Warum Roboter nichts empfinden

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 informationsverarbeitendes 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 Informationsverarbeitung 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 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.[1]

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.

---

1 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.)

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.[2]

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 nicht durch denselben Antrieb verursacht wie die Dynamik des Ausgangssystems.

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 physikalische, die neuronale und die geistige Ebene. In meiner Arbeit *Die Begründung der Willensfreiheit* 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.[3]

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:

---

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

3 *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.

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.[4] **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.

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.

Heinz Heinzmann

August 2021

4 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.)