soweit es die Gesetze der Physik zulassen.

Nun aber zu uns selbst:

Wie verstehen wir? ***Verschafft uns die Fähigkeit zu empfinden dabei irgendeinen Vorteil?***

Die Antwort ist: ***Ja, das tut sie, und zwar in einem Ausmaß, das uns niemals bewusst wird.***

Wie erfahren wir die Welt?

Wenn ein Kind damit beginnt, die Welt zu erkunden, wird es dabei *vollständig* von Empfindung geleitet. Die ersten dabei aktiven Empfindungen sind [*angenehm – unangenehm*] und [*Begehren – Ablehnung*]. Aber auch wenn das Kind zunächst ausschließlich durch diese Empfindungen gesteuert wird, entsteht doch sofort jene *untrennbare Verbindung* von Empfindung

und Information, die wir als *geistigen Zustand* bestimmt haben, weil ja die empfindungsgesteuerte Handlung in jedem Fall mit Informationsgewinn verknüpft ist.

Auch später, wenn im Lauf des Heranwachsens der Informationsanteil zunimmt, bleibt aber Empfindung immer das treibende und steuernde Element.[30]

Der im Rahmen unserer Betrachtung entscheidende Sachverhalt ist jedoch dieser:

*Obwohl Empfindung **nicht definierbar** ist, ist jede Empfindung **ein Allgemeines***.

Betrachten wir als Beispiel wieder eine Farb-Empfindung: Die Empfindung *grün* ist zwar nicht definierbar, aber alle Ereignisse, die die Empfindung *grün* auslösen, können ihr zugeordnet werden.

Der Grad der Allgemeinheit von *Empfindungen* ist außerordentlich hoch. Im Fall der oben genannten Empfindung [*angenehm – unangenehm*] ist er sogar dem Grad der Allgemeinheit der Spitze der Pyramide *logischer* Verallgemeinerung vergleichbar:

Beim logischen Fortschreiten zum Allgemeinen hin landet man zuletzt beim Allgemeinsten: dem *reinen Sein*, das *alles Seiende* beinhaltet.

Das gleiche gilt aber auch von der Empfindung [*angenehm – unangenehm*]: jedes überhaupt mögliche erfahrbare Ereignis lässt sich dieser Empfindung zuordnen, und das trifft sogar auf Ereignisse zu, die nicht existieren, sondern nur denkbar oder vorstellbar sind.

Im Gegensatz zum ***logisch*** Allgemeinsten, das inhaltlich *vollkommen leer* ist, da ihm ja jede Eigenschaft fehlt, ist aber dieses ***qualitativ*** Allgemeinste keineswegs leer: es enthält genau jene Ereignisse, die es ausgelöst haben: wenn sie einmal erfahren worden sind, bleiben sie mit der Empfindung dauerhaft verbunden, und *potentiell* enthält es die unendliche Vielfalt von Ereignissen, die es auslösen *könnten*.

Andere Empfindungen bzw. Qualitäten, wie [*warm – kalt*], oder [*trocken – feucht*], sind hinsichtlich ihrer Verbindung mit Information deutlich spezifischer, haben aber immer noch einen hohen Grad von Allgemeinheit.

Es ist zu beachten, dass dies keine Allgemeinheit gemäß der üblichen, *logischen* Definition ist: in *diesem* Sinn bleiben Empfindungen immer leer, da sie ja nicht definierbar sind und somit keinen logischen Gehalt (Informationsgehalt) haben können.

Was bedeutet es nun, dass diese Art des Allgemeinen – das ***qualitativ Allgemeine*** – bei unserer Erfahrung und bei der Entwicklung der Beziehung zur Welt eine derart dominierende Rolle spielt?

Man stelle sich einen Raum vor, dessen Koordinaten *menschlichen Empfindungen* entsprechen.[31]

Am Anfang unseres Lebens sind unsere Erlebnis-Zustände (die noch keine geistigen Zustände sind) Vektoren in diesem Raum, aber nur für sehr kurze Zeit, denn – wie oben erwähnt – führt jede empfindungsgesteuerte Handlung zu einer Erfahrung, die Information enthält.

Der Raum unserer Erlebnis-Zustände verändert sich also fortwährend: die Zahl seiner Dimensionen nimmt permanent zu, weil neue, *informationstragende* Koordinaten hinzukommen: *Erlebnis-Zustände* werden zu *geistigen Zuständen*.

---

30 Zur Erinnerung: In unserer ontologischen Analyse (Teil 2, Abschnitt 2.2) haben wir Empfindung als *Substanz* des Geistes und damit zugleich als dasjenige bestimmt, was *Ursache* der Dynamik des neuronalen Netzes ist.

31 Die Stärke der Empfindungen ist zwar nicht direkt messbar, lässt sich aber doch aus den zugehörigen physiologischen Reaktionen abschätzen.

Der Raum geistiger Zustände entfaltet sich immer weiter. Empfindung und Information gehen komplexe Verbindungen ein. Die Empfindung [*angenehm – unangenehm*], die anfangs vor allem triebgesteuert war, verbindet sich zunehmend mit Sachverhalten und Zielen. Es entsteht *Intentionalität.*

Da wir zu logischen Verallgemeinerungen und Schlussfolgerungen fähig sind, gibt es in diesem Raum auch Wege, deren Verlauf rein logisch bestimmt ist – sie bilden jedoch die Ausnahme. Meist halten wir uns in Bereichen auf, die ebenso durch *Empfindung* strukturiert sind wie durch *Logik* und *Information*.

Dies sind die Bereiche der Phantasie, der Kunst, aber auch die Bereiche des Ausprobierens und Rätsellösens.

Die Denkweisen und Verhaltensstrategien, die sich durch das Zusammenwirken von Empfindung und Information herausbilden, sind zufälligem Verhalten weit überlegen, weil sie sich ja permanent – im täglichen Leben und Überleben – bewähren müssen.

Durch diese Betrachtungsweise wird nicht nur klar, worin der Unterschied zwischen unserem Denken und dem Denken von KI-Systemen besteht, sondern auch, welchen entscheidenden und uneinholbaren Vorteil uns die Fähigkeit zu empfinden verschafft:

Nur dann, wenn Denken in einem Raum der soeben skizzierten Art stattfindet, kann *Neues* produziert werden, und nur ein System mit einer solchen Art von Denken ist imstande, *neue Sachverhalte* zu integrieren und auf sie adäquat zu reagieren.

Beides ist im Grunde selbstverständlich: während in einem ausschließlich durch Logik und Wahrscheinlichkeit strukturierten Raum *Neues,* das "weit genug" *außerhalb* dieser Struktur liegt, weder erkannt noch produziert werden kann, ist ein Raum, dessen Struktur *auch* von Empfindung bestimmt wird, von dieser Beschränkung frei – das *qualitativ Allgemeine* enthält ja, wie oben dargestellt, nicht nur alles Existierende, sondern auch alles überhaupt Mögliche, Vorstellbare und Denkbare.

Mit anderen Worten:

Es gibt nichts, was *außerhalb* dieser aus Logik *und* Empfindung errichteten Struktur liegt.

Alles Neue kann integriert und verstanden, aber auch produziert werden.

Kurz zusammengefasst, lautet unser Ergebnis also wie folgt:

KI-Systeme werden auch in Zukunft nicht dazu imstande sein, Neues zu erkennen oder zu schaffen. [32]

Für uns gilt diese Beschränkung nicht: wir sind zu beidem fähig.

Wir können also nicht darauf hoffen, dass uns künftige superintelligente KI-Systeme "die Welt erklären" – das müssen wir auch weiterhin selbst versuchen.

Sie werden uns überhaupt nichts Wichtiges *erklären*, sondern nur genau das tun, was sie schon jetzt so gut können: in bekannten, endlichen Szenarien, deren Elemente und Übergänge definierbar sind, mögliche Strukturen und Zusammenhänge zu bestimmen – genauso, wie wir das schon durch das wunderbare Beispiel der Eiweißfaltung erfahren haben.

32 Mit Ausnahme des Neuen, das aus schon Vorhandenem besteht oder daraus ableitbar ist (wie der 37. Zug aus der zweiten Partie zwischen Lee Sidol und AlphaGo).

Zuletzt, noch einmal das Wichtigste

Fakten:

1. KI-Systeme können weder ***wahrnehmen*** noch ***fühlen*** noch ***wollen***.

2. Denken kann nicht der ***neuronalen Aktivität*** gleichgesetzt werden. Es muss auf der ***Ebene neuronaler Ensembles*** stattfinden.

3. Denken muss ***kausal*** sein – ansonsten ist es kein Denken. Eine notwendige Bedingung dafür ist, dass die physikalische Kausalität im System ***unvollständig*** ist. Somit ist auf Basis gegenwärtiger Hardware Denken ***ausgeschlossen***.

Einschränkungen:

1. ***Symbolic AI*** errichtet eine logische Struktur.

Die Welt ist ***nicht berechenbar*** – sie transzendiert jedes logische (mathematische) System. Dasselbe gilt für das Denken.

Also ist symbolic AI notwendig unvollständig, und darauf basierende KI-Systeme sind nur eingeschränkt fähig zu denken.

2. Die Leistung von ***lernfähigen neuronalen Netzen*** wird durch das Auffinden des Minimums einer (hochdimensionalen) Funktion optimiert, deren Wert der Abweichung vom Sollwert entspricht.

[*Auffinden des Minimums*] bedeutet [*Einbeziehen der strukturellen und formalen Bedingungen der angestrebten Leistung*].

Warum ist das möglich? Weil hinreichend große Datenmengen einen wesentlichen Teil dieser Bedingungen *enthalten*.

Denken und Verstehen werden also weder benötigt noch erzeugt. ***In neuronalen Netzen dieser Art existieren sie nicht***.

3. *Symbolic AI* und lernfähige neuronale Netze zu ***kombinieren*** kann die Leistung verbessern, ***die grundsätzlichen Beschränkungen bleiben jedoch bestehen***.

Heinz Heinzmann

Wien, August 2024