# KANN DIE KÜNSTLICHE INTELLIGENZ-FORSCHUNG FRAGEN DER PHILOSOPHIE BEANTWORTEN?

Ansgar Beckermann
Philosophisches Seminar
Georg-August-Universität
Nikolausberger Weg 9c
D-3400 Göttingen

Abstract:
The main question which links philosophy with AI-research already occurs in Descartes: Is it possible to construct machines which have the same intellectual capacities that man has, namely those of thinking and speaking. In my opinion this question can only be answered by AI-research. But even if the answer to this question should be positive, other questions will remain: would such machines have consciousness, subjectivity, intentionality? These questions will have to be answered primarily by the philosopher. For the answers depend to a large extent on the clarification of the involved concepts. Even in this field, however, interdisciplinary cooperation could yield interesting results.

1. Wenn man wissen will, ob die KI-Forschung Fragen der Philosophie beantworten kann, muß man zunächst klären, welche Fragen der Philosophie dabei gemeint sein sollen. Ich werde deshalb zu Beginn versuchen, klar zu machen, an welche Fragen ich bei der Formulierung des Titels gedacht habe. D.h. genauer: zuerst werde ich darstellen, wie sich diese Fragen im Laufe der Geschichte der Philosophie entwickelt haben, dann werde ich auf die Antworten eingehen, die auf diese Fragen bisher gegeben worden sind, und schließlich werde ich dann eine Vermutung darüber äußern, welche dieser Fragen möglicherweise von der KI-Forschung beantwortet werden können und mit welchen die Philosophie selber zu Rande kommen muß.

Etwas sollte ich jedoch gleich noch hinzufügen. Meinem eigenen Arbeitsgebiet entsprechend werde ich hier im wesentlichen Fragen ansprechen, die aus dem Gebiet der Philosophischen Psychologie stammen. (Nur am Schluß wird auch das Gebiet der Sprachphilosophie kurz gestreift). Philosophische Psychologie, das ist die philosophische Teildisziplin, die sich mit den Problemen der Seele beschäftigt. Ihre Fragen sind z.B.: Gibt es - zumindest beim Menschen - eine vom Körper unabhängige immaterielle Seele? Oder ist es vielleicht denkbar, daß auch Menschen nur Maschinen bestimmter Art sind? Wie verhalten sich generell körperliche und seelische bzw. geistige Phänomene zueinander? Usw.

2. Nach dieser Vorbemerkung möchte ich nun mit einem kurzen Ausflug in die Geschichte der Philosophischen Psychologie beginnen. In der Antike war die Seele ganz allgemein das, was das Lebende vom Toten unterscheidet. D.h. in der Antike war die Seele nicht nur für das Denken, Wollen, Wahrnehmen, Fühlen usw. verantwortlich, sondern auch für die Vorgänge der Ernährung und des Wachstums bei Pflanzen und Tieren, für die Vorgänge der Fortbewegung bei den Tieren und für alle ähnlichen Lebensprozesse. Diese Grundidee war wirksam bis hinein in die Biologie und die Physiologie. In diesen Wissenschaften wurden zur Erklärung der entsprechenden Phänomene immer wieder spezielle seelische Kräfte und Fähigkeiten angenommen. Ernährung und Wachstum z.B. wurden durch eine spezielle Fähigkeit der verschiedenen Körperteile erklärt, aus dem Nahrungsbrei bzw. aus dem Blut die für sie wichtigen Bestandteile herauszuziehen (facultas attrahens). Das Schlagen des Herzens wurde erklärt durch eine spezifische vis pulsifica. Und alle diese Fähigkeiten wurden als körperlich nicht begründbar der Seele zugeschrieben.

Diese Art der "wissenschaftlichen" Erklärung wurde erst in der Neuzeit heftig kritisiert. Galilei wäre da zu nennen; aber auch René Descartes. Descartes ist vielleicht besser bekannt als der Philosoph des "cogito, ergo sum" oder als der eigentliche Erfinder des strengen Dualismus von Körper und Seele. Doch das ist nur eine Seite seines philosophischen Denkens. Denn in seiner Naturphilosophie ist Descartes ein begeisterter Anhänger der von Galilei ausgehenden neuen Wissenschaft, und in gewissem Sinne kann man ihn sogar als radikalen Materialisten ansehen. Auf jeden Fall zählt er zu den vehementesten Kritikern der alten, durch die Ideen des Aristoteles geprägten Biologie und Physiologie. Für ihn ist die Bezugnahme auf spezielle Vermögen und Kräfte der Seele zur Erklärung der vitalen Vorgänge in einem Lebewesen

weder ein sinnvolles noch ein notwendiges Unterfangen. Es ist nicht sinnvoll: denn diese speziellen Vermögen und Fähigkeiten erklären nichts; sie bezeichnen eher Lücken der Erklärung. Und es ist auch nicht notwendig: denn nach Descartes' Auffassung sind alle Vorgänge in einem lebenden Körper ebenso mechanisch erklärbar wie die Ereignisse in der unbelebten Natur. Descartes bezieht im Hinblick auf Biologie und Physiologie also einen unerwartet materialistischen Standpunkt, und man kann ihn deshalb wohl zu Recht als einen der ersten konsequenten Verfechter einer einheitlichen Naturauffassung bezeichnen. Die überkommene Zweiteilung der Natur in einen belebten und einen unbelebten Bereich läßt er nicht mehr gelten. Denn seiner Meinung nach sind auch alle Lebensvorgänge - wie das Wachstum, die Bewegung, die Wahrnehmung und die Fortpflanzung der Pflanzen und Tiere - rein mechanische Vorgänge, die sich aufgrund der in der ganzen Natur in gleicher Weise geltenden Gesetze allein aus dem Aufbau und der Anordnung der in einem Lebewesen enthaltenen Teile ergeben. Und dementsprechend unterscheiden sich die belebten von den unbelebten Dingen für Descartes nicht dadurch, daß sie eine Seele haben. Der Unterschied, der zwischen einem lebenden und einem toten Wesen besteht, ist vielmehr der gleiche, der zwischen einer funktionsfähigen und einer nicht mehr funktionsfähigen - also eine kaputten - Maschine besteht. Crombie schreibt in seinem Buch Von Augustinus bis Galilei: "(Descartes') großer Wurf war die eine umfassende Theorie: der Körper ist eine Maschine; alle seine Tätigkeiten lassen sich mit denselben physikalischen Prinzipien und Gesetzen erklären, die für die unbelebte Welt gelten" (Crombie 1977, 470).

Allerdings - und damit kommen wir zurück zum Dualismus von Körper und Seele: die mechanische Erklärbarkeit aller Lebensvorgänge hat für Descartes da eine Grenze, wo beim Menschen die Fähigkeit, zu denken und zu sprechen, ins Spiel kommt. Tiere sind seiner Auffassung nach zur Gänze Maschinen, Menschen aber nicht. Leider gibt es nur wenige Textstellen, in denen Descartes diese Auffassung begründet. Aber z.B. im Discours de la méthode schreibt er:

> "Wenn es Maschinen mit den Organen und der Gestalt eines Affen oder eines anderen vernunftlosen Tieres gäbe, so hätten wir gar kein Mittel zu erkennen, daß sie nicht von genau derselben Natur wie diese Tiere wären; gäbe es dagegen Maschinen, die unseren Körpern ähnlich wären und unsere Handlungen insoweit nachahmten, wie dies für Maschinen wahrscheinlich möglich ist, so hätten wir

immer zwei ganz sichere Mittel, um zu erkennen, daß sie keineswegs wahre Menschen sind. Erstens könnten sie nämlich niemals Worte oder andere Zeichen dadurch gebrauchen, daß sie sie zusammenstellen, wie wir es tun, um anderen unsere Gedanken mitzuteilen ... (Und zweitens:) Sollten diese Maschinen auch manches ebenso gut oder sogar besser verrichten als irgendeiner von uns, so würden sie doch zweifellos bei vielem anderen versagen, wodurch offen zutage tritt, daß sie nicht aus Einsicht (connaissance) handeln, sondern nur aufgrund der Einrichtung ihrer Organe. Denn die Vernunft (raison) ist ein Universalinstrument, das bei allen Gelegenheiten zu Diensten steht, während diese Organe für jede besondere Handlung einer besonderen Einrichtung bedürfen .." (Discours 5.10, PhB 261 92 f., AT VI 57; Hervorh. vom Verf.)

Aus zwei Gründen können Descartes zufolge also Menschen - als ganze betrachtet - keine Maschinen sein:

- Maschinen können im Gegensatz zum Menschen nicht vernünftig sprechen. D.h. sie können Worte nicht so variationsreich und situationsbezogen zusammenfügen wie wir, wenn wir uns mitteilen wollen.

- Maschinen können zwar in einzelnen Situationen erstaunliche Dinge leisten; aber nur Menschen sind in der Lage, in den verschiedensten Situationen den Umständen angepaßt vernünftig zu handeln. "Denn die Vernunft ist ein Universalinstrument".

Auf eine kurze Formel gebracht: Sprache und Intelligenz sind die beiden Merkmale, die nach Descartes den Menschen von jeder möglichen Maschine unterscheiden. Dabei weiß Descartes natürlich auch, daß es Maschinen geben kann, die in der Lage sind, bestimmte Wörter zu äußern, und die sogar in der Lage sind, diese Wörter immer nur bei bestimmten Gelegenheiten zu äußern. Auch er konnte sich Maschinen vorstellen, die "Guten Tag" sagen, wenn man sie an einer bestimmten Stelle berührt, oder die laut "Aua" schreien, wenn man sie unsanft schüttelt. Aber, so schreibt er: "... man kann sich nicht vorstellen, daß (diese Maschine) die Worte auf verschiedene Weise zusammenordnet, um auf die Bedeutung all dessen, was in ihrer Gegenwart laut werden mag, zu antworten, wie es der stumpfsinnigste Mensch kann" (ebd.) D.h., Descartes zufolge können Maschinen nicht so sprechen, wie wir das können, da zum Sprechenkönnen gehört, daß man in der Lage ist, auf alles, was einem zu Ohren kommt, angemessen verbal zu reagieren.

Auch im Hinblick auf die Frage der Intelligenz argumentiert Descartes ähnlich, wobei er diesmal allerdings den Vergleich mit verschiedenen Tieren zieht (die freilich seiner Meinung nach tatsächlich als Maschinen angesehen werden können). Sicher, so Descartes, gibt es Tiere, die auf manchen Gebieten weit mehr Geschicklichkeit zeigen als wir. Aber diese Tiere vollbringen diese Höchstleistungen immer nur auf einem oder auf sehr wenigen Gebieten; auf allen anderen zeigen sie schlechtere Leistungen als wir. Wenn sie aber wirklich denken und ihr Handeln an ihrem Denken ausrichten könnten, dann müßten sie dies auf allen Gebieten tun können und nicht nur in einem bestimmten eng umgrenzten Bereich. "Der Tatbestand also, daß sie es besser machen als wir, beweist nicht, daß sie Geist (esprit) haben; denn wenn man es so nimmt, dann hätten sie mehr als irgendeiner von uns und würden es in jeder Beziehung besser machen. Aber sie haben im Gegenteil gar keinen, und es ist die Natur, die in ihnen je nach der Einrichtung ihrer Organe wirkt, ebenso wie offensichtlich eine Uhr, die nur aus Rädern und Federn gebaut ist, genauer die Stunden zählen und die Zeit messen kann als wir mit all unserer Klugheit" (Discours 5.11, PhB 261 96 f., AT VI 59). Kurz gefaßt kann man Descartes' These also so formulieren.

DESCARTES:

> 1. Tiere sind vollständig als Maschinen begreifbar, d.h. als rein materielle Systeme, deren Verhalten sich allein aus ihren Teilen und der Anordnung dieser Teile ergibt, wobei das Verhalten der Teile selbst allein durch die allgemein geltenden Gesetze der Physik bestimmt ist.
>
> 2. Menschen dagegen können keine rein materiellen Systeme sein, d.h. sie müssen außer einem materiellen Körper auch noch eine immaterielle Seele besitzen, da sie über Fähigkeiten verfügen (Intelligenz und Sprache), die keine Maschine, d.h. kein rein materielles System, besitzen kann.

Bevor wir uns nach diesem Ausflug in die Geschichte der Philosophischen Psychologie der neueren Diskussion zuwenden, möchte ich zu dieser These noch drei Bemerkungen machen.

Erstens: Die erste Teilthese Descartes', daß Tiere vollständig als Maschinen begreifbar seien, hatte nicht nur schwerwiegende Folgen (letzten Endes hängt noch die heutige Diskussion um die Zulässigkeit

von Tierversuchen mit dieser These zusammen), sie war zu seiner Zeit auch kaum mehr als eine kühne Spekulation. Descartes kannte zwar einige Maschinen (er selbst erwähnt Uhren, kunstvolle Wasserspiele und Mühlen); aber die Prinzipien, nach denen diese Maschinen funktionieren, und Descartes' eigene simple Modellvorstellungen von dem, was in einem lebenden Körper geschieht, waren weit davon entfernt, physiologische Vorgänge wirklich erklären zu können. Immerhin hatte Descartes keinen vernünftigen Begriff von Chemie, und er kannte auch die Phänomene der Elektrizität noch nicht. Sie können sich also leicht vorstellen, wie weit man ohne diese Kenntnisse in der Physiologie kommen kann.

Zweitens: Daß Descartes die Grenzlinie gerade zwischen Tieren (alle höheren Säugetiere eingeschlossen) und Menschen zieht, hat nicht nur bei vielen seiner Zeitgenossen und unmittelbaren Nachfolger zu Widerspruch geführt. Auch für uns heute ist diese Grenzziehung aus einer ganzen Reihe von Gründen unplausibel. Die Ideen der Darwinschen Evolutionstheorie z.B. sprechen nicht gerade für diese Trennungslinie. Außerdem - und das mag hier noch interessanter sein - ist mit der ersten Teilthese Descartes' die Überzeugung verbunden, daß nicht nur die Ernährungs- und Verdauungsprozesse, sondern auch alle Phänomene der tierischen Wahrnehmung und des tierischen "Denkens" rein maschinell erklärt werden können. Diese These mag vielleicht KI-Forscher freuen, die auf dem Gebiet der Computer Vision tätig sind. Aber angesichts der Tatsache, daß es sehr viel schwieriger zu sein scheint, ein funktionierendes System der visuellen Wahrnehmung als z.B. einen Schachcomputer zu programmieren, liegt zumindest die Frage nahe, ob Descartes die Wahrnehmungs- und auch die anderen kognitiven Prozesse von Tieren nicht einfach unterschätzt hat.

Drittens: Wenn Descartes zur Erklärung der Fähigkeiten des Denkens und Sprechens auf eine immaterielle Seele zurückgreift, dann bedient er sich damit eines Kunstgriffs, den er im Hinblick auf die nach-aristotelischen Erklärungen physiologischer Vorgänge selbst nachdrücklich kritisiert hat. Denn die Annahme einer immateriellen Seele erklärt in der Psychologie im Grunde ebenso wenig wie die Annahme spezieller Kräfte und Fähigkeiten in der Biologie.

3. Auf jeden Fall - und damit kommen wir nun endgültig zur neueren Diskussion im Bereich der Cognitive Science - geht aus der These Descartes' hervor, daß er schon im 17. Jahrhundert - lange also, bevor an KI überhaupt gedacht werden konnte - ein engagierter Kritiker der KI-Forschung war bzw. ein engagierter Kritiker bestimmter Auffassungen, die von einigen Vertretern der Cognitive Science verfochten werden - Auffassungen, die John Searle in seinem Aufsatz "Minds, Brains, and Programs" "starke KI" genannt hat. "Schwache KI", das ist nach Searle nur die Auffassung, daß die Arbeit mit dem Computer ein nützliches Instrument bei der Erforschung psychischer Vorgänge sein kann. "Starke KI" dagegen steht bei ihm für die sehr viel weiter gehende These, daß der Computer nicht nur ein nützliches Instrument sein kann, sondern daß der geeignet programmierte Computer tatsächlich in dem Sinne selbst einen Geist besitzt bzw. ein geistiges Wesen ist, als man von ihm im wörtlichen Sinn sagen kann, er verstehe etwas, sei intelligent und habe auch die erforderlichen kognitiven Zustände. Der "starken KI" zufolge sind die entsprechenden Programme nicht nur Mittel, um psychologische Erklärungen zu testen, sie sind vielmehr selbst die gesuchten psychologischen Erklärungen.

Searle sagt nicht, welche Autoren er für Vertreter der "starken KI" hält. Soweit ich sehen kann, gehören aber sicher Newell und Simon dazu. Denn in dem Aufsatz "Computer Science as Empirical Inquiry" entwickeln diese beiden Autoren eine Theorie physikalischer Symbolsysteme, die in der Tat auf das genaue Gegenteil der Descartesschen These hinausläuft. Was sind nach Newell und Simon nun physikalische Symbolsysteme?

Zunächst einmal wollen sie mit dem Adjektiv "physikalisch" ausdrücken, daß solche Systeme den Gesetzen der Physik gehorchen und daß sie daher auch durch von Menschen hergestelle Maschinen realisiert werden können. Im einzelnen besteht jedes physikalische Symbolsystem aus einer Menge von Symbolen - physikalischen Mustern, die insbesondere auch als Komponenten in Ausdrücken (oder wie Newell und Simon sagen: Symbolstrukturen) auftreten können. Zu jedem Zeitpunkt enthält ein physikalisches Symbolsystem eine bestimmte Menge solcher Symbolstrukturen. Außerdem verfügt das System auch noch über eine Menge von Prozessen, mit denen aus schon vorhandenen Symbolstrukturen neue Symbolstrukturen erzeugt werden können: Prozesse des Neuanlegens, des Veränderns, des Kopierens und des Löschens.

Zwei Begriffe sind für solche Strukturen von Ausdrücken, Symbolen und Objekten zentral: Bezeichnung und Interpretation. Newell und Simon schreiben nicht ganz klar:

Ein Ausdruck bezeichnet ein Objekt, wenn das System das Objekt beeinflussen kann oder sich in einer vom Objekt abhängigen Weise verhalten kann, falls der Ausdruck gegeben ist.

Das System kann einen Ausdruck interpretieren, wenn der Ausdruck einen Prozeß bezeichnet und das System der Prozeß ausführen kann, falls der Ausdruck gegeben ist.

Wichtig ist also, wie der letzte Punkt zeigt, daß in einem physikalischen Symbolsystem Symbolstrukturen vorkommen, die systemeigene Prozesse bezeichnen, und daß das System diese Prozesse selbst aufrufen und ausführen kann, wenn diese Symbolstrukturen vorliegen. Außerdem müssen nach Newell und Simon für physikalische Symbolsysteme noch einige Vollständigkeits- und Abgeschlossenheitsbedingungen erfüllt sein, die hier nur kurz erwähnt werden sollen:

- Ein Symbol kann einen beliebigen Ausdruck bezeichnen.
- Für jeden systemeigenen Prozeß existiert ein Ausdruck, der ihn bezeichnet.
- Es gibt Prozesse, um beliebige Ausdrücke zu erzeugen und um beliebige Ausdrücke in beliebiger Weise zu modifizieren.
- Ausdrücke sind stabil; wenn sie erzeugt sind, bleiben sie erhalten, bis sie modifiziert oder gelöscht werden.
- Die Zahl der Ausdrücke, die in einem System enthalten sein können, ist wesentlich unbeschränkt.

Newell und Simon selbst schreiben:

> "The type of system we have just defined is not unfamiliar to computer scientists. It bears a strong family resemblance to all general purpose computers. If a symbol-manipulaton language, such as LISP, is taken as defining a machine, then the kinship becomes truely brotherly." (Haugeland 1981, 41)

Ich glaube, daß diese Bemerkung klarer sagt, auf was Newell und Simon hinaus wollen, als alle vorherigen "Definitionen". Aber wie dem auch sei, ihre zentrale These lautet folgendermaßen.

NEWELL und SIMON:

> 1. Man kann physikalische Symbolsysteme so organisieren, daß sie allgemeine Intelligenz zeigen.
>
> 2. Alle Systeme, die allgemeine Intelligenz zeigen, werden sich bei genauer Analyse als physikalische Symbolsysteme erweisen.

Dabei soll der Ausdruck "allgemeine Intelligenz" den Bereich der Intelligenz bezeichnen, den wir bei menschlichem Handeln sehen. Er soll besagen, daß das System in jeder realen Situation seinen eigenen Zielen gemäß und den Erfordernissen der Umwelt angepaßt handeln kann - zumindest innerhalb gewisser Grenzen der Geschwindigkeit und Komplexität.

Ich denke, es ist klar, daß diese These von Newell und Simon der oben erläuterten Descartesschen These diametral entgegengesetzt ist. Descartes sagt, kein physikalisches System ist zu allgemeinem intelligentem Handeln fähig, also ist der Mensch kein physikalisches System; Newell und Simon sagen, jedes zu allgemeinem intelligentem Handeln fähige System - also auch der Mensch - wird sich bei genauer Analyse als eine bestimmte Art von physikalischem System, nämlich als physikalisches Symbolsystem, erweisen. Wer hat recht?

4. Bevor wir auf diese zentrale Frage weiter eingehen, scheint es mir sinnvoll, zunächst noch einen Moment innezuhalten und kurz zu untersuchen, was eigentlich hinter dem von Newell und Simon in die Diskussion eingeführten Begriff des physikalischen Symbolsystems steht. Dies läßt sich, wie mir scheint, mit einem einfachen Beispiel am besten verdeutlichen. Nehmen wir etwa das Problem, in einem Zwei-Personen-Brett-Spiel den günstigsten nächsten Zug zu finden. Dieses Problem kann man bekanntlich z.B. so angehen:

a) Man überlegt sich, welche legalen Züge man bei der gegebenen Stellung überhaupt zur Verfügung hat; die durch diese Züge entstehenden Stellungen faßt man in einer Liste L zusammen.
b) Man überlegt sich für jede in der Liste L enthaltene Stellung i, welche Züge der Gegner in dieser Stellung machen kann, und faßt die aus diesen Zügen resultierenden Stellungen für jedes i in einer Liste Li zusammen.

c) Man bewertet alle auf diese Weise entstandenen Stellungen (d.h. alle x, für die es ein i gibt, so daß gilt $x \in Li$) mit Hilfe einer vorgegebenen Funktion f.
d) Man sucht für jedes i in der Liste Li die Stellung mit dem kleinsten f-Wert; mit diesem Wert bewertet man die Stellung i.
e) Man sucht in der Liste L die Stellung, die im vorherigen Schritt am höchsten bewertet wurde, und führt den zu dieser Stellung führenden Zug aus.

Wenn man die Schritte in einer etwas anderen Reihenfolge durchführt, kann man das so beschriebene Verfahren auch durch das in Abb. 1 gezeigte Flußdiagramm darstellen.

Wie es üblich ist, sind die einzelne Schritte in dem durch dieses Diagramm dargestellten Algorithmus als Operations- oder Testanweisungen formuliert, so wie man eben Anweisungen an einen Menschen formuliert, von dem man möchte, daß er bestimmte Handlungen ausführt. Und das gesamte Diagramm besagt, in welcher Reihenfolge diese Teilhandlungen ausgeführt werden sollen, wenn man zu dem gewünschten Resultat gelangen will. Entscheidend ist nun aber, daß insbesondere die metamathematischen Überlegungen Turings und die später anschließende praktische Entwicklung von Computern in den 30er-, 40er- und 50er-Jahren zu dem Ergebnis führten, daß nicht nur Menschen, sondern auch Maschinen diese Aufgabe erledigen, d.h. die in dem Flußdiagramm aufgeführten Handlungen in der angegebenen Reihenfolge durchführen und somit zu demselben Ergebnis wie ein Mensch kommen können. Wie ist das zu verstehen? Wie kann eine Maschine z.B. die zuvor gestellte Aufgabe erledigen?

Die Antwort auf diese Frage lautet, soweit ich sehen kann, daß dafür drei Dinge nötig sind.

a) Die Maschine muß in der Lage sein, physikalische Muster zu speichern und zu manipulieren, für die folgendes gilt: es gibt eine Funktion, die jedem dieser Muster in eindeutiger Weise eine Stellung in dem angenommenen Brettspiel zuordnet.

b) In der Maschine muß für jede im Flußdiagramm angesprochene Teilhandlung ein Prozeß vorhanden sein, der die physikalischen Muster in entsprechender Weise manipuliert. D.h. es muß einen Prozeß geben, der zu einem gegebenen physikalischen Muster ein neues Muster erzeugt, das

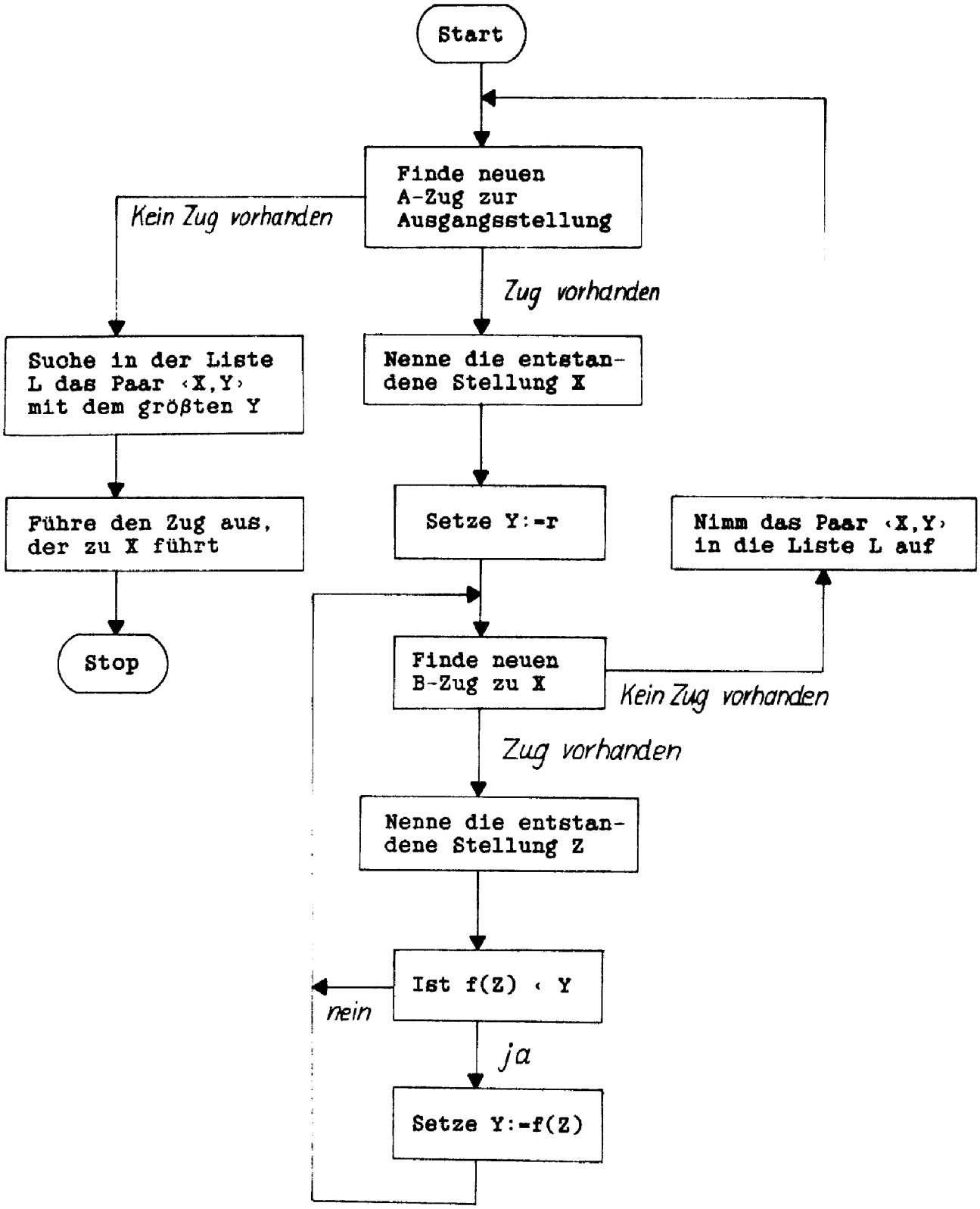


Abb. 1 (Bemerkungen zur Bezeichnung: die beiden Spieler sollen A und B heißen; r sei der größte Wert, mit dem f eine Stellung bewerten kann)

einer neuen durch einen legitimen Zug erzeugten Stellung entspricht; es muß einen Prozeß geben, der einem gegebenen Muster ein Muster zuordnet, das in eindeutiger Weise der natürlichen Zahl entspricht, die die Funktion f der entsprechenden Stellung zuordnet; es muß einen Prozeß geben, der zwei Muster, denen natürliche Zahlen entsprechen, daraufhin untersucht, ob das erste Muster einer kleineren Zahl entspricht als das zweite; usw.

c) Die Maschine muß so konstruiert sein, daß die den Teilhandlungen des Flußdiagramms entsprechenden Prozesse genau in der von dem Diagramm vorgeschriebenen Reihenfolge durchgeführt werden.

Wenn diese drei Bedingungen erfüllt sind, gibt es zwischen der Arbeitsweise der so charakterisierten Maschine und dem Vorgehen eines Menschen, der genau den Anweisungen des Flußdiagramms der Abb. 1 folgt, eine eindeutige Isomorphie. Und genau dann, wenn es eine solche Isomorphie gibt, kann man sagen, daß die Maschine in gewissem Sinne dieselbe Aufgabe wie der Mensch erledigt. Für Maschinen dieser Art sind inzwischen eine ganze Reihe Namen erfunden worden. Im Bereich der Cognitive Science sprechen viele von "informationsverabeitenden Maschinen", Daniel Dennett hat den Ausdruck "semantische Maschinen" geprägt; und Robert Cummins spricht von Maschinen mit "inferentiell charakterisierten Fähigkeiten". Wichtig ist aber nicht der Name, sondern das zuvor erläuterte Prinzip. Und an diesem Prinzip wird meiner Meinung nach deutlich, was hinter der These von Newell und Simon steckt. Denn offenbar gehen diese beiden Autoren von der Idee aus, daß alle Aufgaben, zu deren Erledigung Intelligenz vonnöten ist, mit Hilfe eines Algorithmus gelöst werden können, für den es eine isomorphe Maschine gibt.

6. Nach diesen Erläuterungen nun zurück zu der Frage, was für und was gegen die These von Newell und Simon spricht, genügend ausgestattete physikalische Symbolsysteme seien zu allgemeinem intelligentem Verhalten fähig und alle intelligenten Systeme seien der Art nach physikalische Symbolsysteme. Newell und Simon selbst legen großen Wert darauf zu betonen, daß diese These ihrer Ansicht nach eine empirische These ist - also eine These, über die nicht durch a priori Argumente, sondern nur durch Erfahrungen und Experimente entschieden werden kann.

Als empirische Evidenz für ihre These führen sie dann als erstes die bisherigen Erfolge der KI-Forschung an. Schon früh sei es gelungen,

für eine ganze Reihe von Problemen geeignete Programme zu entwickeln: Ein- und Zwei-Personen-Spiele, Operations-Research Probleme der optimalen Nutzung knapper Resourcen, einfache Probleme des induktiven Schließens. Alle diese Programme zeigten in einem begrenzten Bereich ein gewisses Maß an Intelligenz. Bei diesen einfachen Programmen sei es aber nicht geblieben. Vielmehr habe es einen stetigen Fortschritt sowohl in Richtung auf größere Leistungsfähigkeit in einem gegebenen Bereich als auch in Richtung auf eine Ausweitung der Bereiche gegeben. Die Steigerung der Leistungsfähigkeit von Schachcomputern sei dafür ein gutes Beispiel. Der Fortschritt sei zwar langsam, aber stetig. Und er umfasse immer neue Bereiche. Heute (so schrieben Newell und Simon 1976) gebe es Systeme zum Verstehen und zur Produktion von natürlicher Sprache, Systeme für die Interpretation visueller Szenen usw. Letzten Endes sei also kein Grund zu erkennen, warum es nicht für alle Probleme, deren Lösung intelligentes Handeln erfordert, geeignete Programme geben sollte.

> "If there are limits beyond which the hypothesis will not carry us, they have not yet become apparent." (Haugeland 1981, 48)

In ähnlicher Weise führen Newell und Simon als Evidenz für den zweiten Teil ihrer These die Fortschritte der Kognitiven Psychologie an. Und schließlich fügen sie noch hinzu, daß ein großer Teil der Evidenz für ihre These negativer Art sei, d.h. für diese These spreche insbesondere auch, daß es keine ernst zu nehmende Alternativhypothese darüber gebe, wie intelligentes Handeln - sei es beim Menschen, sei es bei der Maschine - zustande komme.

Soweit also die Pro-Argumente. Wie steht es nun mit der Kontra-Seite? Gegen die These von Newell und Simon sind - besonders auch von Seiten der Philosophie - zwei ganz verschiedene Arten von Argumenten vorgebracht worden. John Haugeland hat sie die hollow shell-Strategie (frei übersetzt: die Strategie der leeren Hülse) und die poor substitute-Strategie (die Strategie des billigen oder bloßen Ersatzes) genannt. Auf die hollow shell-Strategie komme ich später noch zurück. Zunächst einige Bemerkungen zur poor substitute-Strategie.

Das Ziel dieser Strategie ist zu zeigen, daß genau das Gegenteil von dem richtig ist, was Newell und Simon behaupten, daß es nämlich physikalische Symbolsysteme, die im Hinblick auf Intelligenz dasselbe input-output-Verhalten zeigen wie ein Mensch, nicht gibt. Offenbar war

Descartes ein Vertreter dieser Strategie, ein anderer - zeitgenössischer - ist Hubert L. Dreyfus. Welche Argumente sind nun von dieser Seite her vorgetragen worden?

Soweit ich sehen kann, lautet das Hauptargument, daß der von Newell und Simon im Hinblick auf die bisherigen Erfolge der KI-Forschung vertretene Optimismus völlig fehl am Platz ist. Denn aus dem bisherigen Verlauf des Fortschritts könne man keineswegs entnehmen, daß es nur noch eine Frage der Zeit ist, bis die KI-Forscher ein System entwickelt haben, das das Verhalten eines Menschen - zumindest soweit es intelligentes Handeln angeht - vollständig imitieren kann. Ich glaube, daß diese Skepsis durchaus eine vernünftige Basis hat. Es kann schon sein, daß Forscher wie Newell und Simon die Größe der Aufgabe unterschätzt haben. Immerhin finden wir schon bei Descartes die Bemerkung, daß die Vernunft ein Universalinstrument ist, das dem Menschen in allen möglichen Situationen zur Verfügung steht. Es reicht also nicht, für einzelne begrenzte Probleme Lösungen zu finden; wenn die Prognose von Newell und Simon eingelöst werden soll, muß vielmehr ein System geschaffen werden, das mit ebenso vielfältigen und unterschiedlichen Problemen fertig werden kann wie der Mensch.

Dreyfus spricht in diesem Zusammenhang in Anlehnung an eine Formulierung von Bar-Hillel von der Gefahr eines "Fehlschlusses des ersten erfolgreichen Schritts" (Dreyfus 1979, 80). Wenn jemand auf einen Baum steigt, dann hat er damit noch nicht den ersten Schritt für eine Reise zum Mond getan, selbst wenn er dem Mond auf diese Weise ein Stückchen näher gekommen sein sollte. Und ebenso - so Dreyfus - kann es sein, daß ein KI-Forscher, der ein begrenztes Problem gelöst hat, damit dem Problem, ein System mit den intellektuellen Fähigkeiten eines Menschen zu entwickeln, noch keinen Schritt näher gekommen ist. Denn vielleicht sind zur Lösung dieses zweiten Problems ganz andere Methoden erforderlich als zur Lösung des ersten. Dreyfus gibt eine Reihe von Beispielen, die seiner Meinung nach zeigen, daß die bisherigen Methoden der KI-Forschung für die von Newell und Simon gestellte Aufgabe nicht ausreichen. Ich will darauf im einzelnen nicht eingehen; aber ein Beispiel möchte ich doch kurz anführen. Dabei möchte ich mich jedoch vorsorglich gleich dafür entschuldigen, daß sich dieses Beispiel auf ein relativ altes Programm bezieht und also vielleicht durch den Stand der Forschung schon überholt ist. Ich meine den General Problem Solver von Newell, Shaw und Simon.

Der Name dieses Systems scheint zunächst ganz auf der Linie der Descartesschen Idee zu liegen, daß die Vernunft als ein Universalinstrument verstanden werden muß. Aber schon von seinen Schöpfern war er nicht so gemeint. Mit ihm sollte vielmehr nur deutlich gemacht werden, daß es sich hier um ein System handelte, bei dem der Problemlösungsteil allgemein, d.h. unabhängig von spezifischen Eigenschaften einzelner zu lösender Aufgaben formuliert war. Doch wie dem auch sei, meine Kritik am GPS geht nicht dahin, daß hier vielleicht noch nicht das effizienteste Verfahren für die heuristische Suche von Lösungspfaden verwendet wurde. Mein Punkt ist vielmehr, daß man, wenn ich das System richtig verstanden habe, dem GPS alle Details des zu lösenden Problems erst löffelweise eingeben muß, bevor die heuristische Suche überhaupt beginnen kann. Das gilt nicht nur für die Ausgangs- und Zielobjekte, sondern auch für die möglichen Operatoren, für die Bedingungen, die bei der Anwendung von Operatoren nicht verletzt werden dürfen, und insbesondere auch für die Liste und die Rangordnung der den ganzen Prozeß steuernden Differenzen. Alle diese Dinge müssen dem System erst in einer ihm verständlichen Form eingefüttert werden, bevor es losgehen kann. Damit wird jedoch ein Teil der Aufgabe nicht von der Maschine, sondern vom Programmierer gelöst, und zwar nicht der unwichtigste Teil. Denn offensichtlich ist die Strukturierung einer Problemsituation häufig der schwierigste Teil einer zu lösenden Aufgabe. Ein Mensch, der mit einem Problem oder einer Aufgabe konfrontiert wird, muß aber auch mit diesem Teil selbst fertig werden.

Mir ist leider nicht bekannt, ob es inzwischen Systeme gibt, die der genannten Einschränkung nicht mehr unterliegen. Das wäre sicher ein enormer Fortschritt. Ich wollte mit meinem Beispiel auch nur darauf aufmerksam machen, daß man die Schwierigkeit der Aufgabe, ein System zu entwickeln, das - zumindest was intelligentes Verhalten angeht - dieselben Fähigkeiten besitzt wie ein Mensch, nicht unterschätzen sollte. Und ich denke, daß die KI-Forschung das heute auch nicht mehr tut. Immerhin ist zu bedenken, daß ein System, das tatsächlich dieselben Fähigkeiten wie ein Mensch besitzen soll, mindestens die folgenden beiden Bedingungen erfüllen muß:

- ein solches System muß im Prinzip über dieselben Wahrnehmungs- und Handlungsmöglichkeiten wie ein Mensch verfügen

- ein solches System muß die zu lösenden Probleme mindestens in derselben Zeit wie ein Mensch lösen.

Bekanntlich stellt dies z.B. schon beim Sprachverstehen eine große Schwierigkeit dar. Deshalb hat gerade auch dieser letzte Punkt dazu geführt, daß es inzwischen eine ganze Reihe von Philosophen gibt, die es immerhin für möglich halten, daß die Leistungen, zu denen ein Mensch fähig ist, tatsächlich nur von einem menschlichen Gehirn erbracht werden können. Und zu diesen Philosophen gehören auch Autoren wie Daniel Dennett und Douglas R. Hofstadter, die im Prinzip der Idee einer materialistischen Analyse der menschlichen Geistes gar nicht ablehnend gegenüberstehen.

Aber ich will diesen Punkt an dieser Stelle nicht weiter verfolgen. Denn ebenso wie Newell und Simon glaube ich, daß es hier tatsächlich um eine empirische Frage geht, die nicht durch a priori Argumente, sondern nur durch empirische Forschung und durch Experimente entschieden werden kann. Und das ist auch der Grund für meine Auffassung, daß hier eine wichtige Frage der Philosophie vorliegt, die nur von der KI-Forschung beantwortet werden kann: Gibt es Maschinen (Computer oder noch besser Roboter), die über dieselben Fähigkeiten zu intelligentem Verhalten verfügen wie ein erwachsener Mensch?

6. Zum Schluß soll nun auch noch die hollow shell-Strategie kurz zur Sprache kommen. Diese Strategie hat tatsächlich einen deutlich anderen Ansatzpunkt als die poor substitute-Strategie. Denn sie bestreitet nicht, daß es möglicherweise Maschinen gibt, die das Verhalten eines Menschen vollständig imitieren; sie behauptet nur, daß Menschen trotzdem mehr bzw. etwas anderes sind als solche Maschinen. Grundsätzlich umfaßt die hollow shell-Strategie - bezogen auf die These von Newell und Simon - alle Argumente der Art: Selbst wenn es physikalische Symbolsysteme gibt, die dasselbe input-output-Verhalten wie Menschen zeigen, können Menschen doch keine solchen Systeme sein; denn Menschen haben X, und kein physikalisches Symbolsystem hat X. Ich kann mir vorstellen, daß Argumente dieses Typs KI-Forschern eher unverständlich erscheinen. Für uns Philosophen sind sie jedoch besonders interessant, weil sie uns immer wieder dazu zwingen, unseren Begriffsapparat neu zu überdenken. Ich hoffe, daß im folgenden zumindest ansatzweise klar wird, warum das so ist.

Die üblichen Kandidaten für X in der philosophischen Diskussion sind Bewußtsein, Subjektivität und neuerdings dank Searle Intentionalität. (Die Liste ließe sich sicher noch verlängern; aber das scheinen mir doch die wichtigsten Kandidaten zu sein.) Im Fall X=Bewußtsein lautet

das Argument also: Menschen haben Bewußtsein; kein physikalisches Symbolsystem hat Bewußtsein; also sind Menschen keine physikalischen Symbolsysteme. Das Problem ist, daß die KI-Forschung und die Cognitive Science uns zum Problem des Bewußtseins herzlich wenig zu sagen haben. Haugeland bemerkt, daß in der Spezialliteratur der Terminus "Bewußtsein" manchmal schon fast als "schmutziges" Wort angesehen wird (Haugeland 1981, 32). Kann es also nicht sein, daß im Bereich der KI-Forschung und der Cognitive Science etwas für den Menschen sehr Wichtiges ausgeklammert wird? Wenn man über diese Frage nachdenkt, stößt man jedoch sofort auf ein neues Problem: auch die anderen Wissenschaften - die Philosophie eingeschlossen - haben bisher zum Problem des "Bewußtseins" nichts besonders Erhellendes sagen können. Die ganze Sache scheint irgendwie mysteriös und ungreifbar, von welcher Seite man sie auch betrachtet. Und das ist für den Philosophen schon interessant, vielleicht sogar spannend. Denn plötzlich steht er vor der Situation, daß es für einen Begriff, der in der Tradition eine beträchtliche Rolle spielt, weder eine Definition noch auch nur einigermaßen verläßliche Anwendungskriterien gibt. Und dabei schien nichts klarer zu sein als dieser Begriff.

Häufig trifft man in dieser Situation auf die Reaktion: Was Bewußtsein ist, das weiß ich doch von mir selbst; wenn andere Bewußtsein haben, dann haben sie daher genau das, was ich von meinem eigenen Fall als Bewußtsein kenne. Aber diese Art, für psychische Begriffe gewissermaßen nach innen gerichtete Hinweisdefinitionen zu geben, ist spätestens seit Wittgenstein immer wieder heftig kritisiert worden. Andererseits stößt jedoch auch die quasi behavioristische Antwort "Gleiches Verhalten, gleiche psychische Zustände" auf Skepsis. Und ich selbst neige zwar zu dieser zweiten Sichtweise, spüre dabei aber immer noch ein leises Unbehagen. Ist es nicht doch möglich, daß eine Maschine sich zwar genau so verhält wie ich, aber trotzdem nicht dasselbe fühlt und empfindet? Aber wenn das möglich ist, worin besteht dann dieses zusätzliche Element des Fühlens und Empfindens? Es bleibt keine andere Wahl, als das Problem des Bewußtseins noch einmal ganz von vorne zu durchdenken. Und dabei, scheint mir, kann es wieder zu einer fruchtbaren Zusammenarbeit von Philosophie und KI-Forschung kommen. Denn wir wissen über Bewußtsein zwar nicht viel. Aber in irgendeiner Weise hat Bewußtsein offenbar etwas mit einem spezifischen Selbstzugang intelligenter Systeme zu ihren eigenen Zuständen zu tun. Wenn die KI-Forscher Modelle für verschiedene Arten des Selbstzugangs anbieten würden, dann könnten die Philosophen in Auseinandersetzung mit diesen

Modellen versuchen herauszufinden, was für einen Selbstzugang wesentlich ist, der als Bewußtsein qualifiziert werden kann. D.h. die konkreten Modelle könnten den Philosophen dazu dienen, die eigene Begrifflichkeit klarer zu fassen und besser in den Griff zu bekommen.

Ich will die Kritik John Searles an der KI-Forschung benützen, um diesen Punkt noch von einer anderen Seite her zu beleuchten. Searles Ausgangspunkt sind die Dialogsysteme von Roger Schank. Aber er selbst sagt, seine Überlegungen träfen auch auf andere Systeme dieser Art zu, z.B. auf das System SHRDLU von Terry Winograd. Ich nehme also an: auch auf Systeme wie HAM-ANS oder HAM-RPM. Seine These im Hinblick auf alle diese Systeme lautet folgendermaßen.

SEARLE:

> 1. Von keinem der genannten Systeme kann man im Wortsinn sagen, sie verstünden die ihnen gestellten Fragen oder die ihnen vorgelegten Geschichten.
>
> 2. Die Programme dieser Maschinen (und alle ähnlichen Programme) erklären auf keine Weise die menschliche Fähigkeit, Fragen und Geschichten zu verstehen und entsprechenden Antworten zu geben.

Für diese These argumentiert Searle mit einem Gedankenexperiment, das unter dem Namen "Chinese Room" berühmt geworden ist. Er konstruiert eine Situation, in der er, Searle, in Analogie zu bekannten Frage-Antwort-Systemen Folgen von chinesischen Schriftzeichen herstellt, die von Außenstehenden als Antworten auf von ihnen gestellte Fragen aufgefaßt werden können. Im einzelnen schildert Searle diese Situation so:

a) er, Searle, befindet sich in einem abgeschlossenen Raum, in dem sich ein Stapel mit in Chinesisch geschriebenen Texten befindet;

b) er, Searle, versteht kein Wort Chinesisch und ist nicht einmal in der Lage, chinesische von japanischen Schriftzeichen zu unterscheiden;

c) in seinen Raum wird noch ein zweiter Stapel mit chinesischen Texten hereingereicht zusammen mit einer in Englisch geschriebenen (also für Searle verständlichen) Menge von Regeln, die ihn darüber aufklären, wie Texte aus dem ersten Stapel mit Texten aus dem zweiten Stapel in Verbindung gebracht werden können;

d) nun wird noch ein dritter Stapel von chinesischen Zeichen in den Raum gereicht zusammen mit einer weiteren in Englisch geschriebenen Menge von Anweisungen, die es ermöglichen, Elemente des dritten Stapels mit Elementen aus den beiden anderen Stapeln in Verbindung zu bringen, und die Searle außerdem sagen, welche Symbolfolgen er als Reaktion auf bestimmte Symbolfolgen im dritten Stapel zurückgeben soll.

Searle fügt ironisch hinzu:

> "Unknown to me, the people who are giving me all of these symbols call the first batch 'a script', they call the second batch 'a story', and they call the third batch 'questions'. Furthermore, they call the symbols I give them back in respond to the third batch 'answers to the questions', and the set of rules in English that they gave me they call 'the program'." (Haugeland 1981, 284 f.)

Der Punkt dieses Gedankenexperiments liegt auf der Hand. In der angegebenen Situation wird Searle als Reaktion auf ihm in schriftlicher Form gestellte chinesische Fragen, die ihm selbst nur als Folgen von durch ihre graphische Form charakterisierten Symbolen erscheinen, Folgen von solchen Symbolen zurückgeben, die von den Fragestellern als Antworten aufgefaßt werden können. Und das alles, ohne daß Searle selbst deshalb auch nur ein Wort Chinesisch verstehen müßte. Also verstehen auch die Schankschen Dialogsysteme nichts von den Fragen und Geschichten, die in sie hineingefüttert werden.

> "... it seems to me obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. Schank's computer for the same reasons understands nothing of any stories whether in Chinese, English, or whatever, since in the Chinese case the computer is me; and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing." (Haugeland 1981, 285)

Ich glaube nicht, daß Searle mit dieser Argumentation Recht hat, und ich denke, daß schon die ersten Gegenargumente, auf die Searle in seinem Aufsatz selbst noch eingeht, zeigen, daß er es sich zumindest

zu leicht macht. Aber hier möchte ich mehr auf den grundsätzlichen Punkt eingehen, der in Searles Argument deutlich wird. Wenn man es richtig durchdenkt, dann kann Searles These nämlich als direkter Angriff auf die These von Newell und Simon verstanden werden. Denn seine eigentliche Botschaft lautet: Computer können nichts verstehen, gerade weil sie nur physikalische Symbolsysteme sind; denn physikalische Symbolsysteme können nur syntaktisch arbeiten; zum Verstehen gehört jedoch das Erfassen von Bedeutung, also Semantik.

Meiner Meinung nach liegt die Hauptschwäche dieser Botschaft darin, daß Searle so tut, als wüßte er, was Semantik und damit auch was das Verstehen einer sprachlichen Äußerung ist. Tatsächlich scheint mir das jedoch fast ebenso (wenn auch nicht ganz so) unklar zu sein wie die Antwort auf die Frage, was Bewußtsein ist. Andererseits - und darin besteht wohl die Suggestivkraft der Searleschen Auffassung - scheint es wie beim Bewußtsein auch beim Erfassen der Bedeutung einer Nachricht vielen intuitiv klar zu sein, daß es sich hier um ein Phänomen handelt, das irgendwie mehr sein muß als das bloße physikalische Umformen von physikalischen Mustern. Schon Bedeutungen an sich scheinen vielen etwas nicht Physikalisches - spätestens seit Frege in einer anderen Welt Angesiedeltes - zu sein. Und angesichts dieser Vormeinung überrascht uns die KI-Forschung tatsächlich mit einem ganz ungewöhnlichen Semantik-Konzept. Winograd etwa formuliert das so:

> "There has never been a clear definition of what the field of 'semantics' should cover, but attempts to program computers to understand natural language have clarified what a semantic theory has to do ... In practical terms, we need a <u>transducer</u> that can work with a syntactic analyzer, and produce data which is acceptable to a logical deductive system. Given a syntactic parser with a grammar of English, and a deductive system with a base of knowledge about particular subjects, <u>the role of semantics is to fill the gap between them</u>." (Winograd 1972, 28; Hervorh. vom Verf.)

Das ist nun wirklich verblüffend. Denn diese Auffassung scheint im Kern doch auf die These hinauszulaufen, daß das Verstehen eines Satzes darin besteht, daß er in eine interne Repräsentation übersetzt wird. Also ganz grob gesprochen: Verstehen wird gleichgesetzt mit der Übersetzung von einer Sprache in eine andere. Wenn das so ist, dann scheinen aber zumindest zwei Fragen nahezuliegen:

- Muß das System nicht, um die eine in die andere Sprache übersetzen zu können, die erste Sprache zunächst verstanden haben?

- Wie versteht das System denn die zweite Sprache, d.h. seine eigenen inneren Repräsentationen?

Die überraschende Antwort der KI-Forschung auf diese Fragen scheint mir zu sein:

- Die Übersetzung der externen Sprache in die interne Repräsentation geschieht nach rein formalen Regeln.

- Die interne Repräsentation muß selbst nicht wieder verstanden werden. Sie ist in gewisser Weise das Verstehen.

Das ist wohl der eigentliche Clou dieses Semantikkonzepts. Und das ist wohl auch der Punkt, den Searle nicht akzeptieren will. Das Verstehen einer sprachlichen Äußerung - so scheint Searle zu denken - kann nicht darin bestehen, daß sie in eine innere Repräsentation überführt wird. Denn diese ist letzten Endes ein rein physikalisches Muster, hat also selbst keine Bedeutung und kann deshalb auch nicht das Verstehen einer Bedeutung ausmachen. Auf diese Überlegung wiederum scheint mir die Antwort der KI-Forscher zu sein: die interne Repräsentation darf natürlich nicht für sich isoliert als physikalisches Muster betrachtet werden; die Repräsentation einer Bedeutung kann sie nur deshalb sein, weil sie in ein Gesamtsystem in bestimmter Weise integriert ist. Sie kann die Bedeutung eines externen Satzes ausmachen, weil sie erstens über das syntaktische und semantische System so mit der externen Sprache verbunden ist, daß alle externen Sätze mit derselben Bedeutung in dieselbe und alle externen Sätze mit verschiedenen Bedeutungen in verschiedene Repräsentationen überführt werden, und weil aus ihr zweitens mit Hilfe des deduktiven Systems Strukturen gewonnen werden können, die im Hinblick auf ihre Bedeutungen als logische Folgerungen dieser Repräsentation betrachtet werden können.

Diese - zugegebenermaßen etwas verkürzte - Antwort scheint mir jedoch noch nicht ausreichend. Und deshalb scheint mir Searles Argumentation im Hinblick auf die genannten Frage-Antwort-Systeme auch in gewisser Weise berechtigt zu sein. Denn daß ein Satz eine bestimmte Bedeutung hat, beinhaltet auch, daß er sich in bestimmter Weise auf die Welt bezieht. Und ein möglicher Weltbezug der zuvor besprochenen Repräsen-

tationen ist bisher noch gar nicht in den Blick gekommen. Das war schon deshalb ausgeschlossen, weil die zur Debatte stehenden Systeme weder über entsprechende Sensoren noch über entsprechende Effektoren verfügen. Mir scheint jedoch, daß man frühestens dann davon sprechen kann, daß ein System z.B. den Satz "Auf dem Tisch steht eine Vase" versteht, wenn dieses System für den Fall, daß seine optischen Sensoren auf einen Tisch gerichtet sind, auf dem eine Vase steht, eine innere Repräsentation aufbaut, die es dem System gestattet, etwa auf die Frage "Steht auf dem Tisch eine Vase?" mit "Ja" zu antworten.

Aber ich will auch diesen Punkt hier nicht weiter verfolgen. Denn wir müssen zum Ausgangspunkt dieser Überlegungen, zur hollow shell-Strategie, zurückkommen. Diese Strategie provoziert, wie mir scheint, mit ihrem Argument "Menschen haben X; physikalische Symbolsysteme können X nicht haben" Fragen, die in erster Linie an die Philosophie gerichtet sind und die wohl auch nur von der Philosophie selbst beantwortet werden können. Denn diese Fragen beziehen sich ausdrücklich nicht auf äußerlich feststellbare Fähigkeiten, d.h. nicht auf ein bestimmtes input-output-Verhalten. Ich sehe also nicht, daß die KI-Forschung hier direkt etwas zur Klärung beitragen könnte. Die Philosophie dagegen ist angesprochen, weil die Argumente der hollow shell-Strategie erst beurteilt werden können, wenn man weiß, was es heißt, X zu haben. Und für Fragen dieser Art scheint mir die Philosophie zuständig zu sein. Andererseits kann jedoch auch bei solchen Fragen die interdisziplinäre Zusammenarbeit Nutzen bringen. Denn wenn auch die KI-Forschung philosophische Fragen wie "Was ist Bewußtsein?" oder "Was ist Sprachverstehen?" nicht beantworten kann, so führen doch ihre Versuche, z.B. sprachverstehende Systeme oder vielleicht sogar Systeme, die Bewußtsein zeigen, zu konstruieren, zu Ergebnissen, die von der Philosophie diskutiert werden sollten. Denn gerade die Diskussion konkreter Fälle kann bei der Beantwortung der Frage, wann liegt X vor und wann nicht, viel helfen. So wie die zuvor referierte Diskussion meiner Meinung nach viel bei der Beantwortung der Frage helfen kann, was es denn nun wirklich heißt, daß jemand etwas versteht.

Auf einen kurzen Nenner gebracht: Ob es Maschinen mit einem bestimmten input-output-Verhalten gibt, das ist meiner Meinung nach eine Frage, die in erster Linie von der KI-Forschung beantwortet werden muß. Ob - und gegebenenfalls unter welchen Bedingungen - man jedoch Maschinen bestimmte mentale Eigenschaften wie Bewußtsein, Subjektivität und Intentionalität oder bestimmte mentale Zustände wie Wissen, Wollen und

Fühlen zuschreiben kann, das ist eine Frage, die die Philosophie beantworten muß. Allerdings sollte sie sich auch bei der Beantwortung dieser Fragen nicht von einer interdisziplinären Zusammenarbeit mit KI-Forschern und z.B. auch mit Neurophysiologen abhalten lassen.

## Literaturverzeichnis

(Crombie 1977) Crombie, A.C.
Von Augustinus bis Galilei, Munchen 1977

(Cummins 1983) Cummins, R.
The Nature of Psychological Explanation, Cambridge, Mass. 1983

(Dennett 1981) Dennett, D.C.
"Three Kinds of Intentional Psychology", in R.A. Healey (ed.), Reduction, Time and Reality, Cambridge 1981, 37-61

(Descartes 1637) Descartes, R.
Discours de la méthode, OEuvres de Descartes, hrsg. von Ch. Adam und P. Tannery, Paris 1964-1974, Band VI (fr./dt. Ausgabe, übers. und hrsg. von L. Gäbe, Hamburg 1969, Meiner PhB 261)

(Dreyfus 1979) Dreyfus, H. L.
What Computers Can't Do, New York 1979 (dt. Ausg.: Die Grenzen künstlicher Intelligenz, Königstein/Ts. 1985; zitiert nach der dt. Ausgabe)

(Haugeland 1978) Haugeland, J.
"The Nature and Plausibility of Cognitivism", in The Behavioral and Brain Sciences 1 (1978), 215-226 (wieder abgedr. in Haugeland 1981, 243-281)

(Haugeland 1981) Haugeland, J. (ed.)
Mind Design, Cambridge, Mass. 1981

(Newell/Simon 1976) Newell, A. und H.A.Simon
"Computer Science as Empirical Inquiry: Symbols and Search", in Communications of the Association for Computing Machinery 19 (März 1976), 113-126 (wieder abgedr. in Haugeland 1981, 35-66)

(Searle 1980) Searle, J.
"Minds, Brains, and Programs", in The Behavioral and Brain Sciences 3 (1980) (wieder abgedr. in Haugeland 1981, 282-306)

(Winograd 1972) Winograd, T.
"Understanding Natural Language", in Cognitive Psychology 3 (1972), 1-191