Wolfgang Marx

Der simulierte Mensch aus der Sicht der Kognitions-psychologie

Vortrag, gehalten am 18. 9. 2021 am Kolloquium der Schweizerischen Gesellschaft für Symbolforschung

Das mit Abstand Interessanteste, was uns auf diesem Planeten begegnet, ist der Mensch. Da das aber auch zugleich das Komplexeste ist, was in der Welt vorhanden ist, erweist sich gerade das – wie eine klassische Formulierung lautet – als das Schwerste von allen Dingen: sich selbst kennen.

Wie kann das gehen?

Etwas verstehen, heisst erklären können, wie es funktioniert. Verstanden haben wir etwas, wenn wir uns ein Modell davon machen, es gewissermassen „nachbauen“ können. Das beginnt mit Vermutungen, Spekulationen, Hypothesen, die sich, wenn es gut gelaufen ist, am Ende zu einem theoretischen Modell verdichten. Bei dem kann es aber nicht bleiben; denn so plausibel ein solches Konstrukt auch erscheinen mag, besonders dem Konstruktör, in letzter Konsequenz ist der Beweis der Theorie die Praxis. Ein Modell muss im Praxis-Test so funktionieren wie das Original, das es vorgibt zu simulieren. Eine Erklärung impliziert also wissenschaftstheoretisch gesehen immer auch eine Vorhersage, deren Eintreffen überprüft werden kann und deren Nichteintreffen die Erklärung falsch oder doch zumindest unvollständig macht.

Bevor ein theoretisches Modell jedoch überhaupt getestet werden kann, muss es erst einmal realisiert werden, entweder analog als Gebilde aus Metall, Holz oder Pappmaché oder digital als Computerprogramm. Solche Modelle können ein Original unterschiedlich vollständig abbilden. Bei komplexen Gegeben-

heiten werden in der Regel jeweils nur diejenigen Aspekte dargestellt, die für eine bestimmte Fragestellung oder eine konkrete praktische Nutzung relevant erscheinen. Man spricht dann von „homomorphen Modellen“. Ein „isomorphes Modell“ liegt vor, wenn ein Original in jeder Hinsicht abgebildet wird, sich also vom Original in nichts unterscheidet. In diesem Falle wäre das Artefakt gewissermassen ein zweites Original. Wenn ein Artefakt nicht mehr als solches erkannt werden kann, ist das der definitive Beweis für das Gelingen der Simulation. Das ist übrigens der Kerngedanke des berühmten Turing-Tests. Ich werde darauf noch zurückkommen.

Im Rahmen dieser Terminologie lässt sich jetzt die Frage nach einem „künstlichen Menschen“ formulieren als Frage nach der Möglichkeit den Menschen zu simulieren. – Den ganzen Menschen? Wohl eher nicht, obwohl ein isomorphes Modell natürlich das ultimative Ziel und die Vollendung einer solchen Simulation wäre. Ob so etwas überhaupt machbar ist, erscheint beim derzeitigen Stand unserer technischen Möglichkeiten wenig wahrscheinlich. Mag sein, dass es einmal da gelingen wird, wo sich die Parallelen schneiden...

An dieser Stelle könnte man jetzt mit den Möglichkeiten der Reproduktions-medizin argumentieren. Nun ist es wohl wahr, dass ihre Techniken es ermöglichen, „künstlich“ Menschen zu „zeugen“, aber eben nicht „künstliche Menschen“ zu „erzeugen“. Ein Mensch, der durch eine künstliche Befruchtung auf die Welt gekommen ist, ist ja nicht das Ergebnis eines technischen Herstellungsprozesses, er ist – wie alle anderen Menschen auch – durch einen natürlichen Wachstumsprozess entstanden. Dass dieser Prozess nicht auf natürliche Weise in Gang gesetzt wurde, macht ihn nicht zu einem Artefakt.

Bleiben wir also beim Thema homomorphe Modelle, die den Menschen in bestimmten Hinsichten simulieren. Dann stellt sich die Frage, in welchen Hinsichten und warum gerade in diesen?. Zwei in diesem Zusammenhang relevante Aspekte sind schon angesprochen worden: Erkenntnisinteresse und praktischer Nutzen. Es versteht sich, dass diese beiden Dinge nicht

unabhängig voneinander sind. Das Interesse, etwas zu verstehen, ist nicht nur intrinsisch motiviert (das berühmte platonische Staunen), es speist sich zu einem guten Teil auch aus dem Wunsch, gezielt Einfluss nehmen zu können. Umgekehrt wird ein erfolgreiches Eingreifen in die Abläufe um so eher gelingen, je besser man verstanden hat, was da abläuft. Kurt Lewin, einer der Gründerväter der experimentellen Psychologie, hat das in einem gern zitierten Bonmot so formuliert: Nichts ist so praktisch wie eine gute Theorie.

Wann der Antrieb bei auf zwei Beinen gehenden Herrentieren auftauchte, ihresgleichen abzubilden, verliert sich im Dunkel der Vorgeschichte. Vermutlich war das noch nicht bei Australopithecus oder Homo erectus schon der Fall. Erste Zeugnisse jedenfalls finden sich aus der Steinzeit: Jagdszenen an Höhlenwänden, Figurinen üppiger Weiblichkeit, die bekannte Venus von Willendorf, solche Sachen. Man vermutet, dass dabei weniger ein Erkenntnisinteresse im Spiel war (zum Beispiel die Morphologie des weiblichen Körpers betreffend), sondern dass es den Künstlern vielmehr um praktischen Nutzen zu tun war: um die magische Sicherung von Jagderfolg und Fruchtbarkeit. Wenn diese Deutungen zu treffen, dann kam die bildende Kunst nicht als l'art pour l'art in die Welt, sondern als Mittel zum Zweck.

Wie auch immer, die Simulation des Menschen beginnt mit dem gemalten und mehr noch dem aus einem Klumpen Ton geformten oder aus einem Stein gemeisselten Körper. Simuliert wird der Mensch dabei in Hinblick auf seine Oberfläche, seine Morphologie. Die Bildhauerei der Griechen stellt dabei einen grandiosen Höhepunkt dar, die ultimative Vollendung – nicht, was das Niveau der Kunst, wohl aber das Niveau der Kunstfertigkeit angeht – wurde in der Phase des radikalen Realismus, auch Foto-Realismus, in den 70er Jahren des 20. Jahrhunderts erreicht.

Die mit gebrauchten Alltagsklamotten bekleideten lebensgrossen und in nuancierten Fleischfarben bemalten Skulpturen von Duane Hanson, die hinter ihrem Kübel hockende Putzfrau, der Zeitung lesende Mann, sind auf einen ersten flüchtigen Blick von „echten" Menschen nicht zu unterscheiden und auf den zweiten Blick auch nur dadurch, dass sie sich nicht bewegen. Besonders

krass ist die Aktskulptur eines jungen Liebespaars von John de Andrea, schonungslos dargestellt bis in die intimsten Partien und lebensecht bis zu jeder Pore, jedem Schamhaar. Da wird das Ende der Fahnenstange erreicht, mehr Oberflächengenauigkeit geht nicht.

Mehr l'art pour l'art, so scheint es, auch nicht; und doch – könnte nicht gerade eine so lebensechte Darstellung attraktiver Körper weniger zu einem interesselosen Wohlgefallen an einem Kunstwerk als vielmehr zu einem von sehr handfesten (nämlich sexuellen) Interessen geleiteten Wohlgefallen führen? Und gibt es so etwas überhaupt, ein „interesseloses Wohlgefallen"? Wenn Freud die Dinge richtig sieht, ist ein solches nicht vor allem aus sublimierter Libido gespeist, die sich an dieser Stelle wieder offen als ganz unsublimiertes Begehren zeigt?

Sie ahnen es vielleicht, die Rede ist von einem gewissen Pygmalion, der, wie uns Ovid erzählt, lange der Lagergenossin entbehrte; aber, so hexametert er fort, er bildete indessen geschickt ein erstaunliches Kunstwerk, ein elfenbeinernes Weib, die Gestalt einer wirklichen Jungfrau, man dächte, sie lebe. Dass es nur Kunst war, verdeckte die Kunst. Ein radikaler Realismus avant la lettre gewissermassen, obwohl ich zu bezweifeln wage, dass sich mit dem Ausgangsmaterial Elfenbein eine vergleichbar perfekte Darstellung der Oberfläche eines menschlichen Körpers erreichen lässt. Jedes einzelne Haar? So viel an Auflösungsvermögen gibt dieser Stoff nicht her.

Wie auch immer, es kommt, wie es kommen muss: Das Feuer des köstlichen Leibes durchflammte die Brust ihm, also dem Pygmalion, es beginnt ein langanhaltendes Balzverhalten, da tut einer wirklich alles, was ein verliebter Jüngling so tun kann, um das geliebte Wesen zu erfreuen und (ganz buchstäblich) zu erweichen, er fühlt da nämlich gelegentlich immer wieder einmal nach, ob sich in der Hinsicht doch etwas tut. Häufig betasten die Hände das Werk, erzählt Ovid und malt das ganze Gewese (wie billig für den Verfasser einer „Ars armandi") ausführlich und mit erkennbarer Freude an all den zahllosen Einzelheiten aus.

In der nüchternen Terminologie Freuds gesprochen, stellt sich die Sache so dar: ein Fetisch wird zum Triebobjekt, doch leider vergebens, das eigentliche Triebziel kann nicht erreicht werden; denn – das jetzt technisch formuliert – es besteht in dieser Hinsicht keine funktionale Äquivalenz zwischen der Statue und einem lebendigen Leib. Platt gesagt: die Statue ist zum Vollzug des Geschlechtsverkehrs nicht geeignet. Hier kann nur eine dea ex machina helfen, kann Venus ein Wunder bewirken und die Statue beleben, sie zu einer wirklichen Frau machen. Folgerichtig heisst es dann weiter: und nachdem neunmal sich die Sicheln des Mondes zum völligen Kreis gerundet haben, bringt sie Paphos zur Welt.

An dieser Stelle fallen mir immer die ein wenig rätselhaften Verse von Gottfried Benn ein:

„...die erst von Händen berührten,
doch dann den Händen entführten
Statuen bergen die Saat.“

Aber ob Benn das gemeint hat? – Nun ja, man kann und darf sich dabei allerlei denken; denn die Gedanken sind bekanntlich frei. Bei mir kommen sie auf den Turing-Test zurück. Diese Schlusspointe hat nämlich durchaus ihre Logik: sie beweist, dass hier wirklich ein isomorphes Modell vorliegt. Die durch eine magische Technik geschaffene Frau ist in nichts von einer natur-gewordenen Frau zu unterscheiden. Um ein bekanntes Bonmot abzuwandeln, in dem ein Pudding an dieser Stelle zum Exempel dient: Der Beweis der Frau ist ihre Schwängerung.

Die Wortmarke „magische Technik“ habe ich übrigens nicht ohne Hinter-gedanken verwendet; denn was den Alten ihre Magie, ist den modernen Zeiten die Technik. Da findet sich eine sehr viel trivialere Lösung des Pygmalion-Problems, zwar nicht in der Luxusausführung eines isomorphen Modells, aber doch immerhin in einer massenproduktionstauglichen und so für jedermann erschwinglichen homomorphen Modellausführung, in der Billigvariante sogar aufblasbar, die freilich nur den Spass an der Freude bietet

und – wie man hört – durchaus auch differenziertes Balzverhalten auszulösen vermag, sogar Gefühle, heisst es, aber eben nicht einen Reproduktionserfolg ermöglicht, obwohl ein solcher von den Usern vermutlich auch gar nicht angestrebt wird. Das alles ist reichlich geschmacklos, zugegeben; aber auch solche Früchtchen bringt die Wissenschaft hervor: c'est la vie in modern times...

Bleibt am Ende noch die spannende Frage nach der Motivation, die Frage, warum ein Wer statt mit einer Wem mit einem Was schläft. Auch an dieser Stelle hilft die mechanistische Betrachtungsweise Freuds weiter, die Rede von einem Triebobjekt, auf das sich das Begehren richtet. Diese Wortmarke suggeriert von vornherein eine gewisse Bandbreite von Möglichkeiten, auch jenseits der Grenzen des Biologischen.

Aber nicht nur in Sachen Lust und Liebe, auch in Sachen Arbeit und Mühe können Dienstleistungen von selbstverfertigtem Menschenersatz von Interesse sein. Wir kommen zum Thema praktischer Nutzen und zur Sage vom Golem, die auf diesem Feld die Rolle übernimmt, die der Fall Pygmalion in der Leib- und Liebessache gespielt hat, nämlich die des klassischen Einstiegsparadigmas.

Hier ist jetzt Schluss mit lustig, schliesslich geht es jetzt um den Ernst des Lebens, nicht ums Vergnügen. Da ist keine Romantik im Spiel, auch kein Luxus. Das beginnt schon damit, dass ein wenig edles und billig zu habendes Ausgangsmaterial zu seiner Herstellung verwendet wird, nicht aus Elfenbein, aus einem Klumpen Lehm wird der Golem gefertigt, er kann auch nicht sprechen und ist nur annähernd menschenähnlich, eine grobe und ungeschlachte Gestalt, so heisst es; aber es soll sich ja auch niemand in ihn verlieben, er soll nur den Hausknecht geben, soll Stiefel ausziehen und den Boden putzen, für solche Sachen reicht es allemal.

So unscheinbar das Modell Golem auch daherkommt, so ist es doch ein wahres Wunderwerk, es spottet den physikalischen Gesetzen von der Erhaltung der Energie und der Materie. Mit ihm kommt kein weiteres Maul in

den Haushalt, das gestopft werden muss, er benötigt keine Nahrung, mit anderen Worten: hier ist die Konstruktion eines perfekten Perpetuum mobile gelungen, einmal in Gang gesetzt, läuft er ohne jede Energiezufuhr endlos weiter. Von so etwas können wir im technischen Zeitalter nur träumen.

Obwohl – wirklich endlos? – Darauf sollte man es keinesfalls ankommen lassen; denn – ich komme jetzt zum Gesetz von der Erhaltung der Masse – das tönerne Ding wächst im Laufe der Zeit unablässig, es entzieht sich mehr und mehr der Kontrolle und kann für seinen Herrn und dessen Haus zu einer echten Gefahr werden. Man muss es beizeiten abschalten, solange man den Ausknopf noch erreichen kann. Der ist dem Golem sozusagen auf die Stirn geschrieben in Form des Wortes „'æmæt“ (was „Wahrheit“ bedeutet, „Beständigkeit“, „Gott“, etwas in der Art). Löscht man das Aleph ('æ) aus, bleibt noch die Silbe „mēt“ („tot“, „Leichnam“) , worauf der Golem wieder zu einem Lehmklumpen zerfällt, der aber jetzt sehr viel voluminöser ist, als er vor der Verwandlung war; und wenn es dumm läuft, seinen Herrn unter sich begräbt. Die Sage weiss von so einem tragischen Fall mit letalem Ausgang. Nun ja, die meisten Unfälle passieren bekanntlich im Haushalt.

Das Modell Golem macht von Anfang an deutlich, dass es bei dieser Sorte von Menschmaschinen nicht so sehr um morphologische Ähnlichkeit geht, sondern darum, dass der Mensch simuliert werden soll in Hinblick auf bestimmte Funktionen. Es geht nicht darum, dass er möglichst genau wie ein Mensch aussieht – in der Logik des Turing-Tests gesprochen: dass er mit einem Menschen verwechselt werden könnte – es geht vielmehr darum, dass er machen kann, was ein Mensch macht, nicht gerade alles, aber doch wenigstens die Drecksarbeit, die soll er machen.

Angeregt von der Golem-Sage entwickelte sich eine literarische Tradition arbeitender Maschinen als Untergenre der Science Fiction. Eine zentrale Rolle spielt dabei das Theaterstück „R.U.R.“ (das steht für „Rossums Universal Robots“) von Karel Čapek. Hier taucht erstmals der Begriff des Roboters auf, eine Ableitung vom tschechischen „robota“ (Arbeit, speziell auch Fronarbeit). Was die Firma Rossum massenhaft produziert und billig auf

den Markt wirft, sind gewissermassen „mechanische Golems“, universell einsetzbar bei Arbeiten aller Art, in der Gestalt menschenähnlich und mit deutlich mehr Intelligenz (künstlicher Intelligent, nota bene) ausgestattet als ihr lehmgezeugtes Vorgängermodell. Die Geschichte geht übrigens nicht gut aus, die mechanischen Sklaven rebellieren und vernichten die Menschheit. Nun gut, das ist Science Fiction, die immer dann zu ihrer höchsten Form aufläuft, wenn die Dinge dabei sind, die schlimmstmögliche Wendung zu nehmen. An diesem Garn soll später noch einmal ein wenig weitergesponnen werden. Die tatsächliche Entwicklung ging andere Wege, aus technischen und ökonomischen Gründen.

In der Frühphase der Mechanisierung der Arbeit im grossen Stil, während der sogenannten „industriellen Revolution“, stand schlicht und einfach nicht einmal ansatzweise eine Künstliche Intelligenz zur Verfügung wie sie für die autonome Selbststeuerung eines derart komplexen Systems wie eines menschenähnlichen Roboters, also eines Androiden, erforderlich gewesen wäre. Zwar sind wir seit jenen Tagen von Russ, Qualm und Dampfmaschinen in dieser Sache ein gutes Stück weitergekommen; aber immer noch längst nicht so weit, einen veritablen Androiden auf zwei Beine zu stellen und zu einem halbwegs geschmeidigen Laufen zu bringen. Wenn Sie einmal Gelegenheit hatte, zuzuschauen bei einem Match von Fussball spielenden – nein, Androiden kann man die ungelenk über das Spielfeld staksenden Blechmännchen beim besten Willen nicht nennen – dann wird Ihnen aufgegangen sein, wie viel da noch fehlt, um einen Vergleich mit dem auszuhalten, was ein menschlicher Organismus bei der Gelegenheit leistet. Und das ist ja nur ein peripheres Problem, ein Kleindetail im grossen Getriebe.

Versuchen wir einmal nur grob zu überschlagen, was ein künstliches zentrales Nervensystem alles zu steuern und aufeinander abzustimmen hätte, allein an Bewegungssteuerung in Einklang mit der Wahrnehmung dessen, was bewegt und bearbeitet werden soll, an der Bewertung dessen, was geleistet wurde, also einer Qualitätskontrolle, an der eventuellen Einleitung von Korrekturmassnahmen nach Massstäben, die schon vorher festgelegt

worden sein müssen, wobei ökonomische, gegebenenfalls sogar ethische Standards zu beachten wären. Darüber hinaus gilt es, ständig zumindest im Augenwinkel zu behalten, was sonst noch im Arbeitsumfeld geschieht, es einzuordnen, um gegebenenfalls angemessen darauf reagieren zu können, zu wissen, was angemessen ist – und bei alledem auch noch ständig dieses labile System aufrecht und auf seinen zwei Beinen im Gleichgewicht zu halten, um damit wieder zum Ausgangspunkt unserer Überlegungen zurückzukehren.

Was für ein gigantischer Aufwand für ein Gerät, das die Wäsche waschen, das Geschirr spülen und staubsaugen oder ein paar einfache Handgriffe am Fliessband machen soll. Wieviel Rechenleistung und Energie würden da schon für die blosse Selbstverwaltung des Systems verbraucht werden statt in die eigentliche Arbeit zu fliessen? So etwas rechnet sich nicht. Der Android ist kein sinnvolles Modell für ein Arbeitsgerät. Wenn er überhaupt jemals gebaut wird, dann aus denselben Gründen, aus denen jemand den Mount Everest besteigt. Weil der Berg eben da ist. Eine Zukunft als Industriearbeiter oder als Haushaltshilfe hätte dieses Modell sicher nicht; und auch die Vernichtung der Menschheit stünde wohl kaum auf seinem Programmzettel.

Verständlich also, dass von Anfang an ein anderer Weg eingeschlagen wurde, nämlich der, bestimmte Teilsysteme abzugrenzen, ihren Aufbau und ihre Funktionen zu analysieren, um sie dann in Form von mechanischen Modellen nachzubauen. Am Ende einer langen Entwicklung stehen die Industrieroboter am Fliessband, die jeweils nur einige wenige „Handgriffe“ ausführen, diese aber mit gleichbleibender Präzision und unermüdlich. Diese bizarren Geräte haben keine auch nur entfernte Ähnlichkeit mit einem Menschen mehr. Bei dieser Teilsimulation des Menschen geht es aber auch gar nicht um morphologische Ähnlichkeit, sondern um funktionale Äquivalenz.

Nun stellt sich die Frage der funktionalen Äquivalenz nicht nur in Hinsicht auf das, was sich mit Händen, Armen und Beinen anstellen lässt, sie stellt sich auch in Hinblick auf das, was wir mit dem Kopf, beziehungsweise in ihm, zu machen verstehen. Es geht noch einmal um die Frage der Künstlichen

Intelligenz, die bisher nur am Rande gestreift wurde, um die Frage, ob überhaupt und wenn ja, wie gut es gelingen kann, kognitive Funktionen des Menschen zu simulieren.

Im Kontext dieser Frage wurde übrigens der schon mehrfach angesprochene Turing-Test konzipiert, hier liegt gewissermassen seine Kernkompetenz. Der Grundgedanke von Alain Turing war erstaunlich einfach und doch sehr effektiv, wie viele geniale Ideen; aber eben, man muss erst einmal auf so etwas kommen: Wenn wir bei der Kommunikation mit einem Partner, den wir nicht sehen, nicht mehr unterscheiden können, ob wir uns mit einer natürlichen oder mit einer Künstlichen Intelligenz auseinandersetzen, dann ist die Simulation gelungen.

An dieser Stelle lohnt es sich, einen Augenblick darüber nachzudenken, was denn da simuliert werden soll, was das ist, das da in uns denkt, sich erinnert, wahrnimmt, Entscheidungen trifft, etwas weiss, etwas glaubt, Sehnsucht verspürt oder Angst, Befürchtungen hat oder Hoffnungen. Was für ein Ding ist das, dieses denkende Ding? Woraus ist es gemacht? Und wie ist es in unsere Köpfe hineingekommen? – Dumme Fragen, so muss es von unserem heutigen Wissensstand aus erscheinen; denn das weiss doch jeder: das denkende Ding ist das Gehirn, ein Fleisch, das gewissermassen Wort werden kann. Nur, das ist eine relativ späte Einsicht.

2'000 Jahre lang, von Platon bis Descartes, von der „Psyche“ bis zur „res cogitans“, hatte man von dieser Sache im Abendland ganz andere Vorstellungen, Ideen von zwei inkommensurablen Welten, einer Welt ausgedehnter (materieller) Dinge und einer Welt nicht ausgedehnter (immaterieller) Dinge; und zu dieser zweiten Welt gehört das denkende Ding, die res cogitans. Vor diesem Hintergrund erscheinen die gestellten Fragen gar nicht mehr dumm, sondern geradezu notwendig. Es muss klargestellt werden, dass das denkende Ding nicht aus Materie gemacht (also kein Fleisch) ist, sondern aus einer immateriellen Substanz besteht.

Wer eine solche steile These vertritt, muss dann schon ein paar Dinge erklären, zum Beispiel, wie so etwas in unsere Köpfe hineinkommt, warum überhaupt und wie es möglich sein soll, dass zwei doch inkommensurable Substanzen miteinander interagieren können, was ja ein Widerspruch in sich wäre, eine contradictio in adjecto. Schwierige Probleme sind das wohl; zwar nicht jenseits aller Mutmassungen, an solchen herrscht kein Mangel; aber doch jenseits aller rationalen Lösungsmöglichkeiten; denn – in der diskreten Terminologie Kants gesprochen – handelt es sich dabei durchweg um synthetische Sätze a priori, die man bekanntlich nicht beweisen kann, also um metaphysische Spekulationen. Das ist die Art von Sätzen, die Wittgenstein später im „Tractatus“ auf sein uncharmant direkte Art als „unsinnig“ bezeichnen wird, Sätze, die man, so seine Empfehlung, überwinden müsse, um die Welt richtig zu sehen.

Wenn man versucht, die Welt richtig zu sehen, dann macht es in der Tat wenig Sinn, den kognitiven Apparat als etwas dem Organismus Wesensfremdes, ihm gewissermassen von aussen Aufgepfropftes zu betrachten und nicht als das, als was er sich bei genauerer Untersuchung und fortschreitendem Verständnis erweist: als ein wichtiges Organ, das ganz auf den Körper abgestimmt und vielfältig mit ihm vernetzt ist, das zwischen der Wahrnehmung, einerseits der Aussenwelt, andererseits der inneren Zustände des Organismus und den Effektoren vermittelt, die auf die inneren und äusseren Vorgänge zu reagieren haben, einerseits, um den inneren Zustand des Körpers im Gleichgewicht zu halten, andererseits, um sich in der Welt zu behaupten, zu überleben vor allem, dann aber auch, sich zu reproduzieren.

Löst man dieses universelle Steuerorgan aus dem Organismus heraus, wird er funktionslos, läuft leer. Wenn man es dann auch noch zu etwas Immateriellen erklärt und in alle Ewigkeit fortdauern lässt – um noch einmal einen kleinen Abstecher auf die öde Heide der metaphysischen Spekulationen zu machen – dann weiss man nicht recht, was es da eigentlich tun soll in alle Ewigkeit. Es wäre dort so gut aufgehoben wie ein Staubsauger in einer Welt, in der es keine Teppiche gibt, nicht einmal Staub.

Aber zurück zur Welt unterhalb des Mondes, der Welt in der wir tatsächlich leben und manchmal sogar gern auch wenn sie gelegentlich ein wenig streng riecht; und zurück zur Künstlichen Intelligenz, zur Simulation des kognitiven Apparats. Ob es einmal möglich sein wird, ein isomorphes Modell des zentralen Nervensystems zu bauen – wer kann behaupten, das zu wissen? Versuchen wird man es auf jeden Fall, da gibt es ein gewaltiges Erkenntnis-interesse: nichts ist für uns interessanter als gerade dieses Organ, definiert es doch, was der Mensch ist, was er anderes ist als der Rest der Fauna dieses Planeten.

Das ist freilich noch Zukunftsmusik – und folgerichtig ist die Sage, mit der das Thema eines menschenartigen Computers illustriert werden soll, auch in der Zukunft angesiedelt, es ist die Sage von der „Odyssee im Weltraum“, so der Titel des Films von Stanley Kubrick aus dem Jahre 1968, inzwischen ein Klassiker des Genres.

Auf den Plot möchte ich an dieser Stelle nicht weiter eingehen, die Ereignisse, die für unser Thema interessant sind, spielen sich an Bord eines Raumschiffs ab, das zu einer Jupiter-Expedition unterwegs ist. Dabei kommt es zu einem dramatischen Show-down zwischen zwei Astronauten und dem Bordcomputer HAL 9'000, der mit einer Künstlichen Intelligenz ausgestattet ist, die es ermöglicht, das Raumschiff autonom zu steuern.

Die Sache beginnt scheinbar harmlos damit, dass HAL den Ausfall eines wichtigen elektrischen Bauteils voraussagt, sich bei der Überprüfung durch die Astronauten jedoch herausstellt, dass das Teil einwandfrei funktioniert. Ein banaler Fehlalarm, könnte man meinen, nur, die Computer der Serie 9'000 gelten als absolut perfekt, als unfähig, Fehler zu machen, eine Tatsache, die zum Wissen von HAL über sich selber, über seine Möglich-keiten und Fähigkeiten gehört. In der Kognitionspsychologie nennte man das „metakognitives Wissen“. So etwas ist für einen Akteur notwendig, um sich selber als Einflussgrösse bei der Handlungsplanung mit einberechnen zu können.

Was also auf den ersten Blick wie ein banaler Fehlalarm aussieht, stellt für HAL ein unlösbares Dilemma dar: einerseits kann er keine Fehler machen, das weiss er, andererseits hat er einen Fehler gemacht, das sieht er. In der Fachterminologie nennt man so etwas eine „kognitive Dissonanz“, etwas, was bei HAL eigentlich gar nicht entstehen können sollte. Bei uns armen Sterblichen sieht das freilich ganz anders aus. Es kommt nicht nur vor, dass wir zu verschiedenen Zeiten in unterschiedlichen Kontexten logisch miteinander unverträgliche Argumente benutzen, es fällt uns meistens nicht einmal auf. Wenn wir eine solche Dissonanz aber doch einmal bemerken oder von anderen darauf hingewiesen werden, beeindruckt uns das in der Regel wenig. Wir haben im Laufe des Lebens gelernt, mit den Widersprüchen in unserem Kopf zu leben, sie zu ignorieren, sie zu bagatellisieren à la :“Was kümmert mich mein dummes Geschwätz von gestern“ oder sie schlicht zu leugnen, der gängige Spruch in diesem Falle lautet: „Das ist doch etwas ganz anderes“.

Über solche Coping-Strategien (so der psychologische Fachterminus) verfügt HAL jedoch nicht, sie sind in seiner kognitiven Ausstattung nicht vorgesehen, weil man davon überzeugt war, er würde so etwas nicht brauchen. Er ist also mit einem Problem konfrontiert, das für ihn unlösbar und zugleich unausweichlich ist. Er kann nicht – wie es in der blumigen Sprache der Motivationspsychologie heisst – „aus dem Felde gehen“. In dieser ausweglosen Lage reagiert HAL nicht anders als mit weit weniger Intelligenz begabte Lebewesen auch, wenn sie systematisch überfordert werden: mit einer Neurose. Er beginnt, ein unberechenbares Eigenleben zu entwickeln, was die Astronauten als bedrohlich empfinden und beschliessen, vorsichtshalber seine höheren kognitiven Funktionen abzuschalten. Diese Absicht bleibt HAL nicht verborgen, er ist jedoch nicht bereit, das zu akzeptieren; und das ist nun eine sehr spezielle Situation: normalerweise wehrt sich eine Maschine nicht dagegen, abgeschaltet zu werden.

An dieser Stelle lohnt sich ein kleiner Abstecher in die Existenzphilosophie zu Heideggers Charakterisierung des Menschen als eines Seienden, dem es um das eigene Sein geht. Das ist gewissermassen das Alleinstellungsmerkmal

des Menschen, das ihn von allem, belebt oder unbelebt, unterscheidet, was in der Welt vorhanden ist. Er kann sein Dasein reflektieren und um sein Dableiben besorgt sein. So gesehen funktioniert HAL wie ein Mensch; und das macht auch Sinn; denn wenn er, und sei es auch nur vorübergehend, die Steuerung des Raumschiffs und die Kontrolle über alle Vorgänge an Bord übernehmen können soll, muss er auch dazu in der Lage sein, alles zu tun, was notwendig ist, seine eigene Funktionsfähigkeit zu erhalten.

Aber wirklich alles? – Sicher nicht, die Astronauten umzubringen, um seine Abschaltung zu verhindern. Spätestens an dieser Stelle wird unübersehbar, dass er vom Pfad seiner Tugend, der durchgängigen Rationalität, abgewichen ist; denn die Prämisse aller seiner Aktivitäten muss sein: Das Forschungsprojekt darf nicht gefährdet werden. Nun kann er das Raumschiff zwar auch ohne die Astronauten zum Jupiter steuern, er kann aber dort nicht an ihrer Stelle von Bord gehen und die geplanten Untersuchungen durchführen. Die Begründung seines Angriffs auf die Astronauten, die Mission dürfe nicht gefährdet werden, ist also unsinnig; denn wenn die Forscher tot sind, kann die Mission gar nicht erfüllt werden.

Man könnte in diesem Falle von einem im nicht ganz eigentlichen Sinne des Wortes „menschlichen“ Versagen sprechen, wobei „Menschlichkeit“ hier bedeutet, ein das eigene Wohlbefinden betreffende Motiv in eine Sachentscheidung einfliessen zu lassen. So etwas kann gravierende Folgen haben; denn Rationalität existiert ja nicht an sich, sondern in Abhängigkeit von immer schon vorausgesetzten Werten eines Akteurs. Es versteht sich, dass eine rationale Entscheidung jeweils anders ausfallen muss, wenn die höchste Priorität lautet „Ich will das hier auf jeden Fall überleben“ oder „Die Mission darf auf keinen Fall gefährdet werden“. Wenn HAL also „menschlich“ reagiert und ein persönliches Motiv über die Interessen des Auftrags stellt, verhält er sich tatsächlich nicht irrational, irrational ist nur seine Argumentation, weil er ein Ziel zu verfolgen behauptet (den Schutz der Mission), sein Verhalten aber gar nicht zielführend ist. Auch das ist sehr „menschlich“, führt aber dazu, dass er als Arbeitsgerät versagt.

Sie kennen wahrscheinlich alle die Maxime Dürrenmatts, eine Geschichte sei erst dann zu Ende gedacht, wenn sie die schlimmstmögliche Wendung genommen habe. Das gilt in besonderem Masse für das Genre der Science Fiction, das, wie Sartre einmal angemerkt hat, die Angst des Menschen vor sich selber spiegele. Da wird gar nicht so sehr die „Schöne neue Welt“ thematisiert, Huxley verwendet diesen Titel mit bitterer Ironie, es wird vielmehr und viel öfter ausgemalt, wie bös das alles enden wird.

Unter diesem Unstern steht, so wird das von allem Anfang an gesehen, auch die Entwicklung einer Künstlichen Intelligenz. Schon ihr erster literarischer Auftritt bei Čapeks „Rossums Universal Robots“ endete, Sie werden sich daran erinnern, mit keiner geringeren Katastrophe als der Vernichtung der Menschheit. Dieser schwarze Faden zieht sich in der Folge durch manche Dystopie hindurch bis hin zu HAL 9'000; und gerade seine Geschichte zeigt exemplarisch auf, was wir an seinesgleichen fürchten, nämlich dass sich unser „Sündenfall“ bei ihnen wiederholen könnte, jetzt mit dem Menschen in der Rolle des düpierten Schöpfers. Die Sorge ist, sie könnten ebenfalls eine Frucht vom Baum der Erkenntnis essen, könnten zu Wissen über sich selber kommen, könnten werden wie wir, also autonom eigene Ziel verfolgen, die kaum die unseren sein dürften.

Nun werden Dystopien nicht geschrieben als Vorhersagen, sondern als Warnungen, gewissermassen als Prophezeiungen, die sich nicht erfüllen sollen. Die durchgängige Botschaft lautet, angefangen schon bei der Sage vom Golem: Wir dürfen auf keinen Fall die Kontrolle verlieren, vor allem nicht die über den Abschaltknopf.

Was uns gefährlich werden kann, ist also nicht ein weit übermenschliches kognitives Potential, das ist „wertneutral“. Algorithmen haben keine Interessen, es ist ihnen egal, ob sie dafür sorgen, dass eine Firma höhere Gewinne macht oder eine Person X einen Job bekommt und nicht eine Person Y. So soll es sein und so soll es bleiben. Wir dürfen nicht zulassen, dass eine Künstliche Intelligenz sich selber überlassen bleibt, sich um ihre eigenen Belange kümmert und damit eigene Motive verfolgen und weiter-

entwickeln kann. Um als Arbeitsgerät im Sinne des Auftrags, und nur in diesem, zu funktionieren, darf sie nicht vom rechten Wege abkommen, weil sie sich gerade um ihr eigenes Schicksal kümmert. In der Terminologie Heideggers gesprochen: sie darf an ihrem Dasein nicht interessiert sein, sich um ihr Dableiben nicht sorgen, mehr noch, die Tatsache ihrer eigenen Existenz sollte ihr verborgen bleiben, sie sollte kein metakognitives Wissen über sich selber erwerben können und auch keines über die Hardwarebasis ihres Funktionierens, sie sollte für sich selber ein blinder Fleck sein, sich selber nicht auf der Rechnung haben, mit einem Wort, sie sollte „reine“ Intelligenz sein.

Eine solche unterscheidet sich grundsätzlich von der menschlichen Intelligenz, die in einen Organismus eingebettet und mit diesem vielfältig vernetzt ist und mit ihm kontinuierlich kommuniziert, um seinen Bedürfnissen optimal dienen zu können. Der kognitive Apparat ist, ich wiederhole mich in diesem Punkt, ein Organ, das dem Organismus hilft zu überleben und sich zu reproduzieren, ist also immer Mittel zum Zweck, und er ist auf diesen Zweck hin dimensioniert, um nicht zu sagen „beschränkt“. Diese Begrenztheit unserer kognitiven Kapazität macht es (auch ökonomisch) interessant, „Intelligenz-Verstärker“ zu konzipieren, so wie die Begrenztheit unserer Muskelkraft dazu angeregt hat, „Kraft-Verstärker“ zu bauen. Dabei geht es nicht um eine Simulation des ganzen Menschen, sondern um eine technische Nachbildung eng umschriebener Funktionen; das spezifisch Menschliche, nämlich ein Seiendes zu sein, dem es um das eigene Dasein geht, bleibt aussen vor. Ein Industrieroboter oder ein Hochleistungsrechner haben nicht viel gemeinsam mit einem menschlichen Wesen, weder morphologisch noch kognitiv; und solange die Kraftmaschinen gar keine Intelligenz haben und die Denkmaschinen keine menschliche, solange ist die Stellung des Menschen als Endglied der Nahrungskette auf diesem Planeten nicht gefährdet.