Ansgar Beckermann

Der Computer – ein Modell des Geistes?

1. Spätestens seit dem Erscheinen von Turings Aufsatz „Computing Machinery and Intelligence" im Jahre 1950 ist die Idee, daß der Computer ein Modell des Geistes sein könne, aus der modernen Diskussion des Leib-Seele-Problems nicht mehr wegzudenken. Warum ist das so? Was sind die Gründe dafür, daß diese Idee in so kurzer Zeit eine so dominierende Stellung gewinnen konnte?

Wenn man dies verstehen will, ist es meiner Meinung nach sinnvoll, gut dreihundert Jahre zurückzublicken auf die Cartesianische Auffassung von Natur und Geist.[1] Descartes war als Dualist auf der einen Seite bekanntlich ein vehementer Vertreter der These, daß der Geist etwas ganz und gar Unkörperliches sei und sich daher grundsätzlich von allen physischen Dingen unterscheide. Auf der anderen Seite vertrat er jedoch mit ebenso großem Nachdruck die Auffassung, daß die gesamte nichtgeistige Natur bis hin zu den am höchsten entwickelten Tieren völlig nach den Prinzipien der Mechanik erklärt werden könne. Diese zweite These bedeutete einen fundamentalen Bruch mit der aristotelischen Tradition. Denn diese Tradition war durch die Grundannahme geprägt, daß die Eigenschaften und Fähigkeiten, die Lebewesen von unbelebten Dingen unterscheiden, auf keinen Fall auf die physischen Teile dieser Lebewesen und auf deren Anordnung zurückgeführt werden können. Die charakteristischen Fähigkeiten und das charakteristische Verhalten von Lebewesen können ihr zufolge nur durch die Annahme einer Seele erklärt werden, die jedoch anders als bei Descartes nicht als Substanz, sondern als Form, d. h. als organisierendes Prinzip aufgefaßt wurde.

Für Descartes waren Erklärungen durch Formen oder Entelechien jedoch wissenschaftlich unbefriedigend (um das mindeste zu sagen). Und mit dieser Einschätzung hatte er sicher recht. Seiner Meinung

[1] Vgl. zu diesem Abschnitt Beckermann (1989).

nach waren solche Erklärungen aber auch unnötig. Denn auf der Grundlage der zu Beginn der Neuzeit entstehenden neuen Naturwissenschaft war es ihm zufolge durchaus möglich, auch die für Lebewesen charakteristischen Vorgänge rein mechanisch zu erklären. D. h., Descartes war der Meinung, daß sich diese Vorgänge mit der gleichen Notwendigkeit aus den Teilen des menschlichen und tierischen Körpers ergeben, „wie der Mechanismus einer Uhr aus der Kraft, Lage und Gestalt ihrer Gewichte und Räder folgt" (*Discours*, 81 ff.). Und entsprechend versucht er, im *Traité de l'homme* und *La description du corps humain* für den Herzschlag, für die Ernährung, für die Wahrnehmung, für das Gedächtnis und schließlich sogar für die Fortpflanzung mechanische Erklärungen zu liefern, wobei er sich hauptsächlich an drei *Modellen* orientiert: am Modell der Uhr, deren Verhalten vollständig durch das mechanische Zusammenwirken ihrer Gewichte und Räder bestimmt wird; am Modell der Orgel, bei der Register und Tastenanschlag das Öffnen und Schließen der einzelnen Orgelpfeifen bewirken, und schließlich auch an den komplizierten hydraulischen Steuerungssystemen, mit denen die Gartenbaumeister seiner Zeit viele kleine Gartenfiguren zu einer Art von künstlichem Leben zu erwecken verstanden.[2]

Auch für Descartes gibt es jedoch eine *prinzipielle* Grenze für die mechanische Erklärbarkeit der Fähigkeiten von Lebewesen. Und diese Grenze liegt für ihn da, wo beim Menschen die Fähigkeiten des Denkens und Sprechens ins Spiel kommen. Im Teil V des *Discours de la méthode* erklärt Descartes ausdrücklich, daß sich Menschen seiner Meinung nach in zwei Punkten grundsätzlich von jeder Maschine, d. h. von jedem mechanischen System unterscheiden. Erstens nämlich könnten solche mechanischen Systeme „niemals Worte oder andere Zeichen dadurch gebrauchen, daß sie sie zusammenstellen, wie wir es tun, um anderen unsere Gedanken mitzuteilen". Und zweitens würden solche Systeme, auch wenn sie in einigen Punkten sehr gute Leistungen vollbrächten, „doch zweifellos bei vielem anderen versagen, wodurch offen zutage tritt, daß sie nicht aus Einsicht handeln, sondern nur aufgrund der Einrichtung ihrer Organe. Denn die Vernunft ist ein *Universalinstrument*, das bei allen Gelegenheiten zur Verfügung steht, während diese Organe für jede besondere Handlung

[2] Vgl. hierzu Specht (1966, 114 ff.).

einer besonderen Einrichtung bedürfen..." (*Discours*, 92 f. – Hervorh. vom Verf.).

Leider sagt Descartes sehr wenig darüber, warum es seiner Meinung nach für die Fähigkeiten des Denkens und Sprechens keine mechanischen Erklärungen geben kann. Aber es ist wohl besonders der im letzten Satz der gerade zitierten Passage angesprochene *universale Charakter* der Vernunft, der für ihn in diesem Zusammenhang ausschlaggebend war. Auf jeden Fall läßt Descartes keinen Zweifel daran, daß es sich bei den Fähigkeiten des Denkens und Sprechens seiner Meinung nach um im Rahmen einer mechanistischen Naturwissenschaft nicht erklärbare Phänomene handelt.

Schon Descartes' unmittelbare Nachfolger sahen in dieser Position jedoch eher eine Herausforderung. War es nicht vielleicht doch möglich, den *ganzen* Menschen mit all seinen Fähigkeiten mechanisch zu erklären? Besonders die Philosophen der französischen Aufklärung glaubten an die Möglichkeit einer positiven Antwort auf diese Frage. Der Titel des Buches *L'homme machine* (1748) von J. O. de La Mettrie spricht da eine deutliche Sprache. Alle Versuche, Descartes' Erklärungsansatz über dessen eigene Überlegungen hinauszutreiben, mußten jedoch notwendig im bloß Proklamatorischen steckenbleiben, solange es kein *Modell* gab, mit dessen Hilfe sich plausibel machen ließ, daß auch der universale Charakter der Vernunft durch ein rein mechanisches System realisiert sein kann.

Diese Lücke wurde erst durch die Erfindung des Computers geschlossen. Denn Computer können in verschiedener Hinsicht als Universalinstrumente aufgefaßt werden. Und genau darin liegt wohl der Grund dafür, daß viele glauben, mit dem Computer zum ersten Mal ein überzeugendes Modell des Geistes zu besitzen. Ich will das im folgenden etwas genauer erläutern.

2. Das Konzept des Computers, so wie wir ihn heute kennen und wie er inzwischen auf fast jedem Schreibtisch steht, geht auf verschiedene Wurzeln zurück. Aber die wichtigste dieser Wurzeln liegt sicher in den bahnbrechenden Arbeiten des englischen Mathematikers Alan Turing.[3] Worin bestand Turings große Leistung? Wenn man es auf einen kurzen Nenner bringen will, kann man vielleicht sagen: Erstens in dem Nachweis, daß es zu jeder arithmetischen Funktion, die

[3] Besonders Turing (1936/37).

überhaupt berechenbar ist, eine Maschine gibt, die diese Funktion berechnet, und zweitens in dem weit über dieses erste Ergebnis hinausgehenden Nachweis, daß es eine *universelle* Maschine gibt, die den Wert jeder beliebigen berechenbaren Funktion für jedes beliebige Argument berechnet.

Auf die Details der Überlegungen Turings kann ich an dieser Stelle nicht eingehen. Das ist jedoch auch nicht notwendig. Denn schließlich zeigen diese Überlegungen tatsächlich nur, daß es universelle *Rechenmaschinen* gibt. Wenn Descartes die Vernunft als ein Universalinstrument bezeichnet, dann ist damit aber sicher mehr gemeint als nur ein universelles Instrument zur Berechnung arithmetischer Funktionen. Zu den Ergebnissen Turings mußte also noch etwas anderes hinzukommen, um die Idee, der Computer könne ein Modell des Geistes sein, plausibel erscheinen zu lassen. Soweit ich sehen kann, stammt dieses zusätzliche Element aus zwei Quellen: erstens den Ergebnissen der logischen Grundlagenforschung und zweitens der Entdeckung, daß man diese Ergebnisse bei der Programmierung von Computern zu nichtnumerischen Zwecken sinnvoll einsetzen kann.

Die Ergebnisse der logischen Grundlagenforschung sind in diesem Zusammenhang deshalb von Bedeutung, weil sie zeigen, daß man logisches Schließen – ebenso wie das numerische Rechnen – als rein formale Veränderung von (strukturierten) Zeichenreihen durchführen kann. Der formale Charakter der Logik war zwar schon lange vor diesen Ergebnissen bekannt. Aber Anfang der dreißiger Jahre dieses Jahrhunderts konnte Kurt Gödel zum ersten Mal zeigen[4], daß die Prädikatenlogik 1. Stufe vollständig kalkülisierbar ist, d. h., daß es für die Prädikatenlogik 1. Stufe Kalküle K gibt, für die gilt:

1. Eine Formel A ist in der Prädikatenlogik 1. Stufe genau dann allgemeingültig, wenn sie in K beweisbar ist.
2. Eine Formel B folgt in der Prädikatenlogik 1. Stufe genau dann aus den Formeln A_1, ..., A_n, wenn sie in K aus diesen Formeln ableitbar ist.

Auf der Grundlage dieser Ergebnisse ist es nicht schwer nachzuweisen, daß es Algorithmen gibt, die für jede allgemeingültige Formel nach endlich vielen Schritten zu dem Ergebnis „allgemeingültig" führen. Im Falle nichtallgemeingültiger Formeln haben diese Algo-

[4] Gödel (1930).

rithmen zwar den Nachteil, in einigen Fällen nie zu einem Ende zu kommen; aber trotz dieses Nachteils schien die Existenz solcher Algorithmen die Vermutung zu stützen, daß man auch das logische Schließen auf einer Maschine realisieren kann.

Diese Vermutung wurde allerdings erst mit dem Beginn der KI-Forschung in den fünfziger Jahren zu einem konkreten Programm. Zunächst war das, was man seitdem *automatisches Beweisen* nennt, nur ein Teilgebiet der KI-Forschung. Sogar der Bereich des Problemlösens entwickelte sich zunächst ganz unabhängig von der Forschung auf diesem Gebiet. Das lag unter anderem sicher daran, daß die ersten Probleme, die man zu lösen versuchte, mit Brettspielen wie Tic-tac-toe, Dame oder Schach zu tun hatten und daß man zur Lösung dieser Probleme logikunabhängige Methoden verwenden konnte. Schon Ende der 50er Jahre versuchten Newell, Shaw und Simon jedoch, generelle Methoden zur Lösung beliebiger Probleme zu entwickeln und in ein Programm zu integrieren, dem sie den ehrgeizigen Namen GENERAL PROBLEM SOLVER gaben.[5] Doch dieses Programm hielt nicht, was sein Name versprach. Anfang der 60er Jahre wurde zunehmend klar, daß es unmöglich war, komplexere Probleme mit Hilfe dieses Programms zu lösen, ohne auf Spezialwissen über den jeweils spezifischen Problembereich zurückzugreifen.[6]

Fast zur gleichen Zeit entwickelte der Mathematiker J. A. Robinson mit seinem Resolutionsalgorithmus[7] aber eine neue und außerordentlich effektive Methode zum Beweis prädikatenlogischer Formeln. Und da sich dieser Algorithmus leicht auf einem Computer implementieren ließ, schien sich in ihm der Traum eines universalen, d. h. *bereichsunabhängigen* Problemlösers nun doch zu erfüllen. Voraussetzung dafür war allerdings, daß sich zeigen ließ, daß jedes Problem auf die Ableitung einer prädikatenlogischen Formel aus einer Menge von Prämissen reduzierbar ist. Und diese Voraussetzung bildet tatsächlich die Prämisse dessen, was man das ‚Logizistische Programm' in der KI-Forschung nennen könnte. Die Gründe, die für dieses Programm sprachen, lagen auf der einen Seite natürlich in den Erfolgen, die man auf der Grundlage der genannten Prämisse erzielen konnte. Auf der anderen Seite sicher aber auch in der Einfachheit und Allgemeinheit

---

[5] Newell/Shaw/Simon (1960) und Newell/Simon (1963).
[6] Vgl. Schefe (1986, 33).
[7] Robinson (1965).

dieser Prämisse. Denn wenn jede Problemlösung auf die Ableitung einer Formel aus einer Menge von Prämissen reduzierbar ist, dann stellt jedes System mit der Fähigkeit zum automatischen Beweisen – also z. B. jeder Computer, der nach dem Muster des Robinsonschen Resolutionsalgorithmus programmiert ist – tatsächlich eine universale Problemlösungsmaschine im Sinne Descartes' dar. Unter dieser Voraussetzung kann das Problem, zu verstehen, wie die Fähigkeit zu denken in einem rein mechanischen System realisiert sein kann, daher als zumindest im Prinzip gelöst gelten. D. h., unter dieser Voraussetzung ist zwar nicht der Computer als solcher, wohl aber *jedes System mit der Fähigkeit zum automatischen Beweisen ein mögliches Modell des Geistes.*

3. Es dauerte jedoch nicht lange, bis die Probleme des Logizistischen Programms deutlich wurden. Das berühmteste dieser Probleme ist das sog. *Frame-Problem,* das man in einem ersten Schritt vielleicht am besten anhand des sehr illustrativen Beispiels erläutern kann, das Daniel Dennett in (1984) anführt.

R_1 sei ein Roboter, dem man als einziges Ziel einprogrammiert hat, für sich selbst zu sorgen. Dies gelingt ihm auch recht gut – bis er eines Tages mit einem unangenehmen Problem konfrontiert wird. Seine Energiequelle, eine mittelgroße Batterie, befindet sich in einem verschlossenen Raum zusammen mit einer Zeitbombe, die bald explodieren wird. R_1 findet den Raum und den Schlüssel und geht daran, einen Plan zur Rettung seiner Batterie zu entwerfen. Er weiß, daß sich die Batterie auf einem kleinen Wagen befindet, und so kommt R_1 zu dem Schluß, daß die folgende Handlung den gewünschten Effekt haben wird: ZIEHE_HERAUS (WAGEN, RAUM). Unglücklicherweise liegt jedoch auch die Zeitbombe auf dem Wagen. R_1 wußte das zwar; aber er zog daraus nicht den Schluß, daß seine Handlung die Bombe mit der Batterie nach draußen bringen würde. Die fatalen Konsequenzen sind offenkundig. Armer R_1.

„Kein Problem", sagen die Konstrukteure von R_1. „Unser nächster Roboter muß nicht nur die beabsichtigten Wirkungen seiner Handlungen berücksichtigen, sondern auch ihre nicht beabsichtigten Nebeneffekte. D. h., er muß in der Lage sein, alle Effekte abzuleiten, die eine Handlung in einer gegebenen Situation hervorruft." Sie nennen ihr nächstes Modell einen „robot-deducer", kurz R_1D_1, und bringen es in dieselbe Situation, in der R_1 vorher so schrecklich gescheitert war. Auch R_1D_1 kommt bald auf die Idee, daß die

Handlung ZIEHE_HERAUS (WAGEN, RAUM) sein Problem lösen könnte; aber bevor er sich daran machen kann, diese Handlung auszuführen, muß er zunächst noch ableiten, welche anderen Effekte die Ausführung dieser Handlung mit sich bringen würde. Dafür benötigt er eine Menge Zeit. Und als er sich gerade anschickt, zu beweisen, daß die von ihm ins Auge gefaßte Handlung nicht die Farbe der Wände des Raumes ändern würde, ist es zu spät – wieder explodiert die Bombe.

Die Konstrukteure sind etwas konsterniert. Nach einiger Zeit glauben sie aber, doch noch eine Lösung gefunden zu haben. „Es reicht nicht, alle Effekte und Nebeneffekte zu berücksichtigen. Wir müssen dem Roboter auch beibringen, die relevanten von den irrelevanten Effekten zu unterscheiden und die irrelevanten zu ignorieren." Entsprechend programmieren sie ihr nächstes Modell, das sie einen „robot-relevant-deducer" oder kurz: R_2D_1 nennen. Auch R_2D_1 wird zu Testzwecken in dieselbe Situation gebracht. Und die Konstrukteure sind überaus erstaunt, als sie feststellen, daß R_2D_1 vor dem Raum mit der tickenden Bombe sitzt – wie Hamlet, angekränkelt von der Blässe des Gedankens. „Tu endlich etwas", rufen sie ihm zu. „Ich bin doch dabei", antwortet R_2D_1 etwas ungnädig, „ich bin eifrig damit beschäftigt, Tausende von Nebeneffekten zu ignorieren, die ich als irrelevant erkannt habe. Immer wenn ich einen irrelevanten Nebeneffekt abgeleitet habe, setzte ich ihn auf die Liste der Effekte, die ich ignoriere, und..." Mitten im Satz wird R_2D_1 unterbrochen. Wieder explodiert die Bombe, bevor er einen vernünftigen Plan hat, mit dem er sein Problem lösen könnte.

Obwohl das Frame-Problem seit über 20 Jahren bekannt ist[8], gibt es – besonders in der Diskussion zwischen KI-Forschern und Philosophen – bis heute keine Einigkeit über den genauen Gehalt dieses Problems. Es ist nicht klar, worin das Problem überhaupt besteht; es ist nicht klar, ob und, wenn ja, wie man es lösen kann; und es ist nicht klar, was aus der Existenz bzw. der Unlösbarkeit dieses Problems eigentlich folgt. Klar ist, daß das Frame-Problem nur in einem bestimmten Rahmen entsteht, nämlich dann, wenn man versucht, Programme zu schreiben, die es einem Computer oder Roboter ermöglichen, die Veränderungen zu modellieren, die bestimmte Handlungen oder Ereignisse in der Welt hervorrufen. Das Problem,

[8] Ein guter Überblick über die Diskussion findet sich in Janlert (1987).

über das McCarthy und Hayes 1969 zum ersten Mal in der Literatur berichteten und dem sie damals den Namen „Frame-Problem" gaben, ergab sich jedoch nicht aus dieser Aufgabenstellung selbst, sondern aus der besonderen Art, in der sie diese Aufgabe zu lösen versuchten.[9] Sie faßten die verschiedenen möglichen Zustände, die die Welt annehmen kann, als Situationen auf und deuteten Ereignisse und Handlungen als Funktionen von Situationen in Situationen. Eine Situation selbst war dabei die Menge aller der Fakten, die in dieser Situation wahr sind. Zur Berechnung der Folgesituation *s'*, die sich aus der Situation *s* ergibt, wenn in ihr das Ereignis *e* stattfindet, verwendeten McCarthy und Hayes eine Reihe von Axiomen wie z. B.

(1) Wenn *x* ein Lebewesen ist und in der Situation *s* von *y* nach *z* geht, dann ist *x* in der Folgesituation in *z*.

Das Problem, das sich bei diesem Formalismus ergibt, besteht schlicht darin, daß man – zusätzlich zu den Axiomen, die die Veränderungen spezifizieren, zu denen ein bestimmtes Ereignis führt – auch noch eine ziemlich große Zahl von sogenannten „Frame Axiomen" braucht, in denen festgestellt wird, was sich alles nicht ändert, wenn dieses Ereignis stattfindet. Für die Tatsache, daß sich die Farbe eines Objektes bei der Bewegung von *y* nach *z* nicht ändert, z. B. braucht man ein zusätzliches Axiom wie

(2) Wenn *x* in der Situation *s* die Farbe *c* hat und von *y* nach *z* geht, dann hat *x* auch in der Folgesituation die Farbe *c*.

Da die meisten Handlungen und Ereignisse die meisten Fakten nicht verändern, besteht die größte Anzahl der Axiome, die man zur Berechnung der Folgen dieser Handlungen benötigt, aus langweiligen „Frame Axiomen" dieser Art. Und das bedeutet auch, daß der größte Teil der Zeit, die ein entsprechend programmiertes System dafür benötigt, die Folgen einer Handlung zu berechnen, damit verschwendet wird, zu beweisen, was sich alles nicht ändert. Dies ist das ursprüngliche Frame-Problem, das sich im übrigen ja auch bei Dennetts Roboter R_1D_1 zeigte, der mit seinem Problem u. a. deshalb nicht zu Rande kam, weil er seine Zeit damit verschwendete, zu beweisen, daß die Handlung ZIEHE_HERAUS (WAGEN, RAUM) nicht die Farbe der Wände des Raumes ändern würde.

[9] Vgl. zum folgenden McDermott (1987).

KI-Forscher bestehen häufig darauf, dies und nur dies sei das Frame-Problem, und sie betonen dann weiter, daß es für dieses Problem eine ganze Reihe von Lösungsvorschlägen gebe. Einer dieser Vorschläge ist das Programm STRIPS (Fikes und Nilsson, 1971), das zur Lösung des ursprünglichen Frame-Problems eine Strategie verwendet, die John Haugeland die „Strategie der schlafenden Hunde" genannt hat. Das Grundprinzip dieser Strategie ist ebenso einfach wie effektiv. STRIPS behandelt jede Situation als eine eigene Datenstruktur. Um herauszufinden, wie die Nachfolgesituation *s'* aussieht, die durch die Ausführung der Handlung *e* in der Ausgangssituation *s* entsteht, berechnet STRIPS die Veränderungen, die *e* in *s* bewirkt, und nimmt die entsprechenden Änderungen in der Datenstruktur vor. Alles andere wird einfach so gelassen, wie es ist. Zusätzliche Frame-Axiome, mit deren Hilfe abgeleitet werden kann, was sich nicht ändert, sind daher überflüssig.

Es ist nicht ganz klar, ob die „Strategie der schlafenden Hunde" eine in allen Punkten befriedigende Lösung des Frame-Problems in seiner ursprünglichen Form darstellt (vgl. z. B. Fodor, 1987). Aber diese Frage können wir hier getrost beiseite lassen. Viele Philosophen haben nämlich ganz unabhängig davon immer wieder betont, daß ihrer Meinung nach das ursprüngliche Frame-Problem nur die Spitze eines Eisbergs bilde. Unterhalb der Wasseroberfläche befände sich das eigentliche, viel tiefer gehende Problem, das sich nicht so einfach lösen lasse.

Die Argumentation, die hinter dieser Auffassung steht, läßt sich kurz so zusammenfassen. Selbst wenn es – wie etwa bei dem Programm STRIPS – nicht mehr nötig ist, abzuleiten, was sich bei der Ausführung einer Handlung nicht ändert, ist das entscheidende Problem noch nicht gelöst. Denn auch die Zahl der Veränderungen, die eine Handlung bewirkt, kann schon sehr groß sein. Und wenn das so ist, dann wird nicht nur bei dem Versuch, herauszufinden, was sich bei der Ausführung einer Handlung alles *nicht* verändert (dem von Drew McDermott so genannten „inertia problem"[10]), zuviel Zeit verschwendet. Dann kostet auch schon der Versuch zu berechnen, was sich alles *verändert*, so viel Zeit, daß die Lösung eines Problems nicht im Rahmen der zur Verfügung stehenden Zeit gefunden werden kann. Intelligentes Handeln, so diese Philosophen, setzt voraus, daß

[10] McDermott (1987, 117).

ein Problem nicht nur gelöst, sondern daß es in einer angemessenen Zeit gelöst wird. Und dies wiederum ist nur möglich, wenn bei dem Versuch, das Problem zu lösen, nicht alle, sondern nur die *relevanten* Folgen einer Handlung in Betracht gezogen werden.[11]

KI-Forscher haben sich gegen diese Art der Darstellung allerdings manchmal mit dem Einwand gewehrt, das Relevanz-Problem sei keineswegs identisch mit dem Frame-Problem, sondern ein eigenes davon unabhängiges Problem, das im übrigen unter dem Namen „Kontroll-Problem" durchaus bekannt sei. So schreibt z. B. Patrick Hayes in seinem Aufsatz „What the Frame Problem Is and Isn't":

> "Once one has developed some suitable representation of the world about which the reasoner is expected to reason, one needs also to arrange that the system performs deductions which are appropriate for its assigned tasks and doesn't get lost in clouds of valid, but irrelevant conclusions. (It is fairly easy to arrange that it doesn't generate invalid conclusions.) This is variously called the theorem-proving problem, or the control problem, or the search problem, in AI. This is *not* the frame problem either." (1987, 124 – Hervorh. vom Verf.)

Aber hier handelt es sich ganz offensichtlich nur um einen Streit um Worte, d. h. genauer gesagt um einen Streit um die Frage, welcher Name für welches Problem verwendet werden soll. Unbestritten ist, daß es das Frame-Problem oder das Kontroll-Problem oder das Relevanz-Problem tatsächlich gibt und daß es sich dabei um ein außerordentlich schwieriges Problem handelt, für das zur Zeit keine einfache Lösung in Sicht ist. Dies ist auch genau der Punkt des Dennettschen Beispiels. Worauf Dennett aufmerksam macht, ist nämlich gerade, daß die Strategie, erst alle Konsequenzen einer Handlung zu berechnen und dann zu entscheiden, ob sie relevant sind, keine Lösung darstellt. Denn diese Strategie bringt keine Sekunde Zeitersparnis. Die Lösung des Problems muß deshalb darin bestehen, Schlüsse, die zu irrelevanten Konsequenzen führen, erst gar nicht zu ziehen. Und dies scheint nur möglich, wenn man schon vorher weiß, was herauskommt. Eine Situation, die merkwürdig paradox erscheint.

4. Was folgt aus alledem für die Ausgangsfrage, ob der Computer ein Modell des Geistes ist? Nun, zunächst einmal sollte klar gewor-

[11] Vgl. z. B. Pylyshyn (1987 a, x).

den sein, daß Computer zwar universale Rechenmaschinen sind, daß dies aber in diesem Zusammenhang sicher nicht entscheidend ist. Niemand, denke ich, hat je die Auffassung vertreten, daß geistige Fähigkeiten auf die Fähigkeit zurückgeführt werden können, beliebige arithmetische Funktionen zu berechnen. Die Auffassung, daß Computer etwas mit dem Geist zu tun haben könnten, konnte vielmehr erst aufgrund des Nachweises entstehen, daß logisches Schließen als rein formales Verändern von Zeichenreihen auch auf einem Computer realisiert werden kann. Vielleicht sollte die Frage deshalb besser so formuliert werden: „Sind automatische Beweissysteme Modelle des Geistes?" Oder in der Descartesschen Form: „Gibt es intelligente Problemlösungssysteme, deren Fähigkeiten allein auf der Fähigkeit zum automatischen Beweisen beruhen?" So wäre die Frage jedoch allzu eng gestellt. Denn in der KI-Forschung werden zur Problemlösung auch nichtdeduktive Methoden eingesetzt. Das Programm STRIPS hatte ich schon erwähnt. Andere Ansätze, die in diesem Zusammenhang zumindest genannt werden sollen, sind die Versuche zur Entwicklung einer nichtmonotonen Logik bzw. einer Logik des „default reasoning" und der „circumscription"-Ansatz. Ich möchte für die Ausgangsfrage deshalb folgende Formulierung vorschlagen: „Gibt es intelligente Problemlösungssysteme, deren Fähigkeiten allein auf den in der KI-Forschung üblichen Symbolverarbeitungsprozessen beruhen?"

Was folgt aus dem Frame-Problem für die Beantwortung dieser Frage? Zum einen sicher, daß Universalität nicht alles ist. Intelligente Problemlösungssysteme müssen nicht nur bereichsunabhängig, sie müssen auch schnell genug sein, d. h. sie müssen die Fähigkeit besitzen, Probleme nicht nur zu lösen, sondern in einer den Problemen angemessenen Zeit zu lösen. Kann das mit den herkömmlichen Methoden der KI-Forschung geleistet werden? Die Antwort auf diese Frage hängt natürlich davon ab, wie man die Aussichten einer weiteren Verbesserung dieser Methoden einschätzt. Aber sie hängt auch ab von einer Einschätzung der Grundlage, auf der alle diese Methoden aufbauen. Dennett hat dazu eine interessante Bemerkung gemacht:

> "From one point of view, non-monotonic or default logic, circumscription, and temporal logic all appear to be radical improvements to the mindless and clanking deductive approach, but from a slightly different perspective they appear to be more of same, and at least as unrealistic as frameworks for psychological models." (1984, 164).

Der Punkt, auf den Dennett hier abhebt, ist derselbe, den auch John Haugeland in seiner Analyse des Frame-Problems betont: Nicht nur der Grundansatz, Problemlösen auf automatisches Beweisen zurückzuführen, sondern auch alle Versuche zur Verbesserung dieses Grundansatzes gehen von der gemeinsamen Voraussetzung aus, daß das Wissen über die Umwelt, über das ein System verfügen muß, um seine Handlungen sinnvoll planen zu können, in der Form von prädikatenlogischen Formeln repräsentiert sein muß, also in einer quasi-sprachlichen Repräsentationsform. Diese Repräsentationsform hat jedoch den Nachteil, daß die Informationen, die in einem Satz oder einer Menge von Sätzen nur implizit enthalten sind, erst mit Hilfe von – möglicherweise aufwendigen – Ableitungen explizit gemacht werden müssen, um handlungsrelevant werden zu können. Haugeland führt als einfaches Beispiel zwei Sätze über die relative Lage dreier Städte A-Stadt, B-Stadt und C-Stadt an.

(3) A-Stadt liegt 100 km nördlich von B-Stadt.
(4) A-Stadt liegt 200 km nordwestlich von C-Stadt.

Diese beiden Sätze sind mit einer großen Anzahl von Sätzen über die relative Lage von B-Stadt und C-Stadt unvereinbar, z. B. mit dem Satz

(5) B-Stadt liegt 600 km westlich von C- Stadt.

Wenn man diesen Satz aber zu den zwei vorherigen hinzufügt, ergibt sich eine Inkonsistenz nur, wenn es gelingt, aus diesen drei Sätzen und einer Menge von Zusatzannahmen, in denen Wissen über die geometrisch möglichen Anordnungen von drei Städten auf der Erdoberfläche gespeichert ist, einen expliziten Widerspruch abzuleiten. Satzartige Repräsentationen sind im Hinblick auf ihre logischen Konsequenzen opak, könnte man sagen. D. h., diese Konsequenzen sind nicht unmittelbar aus ihnen abzulesen. Und genau deshalb benötigt man deduktive Verfahren, um sie explizit zu machen.

Andere Repräsentationsformen verhalten sich in diesem Punkt viel freundlicher. Dies gilt besonders für quasi-bildhafte Repräsentationen wie etwa Landkarten. Wenn wir die in den Sätzen (3) und (4) explizit enthaltenen Informationen mit Hilfe einer Karte repräsentieren, ergibt sich z. B. die folgende Struktur:

A-Stadt

B-Stadt

C-Stadt

Wenn man diese Repräsentation mit der Repräsentation vergleicht, die aus den Sätzen (3) und (4) gebildet wird, wird der entscheidende Unterschied sofort deutlich. In beiden Fällen wird die relative Lage von B-Stadt und C-Stadt zwar durch die repräsentierten Fakten determiniert. Aber im Falle der Repräsentation, die aus den beiden Sätzen (3) und (4) gebildet wird, gibt es einen scharfen Trennungsstrich zwischen dem, was explizit repräsentiert ist, und dem, was nur implizit im explizit Repräsentierten enthalten ist. Genau aus diesem Grund bedarf es einigen deduktiven Aufwandes, um das nur implizit Repräsentierte an die Oberfläche zu bringen. Im Fall quasi-bildhafter Repräsentationen gibt es diesen scharfen Trennungsstrich dagegen nicht. In der angegebenen Karte gibt es keinen Unterschied zwischen den Repräsentationen der relativen Positionen von A-Stadt und B-Stadt und von A-Stadt und C-Stadt auf der einen und der Repräsentation der relativen Position von B-Stadt und C-Stadt auf der anderen Seite. Wenn man die ersten beiden in die Karte eingetragen hat, ergibt sich die dritte ohne jeden zusätzlichen Rechenaufwand von selbst. Quasi-bildhafte Repräsentationsformen haben den Vorteil, sozusagen von selbst dafür zu sorgen, daß außer den ausdrücklich eingegebenen Fakten auch viele Konsequenzen repräsentiert werden, die sich aus diesen Fakten ergeben. Repräsentationsformen, bei denen aus diesem Grund eine scharfe explizit/implizit-Unterscheidung keinen Sinn macht, nennt Haugeland „komplizit". Und komplizite Repräsentationsformen sind seiner Meinung nach im Zusammenhang mit dem Frame-Problem von entscheidender Bedeutung, da sie die Berechnung der Konsequenzen repräsentierter Fakten in vielen Fällen überflüssig machen.[12]

Wenn Haugeland recht hat, dann scheint das Frame-Problem nichts weiter zu sein als ein Artefakt, das sich allein aus der Annahme ergibt, daß Repräsentationen quasi-sprachlichen Charakter haben müssen. Diese Annahme, so Haugeland, bildet aber das Fundament der gesamten „klassischen" KI-Forschung. Und daraus scheint zwingend zu folgen, daß Systeme, die auf den herkömmlichen Methoden

[12] Haugeland (1987, 88 ff.).

der KI-Forschung beruhen, keine adäquaten Modelle des Geistes sein können, da es auf der Grundlage dieser Methoden keine Lösung für das Frame-Problem gibt.

5. Wenn es stimmt, daß die Annahme, daß Repräsentationen quasi-sprachlichen Charakter haben müssen, wirklich zu den unaufgebbaren Grundannahmen der KI-Forschung gehört, dann ist diese Schlußfolgerung wohl unausweichlich. Aber das ändert nichts daran, daß man den herkömmlichen KI-Systemen in einem anderen Sinne trotzdem einen Modellcharakter nicht absprechen kann. Um dies erläutern zu können, möchte ich noch einmal auf die Überlegungen zurückkommen, mit denen ich diesen Aufsatz begonnen hatte.

Zunächst hatte ich darauf hingewiesen, daß Descartes einen radikalen Bruch mit der aristotelischen Tradition vollzog, als er es sich zum Programm machte, die für Lebewesen charakteristischen Fähigkeiten und Verhaltensweisen rein mechanisch zu erklären. Und ich hatte betont, daß Descartes nur deshalb versuchen konnte, dieses Programm auch durchzuführen, weil es Modelle gab, an denen er sich bei dieser Arbeit orientieren konnte: Uhren, Orgeln und die kunstvollen hydraulischen Steuerungen bestimmter Gartenfiguren. Wenn man sich seine Ausführungen im Detail ansieht, bemerkt man schnell, wie stark Descartes in seinem Denken von diesen und anderen mechanischen Modellen beeinflußt ist. Das Herz zum Beispiel ist für ihn eine Pumpe, die mit mechanischer Kraft dafür sorgt, daß das Blut durch den Körper fließt. Bei der Verdauung wird die Nahrung seiner Meinung nach im Magen zunächst mechanisch zerkleinert und dann, ebenfalls durch mechanische Kraft, durch die Därme bewegt. Dort „treffen die feinsten und bewegtesten Teilchen hier und dort auf eine Unzahl von kleinen Löchern", durch die sie auf dem Wege über die Pfortader zur Leber gelangen. Dabei werden diese Teilchen von den gröberen wie durch ein Sieb getrennt. Es ist nur „die Kleinheit der Löcher, die sie von den gröberen Teilchen scheidet".[13] Ich kann dies hier nicht weiter ausführen. Aber schon diese skizzenhaften Andeutungen machen, wie mir scheint, ausreichend deutlich, daß kaum eine der Erklärungen, die Descartes im einzelnen ausführt, den zu erklärenden Phänomenen wirklich gerecht wird. Und es ist ja auch, wie wir heute wissen, völlig aussichtslos, die Verdauung als einen rein

[13] Descartes (1969, 46).

mechanischen Vorgang der Zerkleinerung und des Aussiebens zu verstehen. Überhaupt ist die Mechanik keine ausreichende Grundlage für die Physiologie. Mit anderen Worten, Descartes mußte bei der Ausführung seines Programms scheitern. Denn die Erklärung physiologischer Vorgänge kann nur auf der Basis einer ausgearbeiteten Chemie erfolgen. Und die stand Descartes noch nicht zur Verfügung.

Trotzdem, und das ist hier für mich das Entscheidende, waren Descartes' Programm und seine Versuche, dieses Programm auszuführen, in einem anderen Sinne außerordentlich erfolgreich. Seine Überlegungen machten deutlich, daß der Versuch, auch die Phänomene des Lebens einer naturwissenschaftlichen Erklärung zugänglich zu machen, nicht von vornherein zum Scheitern verurteilt war. Auch wenn sich im Detail vieles – um nicht zu sagen, alles – als falsch herausstellte, war daher mit den Überlegungen Descartes' eine Tür aufgestoßen. Er überzeugte die Wissenschaftler seiner Zeit ebenso wie spätere Wissenschaftler davon, daß sein Programm im Prinzip durchführbar war. Und nur deshalb konnte überhaupt ein Forschungsprozeß in Gang kommen, von dem wir heute wohl sagen können, daß Descartes' Ziel inzwischen größtenteils erreicht wurde.

Ich denke, daß der Computer, oder besser gesagt: die „klassischen" Programme der KI-Forschung für die wissenschaftliche Erklärung des menschlichen Geistes denselben Modellcharakter haben könnten, den die Uhren, Orgeln und hydraulisch gesteuerten Gartenfiguren für Descartes' Programm der Erklärung des Lebendigen hatten. Sie sind keine realistischen oder adäquaten Modelle geistiger Fähigkeiten. Aber sie nehmen dem Geist die Aura des Unerklärbaren, indem sie andeuten, wie es im Prinzip gehen *könnte*. Vielleicht verhält sich die an den Modellen der KI-Forschung orientierte Kognitionswissenschaft zu einer künftigen adäquaten Theorie geistiger Phänomene so wie die durch mechanische Modelle inspirierte Cartesische Physiologie zu der von der Basis der organischen Chemie ausgehenden modernen Physiologie.

Literatur

Beckermann, A. (1989) „Aristoteles, Descartes und die Beziehungen zwischen Philosophischer Psychologie und Künstlicher-Intelligenz-Forschung", in: Pöppel, E. (Hrsg.), *Gehirn und Bewußtsein*. Weinheim: VCH Verlagsgesellschaft, 105–123.

Boden, M. (ed.) (1990) *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

Dennett, D. (1984) „Cognitive Wheels: The Frame Problem of AI", in: Pylyshyn (1987), 41–64, und in: Boden (1990), 147–170.

Descartes, R. (1960) *Discours de la méthode*. Franz.-deutsch, übers. und hrsg. von L. Gäbe. Hamburg: Felix Meiner.

– (1969) *Über den Menschen* und *Beschreibung des menschlichen Körpers*. Übers. von K. E. Rothschuh. Heidelberg: Lambert Schneider.

Fikes, R. E. and Nilsson, N. J. (1971) "STRIPS: A New Approach to the Application of Theorem Proving in Problem Solving". *Artificial Intelligence* 2, 189–208.

Fodor, J. A. (1987) „Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres", in: Pylyshyn (1987), 139–149.

Gödel, K. (1930) „Die Vollständigkeit der Axiome des logischen Funktionenkalküls". *Monatshefte für Mathematik und Physik* 37, 349–360.

Haugeland, J. (1987) "An Overview of the Frame Problem", in: Pylyshyn (1987), 77–93.

Hayes, P. J. (1987) „What the Frame Problem Is and Isn't", in: Pylyshyn (1987), 123–137.

Janlert, L.-E. (1987) "Modeling Change – The Frame Problem", in: Pylyshyn (1987), 1–40.

LaMettrie, J. O. de (1748) *L'homme machine*. EA Leiden.

McCarthy, J. and Hayes, P. J. (1969) „Some Philosophical Problems from the Standpoint of Artificial Intelligence", in: Meltzer, B. and Michie, D. (eds.) *Machine Intelligence 4*. Edinburgh: Edinburgh University Press, 463–502.

McDermott, D. (1987) „We've Been Framed: Or, Why AI Is Innocent of the Frame Problem", in: Pylyshyn (1987), 113–122.

Newell, A., Shaw, J. C. and Simon, H. (1960) „Report on a General Problem-Solving Program for a Computer", in: *Information Processing: Proceedings of the International Conference on Information Processing*. Paris: UNESCO, 256–264.

Newell, A. and Simon, H. (1963) „GPS, a Program That Simulates Human Thought", in: Feigenbaum, E. and Feldman, J. (eds.). *Computers and Thought*. New York: McGraw Hill, 279–293.

Pylyshyn, Z. W. (ed.) (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Ablex.

– – (1987a) „Preface", in: Pylyshyn (1987), vii-xi.

Robinson, J. A. (1965) „A Machine-Oriented Logic Based on the Resolution Principle". *Journal of the American Association for Computing Machinery* 12, 23–41.

Schefe, P. (1986) *Künstliche Intelligenz – Überblick und Grundlagen*. Mannheim/Wien/Zürich: BI – Wissenschaftsverlag.

Specht, R. (1966) *Descartes*. Reinbek bei Hamburg: Rowohlt.
Turing, A. (1936/37) „On Computable Numbers with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society* 42, 230–265, und 43, 544–546.
– – (1950) „Computing Machinery and Intelligence". *Mind* 59, 433–460. Wiederabgedr. in: Boden (1990), 40–66.