Geert Keil
Was Roboter nicht können
Die Roboterantwort als knapp mißlungene Verteidigung der starken KI-These

1

Theoretiker der *Künstlichen Intelligenz* und deren Wegbegleiter in der Philosophie des Geistes haben in den vergangenen Jahren auf verschiedene Weise auf Kritik am ursprünglichen Theorieziel der KI reagiert. Eine dieser Reaktionen ist die Zurücknahme dieses Theorieziels zugunsten der Verfolgung kleinerformatiger Projekte. Eine andere Reaktion ist die Propagierung konnektionistischer Systeme, die mit ihrer dezentralen Arbeitsweise die neuronalen Netze des menschlichen Gehirns besser simulieren sollen. Eine weitere ist die sogenannte *robot reply*[1], mit der ich mich im folgenden beschäftigen werde. Ich werde in diesem Aufsatz die These vertreten, daß der Roboterantwort eine richtige Intuition zugrunde liegt, von der die Roboterfreunde sich aber zu einer kurzschlüssigen Folgerung verleiten lassen. Im Schlußteil des Aufsatzes werde ich zusätzlich behaupten, daß die Roboterantwort, sei sie nun überzeugend oder nicht, eine Abkehr vom ursprünglichen Theorieziel der KI darstellt.

Was aber ist das ›ursprüngliche Theorieziel‹ der mittlerweile ironisch als GOFAI[2] titulierten KI? In erster Annäherung mag man zu der Antwort neigen, es bestehe in der Konstruktion von Maschinen, deren Leistungen denen eines intelligent handelnden Menschen gleichkommen. Dieses Projekt aber ist kein Theorieziel, sondern ein praktisch-technisches. Allerdings enthält es konzeptuelle und theoretische Anteile. Um sie sichtbar zu machen, muß man die Frage stellen, was mit der Rede von intelligenten Leistungen *gemeint* ist. Der Feld-, Wald- und Wiesendefinition von ›künstlicher Intelligenz‹ zufolge ist eine Maschine bekanntlich dann ›intelligent‹ zu nennen, wenn sie Verhaltensleistungen erbringt, die, würden sie von einem Menschen erbracht,

1 Der Ausdruck stammt wohl von John Searle.
2 ›Good Old-Fashioned Artificial Intelligence‹ (John Haugeland).

Intelligenz erfordern würden. An dieser Bestimmung ist so gut wie alles problematisch. Sie definiert nicht, was Intelligenz ist, sondern setzt ein Verständnis menschlicher Intelligenz voraus, im Rekurs auf welches künstliche Intelligenz bestimmt werden soll. Es ist von ›gleichen Verhaltensleistungen‹ die Rede, nicht aber davon, wie weit sich die Gleichheit erstrecken soll und woran sie sich bemißt. Was die erste Schwäche betrifft, so plädiere ich dafür, großzügig über sie hinwegzusehen. Was Intelligenz ist, ist notorisch schwierig zu bestimmen, und die Frage erscheint im Kontext der jüngeren KI-Debatte auch merkwürdig irrelevant. Würde man heute noch einmal einen neuen Titel für diese Debatte suchen, würde man auf den Intelligenzbegriff vermutlich nicht mehr zurückgreifen. In der Philosophie des Geistes fragt man seit geraumer Zeit nach Bedingungen für die Zuschreibung von ›mentalen Zuständen‹ oder die Einnahme des ›intentionalen Standpunkts‹, was gegenüber der Frage, ob Maschinen intelligent sind, eine zweifache Akzentverschiebung darstellt. Zum einen wird im Zuge des ›semantischen Aufstiegs‹ (Quine) die naiv-realistische Frage nach dem *Besitz* einer Fähigkeit durch die nach den Bedingungen ihrer Zuschreibbarkeit ersetzt. Zum anderen verpflichtet man sich mit dem Zuschreiben mentaler Zustände nicht mehr auf einen bestimmten *Grad* an Intelligenz oder Vernünftigkeit. Die Frage lautet schließlich, ob wir einer Maschine überhaupt Überzeugungen, Wünsche und Absichten zuschreiben sollten. Demgegenüber erscheint die Frage nach dem Intelligenzgrad eines Systems, welches solche Zustände hat, als nachgeordnet. Auch dummen Menschen schreiben wir schließlich mentale Zustände zu – warum dann nicht auch dummen Maschinen? Die KI wäre schon froh, könnte sie eine Maschine bestimmte Leistungen vollbringen lassen, die, würden sie von einem Menschen erbracht, keine außergewöhnliche Intelligenz erfordern würden.

Allzu groß sollte man den Unterschied zwischen der alten und der neuen Formulierung freilich nicht veranschlagen. Einem System mentale Zustände zuzuschreiben verhält sich zu der Auszeichnung des Systems als intelligent oder vernünftig nicht einfach wie eine Vorbedingung. Der Grund dafür ist in der Philosophie des Geistes vor allem von Davidson herausgearbeitet worden. Mentale Zustände werden nicht als einzelne zugeschrieben, sondern *en masse* und systematisch. Ein Systemverhalten, wel-

ches nur die Zuschreibung einer einzigen Überzeugung legitimierte, legitimiert auch diese nicht. Wir hätten keinen Grund, diese ›Überzeugung‹ mit irgendwelchen derjenigen Worte zu beschreiben, mit denen wir unsere eigenen Überzeugungen ausdrücken. Überzeugungen, Wünsche und Absichten sind individuiert über ihre semantischen Gehalte und damit über ihre logischen und inferentiellen Beziehungen zu anderen Überzeugungen, Wünschen und Absichten. Infolgedessen steht ihre Zuschreibung unter gewissen Restriktionen. Man wird beispielsweise niemandem die Überzeugung zuschreiben, daß dort ein roter Tisch steht, dem man nicht auch die weitere Überzeugung zuzuschreiben bereit ist, daß dort ein Tisch steht. (Einzelne solcher Fälle lassen sich mit viel Phantasie konstruieren; wären sie die Regel, würde das fragliche System uninterpretierbar, und man müßte aufhören, es überhaupt als ein geistbegabtes Wesen zu betrachten. Man würde also den probeweise eingenommenen intentionalen Standpunkt wieder aufgeben.[3]) Nennen wir diese Restriktion die der *holistischen Zuschreibungsbedingungen mentaler Prädikate*.

Diese holistischen Zuschreibungsbedingungen sind Davidson zufolge dafür verantwortlich, daß Mengen von Überzeugungen, die wir überhaupt als solche interpretieren können, erstens eine gewisse Kohärenz aufweisen und zweitens zu einem großen Teil wahr sein müssen. Es ist demnach nicht miteinander vereinbar, einem Wesen mentale Zustände zuzuerkennen und es zugleich für radikal unvernünftig oder uninterpretierbar zu halten. Vielmehr müssen die Zuschreibungsbedingungen so angesetzt werden, daß die Klassen der Wesen mit mentalen Zuständen und der vernunftbegabten Wesen sich als koextensiv erweisen. Irrationalität im Einzelfall ist damit nicht ausgeschlossen, im Gegenteil: Ein Verhalten irrational zu nennen setzt voraus, daß man das fragliche System zuvor als Kandidaten für die Beurteilung nach Rationalitätsstandards ausgezeichnet hat. Es ist deshalb hilfreich,

3 »Wenn wir keine Möglichkeit finden, die Äußerungen und das sonstige Verhalten eines Geschöpfs so zu interpretieren, daß dabei eine Menge von Überzeugungen zum Vorschein kommt, die großenteils widerspruchsfrei und nach unseren eigenen Maßstäben wahr sind, haben wir keinen Grund, dieses Geschöpf für ein Wesen zu erachten, das rational ist, Überzeugungen vertritt oder überhaupt etwas sagt« (Davidson, »Radikale Interpretation«, S. 199).

das Prädikat ›rational‹ mit Indizes zu versehen, um einen *weiteren* Sinn, wie er etwa in *animal rationale* vorliegt, von einem *engeren* Sinn, der Äußerungen oder Handlungen eines animal rationale ›irrational‹ zu nennen erlaubt, zu unterscheiden.[4] Rational$_1$ hieße dann soviel wie ›beurteilbar nach Rationalitäts$_2$-standards‹, was für alle Wesen zutreffen muß, denen wir überhaupt mentale Zustände zuschreiben. Rationalität$_2$ ist hingegen eine Sache des Grades. *Wie groß* die Kohärenz der in einer radikalen Interpretation zugeschriebenen Überzeugungsmenge eines Systems sein muß und *wie viele* dieser Überzeugungen wahr sein müssen (um Davidsons Standards aufzugreifen), um die entsprechenden Zuschreibungen zu legitimieren, ist eine offene Frage, die hier nicht beantwortet werden muß. Was nun das Prädikat ›intelligent‹ betrifft, so funktioniert es in der Alltagssprache wie rational$_2$, nicht wie rational$_1$, denn die Charakterisierung eines Verhaltens als ›intelligent‹ setzt die Anwendbarkeit eines entsprechenden Standards schon voraus. (Gegenprobe: Die Qualifizierung eines Verhaltens als ›dumm‹ setzt denselben Standard voraus!) Intelligenztests macht man deshalb mit Personen und nicht mit Gartenschläuchen. Nach Intelligenz fragt man typischerweise, *nachdem* man das Untersuchungsobjekt als Kandidaten für eine entsprechende Beurteilung ausgezeichnet hat.

Nun mag der Fall bei Personen einerseits und Gartenschläuchen andererseits klarer liegen als bei Computern, Robotern, höheren Säugetieren und potentiellen Marsbewohnern. Tatsächlich sind die beiden genannten Schritte auch nicht völlig unabhängig voneinander. Wenn die Rekrutierung eines Kandidaten für den Klub der intelligenten, vernünftigen oder geistbegabten Wesen nicht aufgrund eines kruden biologischen Kriteriums erfolgen soll, bleibt wohl nur die Überprüfung einzelner seiner Äußerungen und Handlungen auf ihre Rationalität$_2$. Solche Proben aufs Exempel erfolgen jeweils im Zuge eines hermeneutischen Vorgriffs: in der Erwartung nämlich, daß die zunächst nur probeweise Einnahme des ›intentionalen Standpunkts‹ (Dennett) erfolgreiche Voraussagen ermöglicht und sich damit nachträglich als gerechtfertigt erweist. Die beiden Schritte sind also rückgekoppelt. Der Prognoseerfolg allein reicht allerdings zur Rechtfertigung des intentionalen Standpunkts nicht aus; Dennett

4 Hierin folge ich Schnädelbach, »Rationalität und Normativität«.

selbst hält die Zuschreibung mentaler Zustände erst dann für legitim, wenn andere, nichtintentionale Erklärungsstrategien versagen. Es ist deshalb auf einer sorgfältigen Begründung dafür zu bestehen, ein System überhaupt als Kandidaten für mentale Zuschreibungen zu akzeptieren. Viele Theoretiker der KI überspringen diesen ersten Schritt und ziehen ihn mit dem zweiten zusammen, in der irrigen Annahme, man könne ein System unmittelbar darauf überprüfen, welche ›Leistungen‹ es erbringt und ob diese intelligent zu nennen seien. Diese Annahme entspringt einem behavioristischen Vorurteil, das die vielbeschworene ›kognitive Wende‹, als deren Kind man die Kognitionswissenschaften ansieht, erstaunlich unbeschadet überdauert hat. Man kann einem Systemverhalten eben nicht ansehen, *als was es zu interpretieren* und welches Vokabular dafür in Anschlag zu bringen ist.

Das Ziel der GOFAI, Maschinen zu konstruieren, deren Leistungen denen eines intelligent handelnden Menschen nicht nachstehen, war mehr oder weniger ausdrücklich mit der Erwartung verbunden, daß sich kognitive oder intellektuelle Kompetenzen in einer Weise auffassen lassen, die eine *separate* Modellierung erlaubt. Der menschliche Geist muß aus seiner Verstrickung mit den übrigen Merkmalen und Eigenschaften des Menschen herausgelöst werden, damit es so etwas wie Künstliche Intelligenz überhaupt geben kann. In der Philosophie des Geistes wurde dieses Programm durch die funktionalistische Auffassung des Mentalen gestützt, die es erlauben sollte, in der Charakterisierung mentaler Prozesse von allem Kontingenten zu abstrahieren. Als kontingent dürfen wir dem Funktionalismus zufolge die meisten physiologischen Eigenschaften des Menschen ansehen. Dieser Auffassung war auch Alan Turing. Einschlägig ist seine Bemerkung über die Irrelevanz der technischen Schwierigkeit, ein von der menschlichen Haut ununterscheidbares Material herzustellen. Es wäre in Turings Augen schlicht unfair, wenn man die KI an solchen Kontingenzen scheitern ließe.[5] Angestrebt ist schließlich die künstliche *Intelligenz*, nicht der künstliche Mensch. Wenn dieses Projekt überhaupt irgendeine Aussicht auf Erfolg haben soll, müssen sich kognitive Fähigkeiten aus ihrer kontingenten Verkörperung in der Biologie des homo sapiens

5 Vgl. Turing, »Computing Machinery and Intelligence«, S. 434.

herauslösen lassen. Auch das *setting* des Turing-Tests war in dieser Absicht gewählt, eine »fairly sharp line between the physical and the intellectual capacities of man«[6] zu ziehen. Wäre dies nicht möglich, behielte der sogenannte Artenchauvinismus recht.

Die GOFAI hat ihr in den fünfziger Jahren formuliertes Ziel der Konstruktion einer Maschine mit menschenähnlicher, allgemeiner Intelligenz selbst bei Anlegung eines behavioristischen Maßstabs bis heute nicht erreicht. Dies wird auch von ihren Sympathisanten längst zugegeben. Die Euphorie der frühen Jahre ist weitgehend abgeklungen, die Diskussionslage ist gekennzeichnet von vielfachen Abschwächungen, Präzisierungen und Modifikationen, von denen oft nicht klar ist, ob man sie als Elaborierungen des ursprünglichen Programms oder aber als Rückzugspositionen ansehen soll. Unter dem Titel ›KI‹ werden heute kleinerformatige Projekte verfolgt als die Entwicklung eines Systems mit menschenähnlich intelligentem Gesamtverhalten. (Daneben gibt es nach wie vor Propheten wie Hans Moravec oder Marvin Minsky, die an den großen Visionen festhalten und sie zum Teil noch in Science-fiction-Manier überbieten. Solche Propheten wird es immer geben, aber sie prägen nicht mehr die innerwissenschaftliche Diskussion der Informatik.)

Manche Kognitionswissenschaftler gehen so weit, die ganze Debatte um die KI für veraltet und allenfalls noch von philosophischem Interesse zu erklären. Über solche Urteile sollten Philosophen sich nicht grämen, denn die mit dem GOFAI-Programm verbundene *philosophische* Frage bleibt eine legitime. Vermutlich wird das Fragen danach, ob man Artefakten kognitive Leistungen zusprechen kann, eher die Disziplin der ›Kognitionswissenschaft‹ überdauern als umgekehrt.[7] Die Roboterant-

6 Ebd.

7 Was an der Rede von der ›Kognitionswissenschaft‹, zumal im Singular, stört, ist die Suggestion, man habe es dort mit einem komplexen, aber einheitlichen Gegenstand namens ›Kognition‹ zu tun, zu dessen gemeinschaftlicher Erforschung sich Disziplinen wie kognitive Psychologie, Linguistik, Informatik, Neurobiologie und vielleicht sogar Philosophie des Geistes und Erkenntnistheorie zusammengeschlossen hätten. Diese Sicht der Dinge ist konstitutionstheoretisch naiv, denn der Ausdruck ›Kognition‹ bezeichnet keinen einheitlichen Gegenstand. Man darf vermuten, daß die Kognitionswissenschaft(en) sich nach dem Abklingen der Interdisziplinaritätseuphorie wieder in ihre Bestandteile auflösen werden. Es wird dann wieder sichtbar wer-

wort als neuerlicher Versuch der Verteidigung der starken KI-These ist selbst ein Indiz für fortwährenden Klärungsbedarf in Grundsatzfragen.

Die Roboterantwort besteht aus zwei Elementen. Sie enthält (a) das Zugeständnis, daß das Systemverhalten eines wie auch immer programmierten konventionellen Digitalrechners mit von-Neumann-Architektur nicht schon menschenähnliche Intelligenz aufweist. Sie enthält (b) die Behauptung, daß es für bestimmte Arten von Maschinen *doch* zur Intelligenz reicht. In die Liga der intelligenten Wesen könnten Maschinen genau dann aufsteigen, wenn sie Roboter sind. Damit ist gemeint: wenn sie über Wahrnehmungskomponenten (Rezeptoren) und Handlungskomponenten (Effektoren) verfügen, mit deren Hilfe sie aktiv in kausale Interaktionen mit ihrer Umwelt eintreten können.

2

Man kann diesen Zug als eine Reaktion auf philosophische Kritik an der KI ansehen, etwa als Antwort auf Searles Argument vom Chinesischen Zimmer. Ebenso ist die Roboterantwort jedoch durch Schwierigkeiten motiviert, auf die die KI in der Praxis zwangsläufig selbst stoßen mußte. Es sind dies Schwierigkeiten der Wissensrepräsentation und des Wissenszugriffs, die man unter dem Titel des *Frame*-Problems[8] zusammenfassen kann: Während ein Mensch in einer komplexen und zugleich dynamischen Umwelt agiert, verändern sich einige Situationsparameter, andere bleiben konstant. Tatsächlich ändert sich in jeder Situation nur weniges von dem, was sich ändern *könnte*. Diese Gewißheit nützt dem Handelnden aber so lange nichts, als er vorher nicht absehen kann, was sich ändern wird und was nicht. Nun verfügen Menschen offenbar über Fähigkeiten des flexiblen Reagierens auf unvorhergesehene Situationsveränderungen. Worin bestehen sie, welche Wissensbasis liegt ihnen zugrunde, und wie lassen sie sich modellieren? – In den siebziger Jahren hat man diese Fragen durch das Aufstellen von ›Skripten‹ oder ›Schemata‹ zu lösen

den, wie vielfältig die Explananda der genannten wissenschaftlichen Disziplinen sind und bleiben.

8 Vgl. Pylyshyn (Hg.), *The Robot's Dilemma*.

versucht. In diesen Skripten sollten die Elemente eines bestimmten Situationstyps vollständig verzeichnet sein, die üblicherweise auftretenden Komplikationen eingeschlossen. So entstanden Drehbücher für Situationstypen wie ›Restaurantbesuch‹ oder ›Geburtstagsparty‹. Der Anspruch dieser Skripte war hochgesteckt: »[T]he restaurant script«, so verkündet dessen Autor Roger Schank, »contains all the information necessary to understand the enormous variability of what can occur in a restaurant.«[9]

Dies war natürlich eine maßlose Übertreibung, wie man sich durch eine einfache Überlegung vor Augen führen kann. Wenn man in einem Restaurant sitzt, kann jederzeit eine Person hereinkommen, mit der man nicht gerechnet hat und auch nicht rechnen konnte, und es kann sich ein Gespräch über einen denkbar entlegenen Redegegenstand entwickeln. Die Akteure müssen dann auf Wissensbestände zurückgreifen, von denen sie vorher nicht wissen konnten, daß sie sie brauchen würden, und die in keinem von Schanks Skripten verzeichnet sind. Es besteht kein Grund zu der Annahme, es könne eine endliche Liste oder ein ›Skript‹ dessen geben, was in einer bestimmten Situation passieren könnte. Situationen begegnen uns in der Welt nicht metaphysisch individuiert, sondern sie sind gegenüber ihrer Umwelt offen. Im Grunde ›gibt‹ es gar keine Situationen, jedenfalls dann nicht, wenn man darunter wohldefinierte Mikrowelten versteht, deren Inventar und deren Regeln sich vollständig verzeichnen lassen. Was wir Situationen nennen, sind unter pragmatischen Gesichtspunkten vorgenommene Abstraktionen. Situationsdefinitionen sind komplexitätsreduzierende Maßnahmen von Soziologen und Lebenswelttheoretikern. Diese Abstraktionen sollte man nicht mit den Umgebungen verwechseln, in denen Menschen handeln. Wenn wir nichts als Schanks Skripte in der Hand hätten, würden wir bei vielen Restaurantbesuchen kläglich scheitern.[10]

Tatsächlich verfügen wir über Kompetenzen, die die Orientierung in einer nicht vollständig spezifizierten Umwelt ermög-

9 Zitiert nach Dreyfus/Dreyfus, »Coping With Change: Why People Can and Computers Can't«, S. 165.

10 Wir befänden uns ungefähr in der Lage des Touristen, der in einem fremdsprachigen Land mit den in seinem Sprachführer aufgeführten Redewendungen auskommen muß, komme, was da wolle.

lichen. Diese Kompetenzen sind von phänomenologischen Philosophen beschrieben worden, und Dreyfus hat diese Überlegungen auf das *Frame*-Problem angewandt. Offenbar ist ein beträchtlicher Teil unseres Wissensbestandes nicht in Form expliziter mentaler Repräsentationen gespeichert und steht dennoch auf Anforderung zur Verfügung. So glauben die meisten von uns, daß Rechtsanwälte typischerweise Schuhe tragen, doch »just when, exactly, did this belief get added to your belief store?«[11] Solches ›Wissen‹ läßt sich nicht in einer Überzeugungsliste festhalten, denn lebensweltliches Hintergrundwissen besteht überhaupt nicht aus diskreten Wissenselementen. Große Teile davon werden besser als ein praktisches Können beschrieben denn als ein explizites propositionales Wissen. Dieses *knowing how* ist aufgrund seines dispositionalen Charakters und seiner holistischen Struktur schwierig zu analysieren. Irgendwie bringen wir es fertig, zu den entsprechenden Gelegenheiten Hintergrundüberzeugungen zu aktualisieren, die vermutlich nirgends separat abgespeichert sind, und diese so zu verwerten, daß wir auch bei auftretenden Komplikationen unsere Handlungsziele erreichen.

Die GOFAI hat die Standardisierung der Situationen, in denen sie ihre Elaborate agieren ließ, ihren Kritikern gern als ›ersten Schritt‹ verkauft, dem weitere folgen sollten. Wird man eine Maschine, die schon bescheidene Intelligenz im befehlsgesteuerten Umschichten von bunten Bauklötzen bewiesen hat, nicht auch zu anspruchsvolleren Leistungen anleiten können? Und entwikkelt sich nicht auch menschliche Intelligenz in jeder Ontogenese von neuem aus bescheidenen Anfängen? – Wenn das Gesagte richtig ist, besteht zwischen den beiden Fällen ein eklatanter Unterschied, der die Gradualisierungsstrategie fragwürdig macht: Die wirkliche Welt, in der Menschenkinder allmählich ihre Fähigkeiten entwickeln, ist nicht aus Mikrowelten zusammengesetzt. Sie ist immer schon ganz da, auch wenn sie jeweils nur ausschnitthaft erfahren wird. Eine Mikrowelt hingegen ist ein geschlossener Kosmos. Ihr Inventar und ihre Regeln sind vor dem Chaos ›da draußen‹ geschützt. In diesem Sinne ist ein Restaurant natürlich keine Mikrowelt. Man betrachte demgegenüber das Schachspiel: Es ist vorab klar, daß es für die Schachregeln keinen Unterschied macht, aus welchem Material die

11 Dennett, »Science, Philosophy, and Interpretation«, S. 537.

Figuren bestehen oder was die Spieler gefrühstückt haben. Solche Mikrowelten gibt es aber nur dort, wo man die nötigen Abschottungen via Konvention *hergestellt* hat. Abgeschottet sind auch nur die konstitutiven Regeln des Schachspiels, nicht die strategischen Regeln und schon gar nicht das gesamte Spiel als soziale Veranstaltung. Schon bei der Entwicklung von Expertensystemen (zum Beispiel zur medizinischen Diagnostik) ist es ungleich schwieriger, das Wissen über die jeweiligen Realitätsbereiche gegenüber anderem möglicherweise relevanten Wissen zu isolieren.

Es ist deshalb eine fragwürdige Strategie, Maschinen in konstruierten Mikrowelten agieren zu lassen und dann in einem zweiten Schritt die dort erfolgreichen Problemlösungsverfahren allmählich auf größere Fragmente der wirklichen Welt übertragen zu wollen. Mikrowelten sind keine Fragmente der wirklichen Welt, sie sind Konstrukte. Der Unterschied zwischen einer Mikrowelt und der Welt, in der wir leben, ist kein gradueller.

Descartes hat große Weitsicht bewiesen, als er den wesentlichen Unterschied zwischen Mensch und Maschine in der *Universalität* der menschlichen Vernunft erblickte. Die Vernunft sei ein ›Universalinstrument‹, das auf keine bestimmte Aufgabe zugeschnitten ist. Selbst wenn Maschinen »viele Dinge ebensogut oder vielleicht besser als einer von uns machten«[12], bewiesen sie damit keine menschenähnliche Intelligenz, denn sie würden »unausbleiblich in einigen anderen fehlen«[13] und dadurch zeigen, daß sie »nicht nach Einsicht«[14] handeln. Es ist die Universalität der menschlichen Vernunft, die sie so leistungsfähig macht und ihre Modellierung so schwierig. Wenn man wirklich auf eine menschenähnliche Intelligenz hinauswill, kann man deshalb nicht auf in Mikrowelten erbrachte ›Leistungen‹ verweisen und das *Frame*-Problem erst einmal vertagen. Man sollte die begrifflichen Schwierigkeiten nicht unterschätzen, eine isoliert ausgeübte ›Tätigkeit‹ oder ›Leistung‹ mit einer solchen zu vergleichen, die Teil eines umfassenderen Repertoires ist. Der Bezug auf ein solches Repertoire könnte ja, zum Schrecken für den Behaviorismus, in die Formulierung der Identitätsbedingungen für die einzelne ›Leistung‹ eingehen.

12 Descartes, *Abhandlung über die Methode des richtigen Vernunftgebrauchs*, S. 53.
13 Ebd.
14 Ebd.

Phänomenologische Philosophen haben behauptet, daß ein Teil des zur Bewältigung unvorhersehbarer Situationen erforderlichen ›Wissens‹ überhaupt nicht in unseren Köpfen aufbewahrt sei, sondern *in der Welt*. Falls man dieser Behauptung irgendeinen Sinn abgewinnen kann, ist es vielleicht dieser: Unser *knowing how* bezüglich der Orientierung in nicht vollständig spezifizierten Situationen beinhaltet nicht nur die Fähigkeit, Wissen zu aktualisieren, von dem wir vorab nicht wissen konnten, daß wir es brauchen würden. Wir sind überdies in der Lage, der Handlungssituation beim Auftreten von Komplikationen neue Informationen zu entnehmen und sie mit dem schon Gewußten zu verknüpfen. Es muß also bei Handlungsbeginn nicht alles erforderliche Wissen schon vorliegen. Es wird daher ein wissensbasiertes System mit einer endlichen Basis von Wissenselementen, und sei sie noch so groß, immer einem solchen System unterlegen sein, das in der Lage ist, sich selbständig neue Informationen aus seiner Umwelt zu verschaffen. Mit anderen Worten: »[T]he cheapest store of information about the real world is the real world.«[15] Die ›wirkliche Welt‹ ist nicht nur der billigste, sondern in vielen Fällen der einzige verfügbare Wissensspeicher – diese Überlegung stellt zugleich das Hauptmotiv für den Übergang zur Roboterantwort dar. Margaret Boden hat in diesem Sinne schon 1977 die Vermutung ausgesprochen, daß »many tasks might be feasible only for an active and percipient robot«.[16] Das Sich-Bewährenmüssen in der rauhen Luft der nicht-virtuellen Welt ist einerseits der härteste und am wenigsten angreifbare Intelligenztest. Andererseits stehen erst dort die Mittel bereit, ihn zu bestehen.

3

Es sind zwei Arten von Fähigkeiten, die Vertreter der Roboterantwort für erforderlich halten, um Maschinen in die Liga der intelligenten Wesen aufsteigen zu lassen. Zum einen müßten Maschinen gewisse *Wahrnehmungs*fähigkeiten besitzen, um den besten Informationsspeicher über die wirkliche Welt auch anzapfen zu können. Zum anderen müßten sie aktiv in kausale Interaktionen mit ihrer Umwelt eintreten können, kurz: sie müßten *handeln* können.

15 Boden, *Artificial Intelligence and Natural Man*, S. 438.
16 Ebd.

Ich werde die Roboterantwort in Auseinandersetzung mit den Argumenten diskutieren, die Beckermann, Rheinwald, Dennett und Tetens für sie angeführt haben.

Beckermann und Rheinwald wenden die Roboterantwort auf das Problem des Sprachverstehens an. Im Unterschied zu einer bloßen ›syntaktischen Maschine‹, deren Operationen noch von einer äußeren Instanz semantisch interpretiert werden müssen, könne ein Roboter eine ›semantische Maschine‹ sein, die für ihre Interpretation selbst sorgt. Um einen natürlichsprachigen Satz *S* zu verstehen, muß eine Maschine, so Beckermann, in der Lage sein, das Erfülltsein von dessen Wahrheitsbedingungen festzustellen. Sie muß also »Situationen, in denen *S* wahr ist, von Situationen [...] unterscheiden, in denen das nicht der Fall ist«.[17] Dafür muß die Maschine über eine Wahrnehmungskomponente verfügen, mit deren Hilfe sie Situationsmerkmale erfassen kann, die sie dann mit ihrer Datenbasis abgleicht. Diese Aufgabe darf man ihr nicht abnehmen, indem man etwa das Wissen um das Erfülltsein einer großen Anzahl von Wahrheitsbedingungen vorab programmiert, denn dies wäre ein Rückfall hinter die Einsicht in das *Frame*-Problem.

Andere Autoren haben die Roboterantwort anhand der Kontroverse um die Zuschreibung *mentaler Repräsentationen* durchgespielt. Wir sollten einem System, so Dennett[18], erst dann mentale Repräsentationen zuschreiben, wenn seine kausale Einbettung in die Umwelt eine gewisse Komplexität erreicht hat. Dies sei eine Sache des Grades. Die kausalen Kontakte eines Taschenrechners zu seiner Umwelt sind noch sehr dürftig. Der Zustand eines Thermometers ist schon enger mit Zuständen der Außenwelt korreliert, der eines Thermostaten noch enger. Als nächstes könnte man einen Thermostaten mit mehreren Meßfühlern ausstatten, dann vielleicht mit einer Infrarotkamera, die ihm meldet, ob überhaupt Personen im Raum sind, dann mit einem Spracherkennungssystem, das Äußerungen wie ›Es ist zu kalt hier‹ erfassen und darauf reagieren kann. Wir verschaffen dem System immer mehr Möglichkeiten, Informationen aus seiner Umgebung zu gewinnen und diese in differenziertes Systemverhalten umzu-

17 Beckermann, »Semantische Maschinen«, S. 207.
18 Vgl. Dennett, »True Believers«, S. 29 ff.

setzen. Die Korrelationen seiner internen Zustände mit Zuständen der Außenwelt werden immer vielfältiger, bis wir diese Beziehung schließlich als eine der Repräsentation bezeichnen werden.[19] Zugleich verursacht das Systemverhalten der Maschine immer spezifischere Veränderungen in der Außenwelt. Die Roboterantwort läuft dann auf die Behauptung hinaus,

»daß eine Maschine nur dann über interne Zustände verfügt, die Repräsentationen *für* die Maschine selber darstellen, wenn [...] der Regler [...] einer solchen Maschine rückgekoppelt ist mit Perzeptoren und Effektoren, so daß sich die Maschine entsprechend zielgerichtet-umweltbezogen verhält. Oder, um es auf eine kurze Formel zu bringen: *Nur Roboter sind informationsverarbeitende und informationsverstehende Maschinen.*«[20]

Den Übergang vom konventionellen Digitalrechner zum Roboter kann man nicht als bloße Erweiterung der Eingabe- und Ausgabeeinheit auffassen. Diese Einheiten sind ja (in Form von Tastatur, Lochkartenstanzer, Bildschirm etc.) auch in jeder Maschine mit von-Neumann-Architektur vorhanden. Die Pointe der Roboterantwort besteht in der Emanzipation der Tätigkeit des Roboters von den Vorgaben und Zuschreibungen seines Konstrukteurs: Auf der Input-Seite tritt an die Stelle der Fütterung die selbständige Nahrungsaufnahme, auf der Output-Seite ist kein Interpret mehr nötig, der die kausalen Interaktionen des Roboters mit der Umwelt deutet, da der Roboter mit seinem ›zielgerichtet-umweltbezogenen‹ Verhalten (Tetens) dafür selber sorgt.

John Searle, auf dessen Argument vom Chinesischen Zimmer die Roboterantwort verschiedentlich gemünzt worden ist, zeigt sich von diesem Zug nicht beeindruckt. Er bestreitet, daß robotische Fertigkeiten irgendeinen Unterschied für die Frage ausmachen, ob ein System wirklich eine Sprache versteht oder wirklich intentionale Zustände hat. Tatsächlich habe nichts von dem, womit der Roboter in der Umwelt interagiert, eine Bedeutung *für* den Roboter. Searle: »Die kausale Interaktion zwischen dem Roboter und der Welt ist unerheblich, solange sie nicht in dem einen oder anderen Geist repräsentiert ist.«[21] Searle bestreitet, daß die Roboterantwort den Unterschied zwischen echter und bloß abgeleiteter Intentionalität, an dem ihm so viel gelegen ist, hinfällig macht. Auch der Roboter verfüge, *qua* Digitalcomputer, nur über eine Syntax, nicht aber über eine Semantik.

19 Vgl. ebd., S. 31.
20 Tetens, »Maschinen, die ›Geist haben‹«, S. 10.
21 Searle, *Geist, Hirn und Wissenschaft*, S. 34.

Diese Überlegungen reichen aber meines Erachtens nicht aus, den ›Fehler‹ der Roboterantwort deutlich zu machen. Vielleicht lohnt ein neuer Versuch.

4

Der philosophischen Frage, ob Maschinen denken können, ist in den letzten Jahrzehnten dasselbe widerfahren wie anderen philosophischen Fragen: Sie mußte sich den ›semantischen Aufstieg‹ gefallen lassen. Viele Autoren halten sie heute für zu vage und unqualifiziert, als daß man sie überhaupt in interessanter Weise beantworten könnte. Sie ist deshalb durch Fragen abgelöst worden wie: Können mentale Begriffe sinnvoll auf maschinale Prozesse angewandt werden? Welche Arten von Gründen kann man dafür oder dagegen anführen? Welche Bedingungen müßten erfüllt sein? Ist die Frage nach der Zuschreibbarkeit mentaler Zustände überhaupt eine Tatsachenfrage, gibt es hier ein *fact of the matter*? Oder sollten wir diese Zuschreibungen instrumentalistisch behandeln? Rechtfertigt sich die Einnahme des intentionalen Standpunkts allein durch ihren Erklärungserfolg?

Wenn wir nun einem Roboter Wahrnehmungs- und Handlungskompetenzen zuschreiben, dann sollten wir analoge Fragen stellen. Wir sollten das erste, naive Stadium so schnell wie möglich überspringen und auch hier den semantischen Aufstieg vollziehen. Was *heißt* es eigentlich, einem System Wahrnehmungs- und Handlungsfähigkeiten zuzuschreiben? Welche Bedingungen müssen erfüllt sein, damit wir solche Zuschreibungen als wörtlich wahr ansehen können?

Die Vertreter der Roboterantwort nehmen offenkundig an, daß die Zuschreibung von Wahrnehmungs- und Handlungskompetenzen weniger problematisch sei als die Zuschreibung von Denkleistungen oder von mentalen Zuständen. Schließlich soll die Zuschreibung der ersten Art solche der letzteren legitimieren. Dies ist nur möglich, wenn Gründe zur Skepsis, die im einen Fall vorliegen, im anderen wegfallen. Von dieser Unterstellung lebt die Roboterantwort, und sie hat auch einiges für sich. Wie eine

Maschine sich *verhält*, wie sie sich bewegt, auf welche Veränderungen in ihrer Umwelt sie reagiert, wie sie mit Gegenständen manipuliert, das scheint doch offen zutage zu liegen. Entweder erbringt sie eine fragliche Leistung oder sie erbringt sie nicht. Hier müssen wir nicht erst unbeobachtbare innere Zustände fingieren, um das Systemverhalten zu erklären, sondern wir lassen uns umgekehrt vom beobachtbaren Systemverhalten beeindrukken, von dessen Vielfalt und Plastizität, von dessen »Zielgerichtetheit und Umweltbezogenheit« (Tetens), und nehmen *dies* dann zum Anlaß, dem System mentale Zustände zuzuschreiben. Mit einem Wort: Wir können zunächst gute Behavioristen bleiben. Angesichts des wahrnehmbaren Verhaltens eines Roboters müssen wir nicht schon von vornherein den intentionalen Standpunkt einnehmen, den es doch erst zu legitimieren gilt.

Diese Sicht der Dinge ist auf den ersten Blick bestechend und auf den zweiten naiv. Ich möchte die ihr zugrunde liegende Annahme im folgenden angreifen: daß es weniger problematisch sei, einen Roboter wahrnehmend und handelnd zu nennen als eine Maschine denkend, intelligent oder geistbegabt.

Was *heißt* es also, einem System Wahrnehmungs- und Handlungsfähigkeit zuzuschreiben? – Den Begriff der Wahrnehmungsfähigkeit, um mit diesem zu beginnen, führt Rosemarie Rheinwald in ihrer Version der Roboterantwort folgendermaßen ein:

»›Wahrnehmungsfähigkeit‹ soll hier nicht mit hohem erkenntnistheoretischem Anspruch gebraucht werden. Unter ›Wahrnehmungsfähigkeit‹ verstehe ich nur die Fähigkeit, bestimmte Gegenstände und Situationen mittels kausaler Mechanismen voneinander zu unterscheiden.«[22]

Das »Unterscheiden mittels kausaler Mechanismen«[23] erläutert Rheinwald am Beispiel des Thermometers. Dieses besitze »die Fähigkeit, (innerhalb der Meßgenauigkeit) *beliebige* Temperaturen zwischen 0 und 20 °C zu unterscheiden«.[24] – Offenkundig ist Rheinwald bestrebt, den Begriff des Wahrnehmens möglichst anspruchslos einzuführen. Das Gesagte bezeichnet bislang nur den Umstand, daß das physikalische Verhalten des Thermometers unterschiedlich ausfällt in Abhängigkeit von der Temperatur,

22 Rheinwald, »Können Maschinen eine Sprache sprechen?«, S. 42.
23 Ebd., S. 44.
24 Ebd.

die in seiner Umgebung herrscht. Ein solches Systemverhalten legitimiert sicherlich noch nicht die Zuschreibung mentaler Zustände oder die Einnahme des intentionalen Standpunkts in heuristischer Absicht. Unabhängig davon sollten wir uns fragen, ob diese Erläuterung von ›Wahrnehmungsfähigkeit‹ nicht *zu* anspruchslos ist. Einigkeit sollte darüber bestehen, daß ›wahrnehmen‹ weder gleichbedeutend noch koextensiv ist mit ›kausal affiziert werden‹. In jedem Einzelfall von Wahrnehmung dürfte zwar ein kausales Affiziertwerden vorliegen, aber das Umgekehrte gilt nicht. Wir bezeichnen nicht jedes kausale Affiziertwerden als Wahrnehmen. So würden wir den Wellensaum, den das ablaufende Wasser auf dem Ufersand zurückläßt, nicht als Indiz dafür ansehen, daß der Strand die Welle ›wahrgenommen‹ hat. Gleichwohl fällt natürlich das Systemverhalten des Strandes unterschiedlich aus in Abhängigkeit von den Wellen, mit denen er physischen Kontakt hatte. Rheinwald schuldet uns eine Erklärung dafür, warum wir manche kausale Affektionen, die einen Unterschied im Systemverhalten machen, Wahrnehmungen nennen und andere nicht. (Man erwidere nicht, *ich* schuldete eine Erklärung dafür, die Fälle überhaupt zu unterscheiden. Mag auch die Berufung auf den normalen Sprachgebrauch philosophisch unpopulär geworden sein, so sollte doch, wer von diesem Sprachgebrauch eklatant abweicht, zumindest die Beweislast tragen.)

Zu klären wäre, ob sich ein Begriff der Wahrnehmung erläutern läßt, der nicht jedem beliebigen mit Umweltveränderungen kovariierenden System Wahrnehmungen zuzuschreiben nötigt, dies aber bei Robotern erlaubt. Ich werde diese Frage auf sich beruhen lassen, weil ich meine Kritik an der Roboterantwort am Begriff des Handelns durchspielen möchte und nicht am Begriff des Wahrnehmens.[25] Viel verloren ist damit nicht, weil der Fall des

25 Ein möglicher Vorschlag lautet, daß wir überall dort, wo eine strikte nomologische Korrelation zwischen Systemverhalten und Zuständen der Außenwelt besteht, *nicht* von Wahrnehmen sprechen. Wo wir das Wahrnehmungsvokabular in Anschlag bringen, müsse Fehlrepräsentation – im Sinne eines fehlerhaften Urteils über das Wahrgenommene – möglich sein. Wenn ich behaupte, ein Kaninchen wahrgenommen zu haben, muß es auch der Fall sein können, daß ich etwas wahrgenommen habe, das nur so *aussah* wie ein Kaninchen. Andernfalls, also beim Vorliegen einer strikten Korrelation, bestünde über-

Wahrnehmens und der des Handelns aus meiner Sicht für den Roboterfreund gleichartige Probleme aufwerfen.

Eingangs war von der Feld-, Wald- und Wiesendefinition von ›Künstlicher Intelligenz‹ die Rede, in der die Leistungen einer Maschine anhand von deren Fähigkeit beurteilt werden, intelligenzerfordernde menschliche Leistungen zu ersetzen. Eine etwas genauere Formulierung dieses Ansatzes findet sich bei Tetens:

»Es ist das Ziel der Künstlichen Intelligenz, Maschinen zu bauen, die Verhaltensleistungen erbringen, die wir beim Menschen (alltags-)psychologisch durch Rekurs auf mentale Zustände erklären und vorhersagen würden.«[26]

Hier fällt zunächst die Richtung der Analogie auf. Es wird heute oft vergessen, daß die klassische KI nicht mit dem Computermodell des Geistes beginnt, sondern mit dem Geistmodell des Computers. Die Maschine wird beschrieben, als besäße sie etwas, von dessen Vorhandensein wir beim Menschen überzeugt sind: mentale Zustände. Das Computermodell des Geistes, das den menschlichen Geist seinerseits als einen Computer auffaßt, kommt erst später durch eine Rückanwendung ins Spiel[27], wenn

haupt kein Anlaß, das Wahrnehmungs- oder das Repräsentationsvokabular zu verwenden. Wir würden einfach, wie wir es bei Thermometern tun, am Zustand des Systems dessen kausale Quelle in der Außenwelt ablesen. – Umstritten ist, wie sich dieser Vorschlag auf Wesen (zum Beispiel Tiere) anwenden läßt, die aufgrund des Fehlens symbolischer Repräsentationen nicht in dem hier angenommenen Sinne Fehlurteile abgeben können.

26 Tetens, »Maschinen, die ›Geist haben‹«, S. 3; vgl. Tetens, *Geist, Gehirn, Maschine*, S. 106-121, 148.

27 Searles Unterscheidung zwischen der ›starken‹ und der ›schwachen‹ KI-These bezieht sich auf Auffassungen, in denen diese Umkehrung der Übertragungsrichtung schon stattgefunden hat. Unter ›starker KI‹ versteht Searle die Auffassung, einen Geist zu haben sei nichts anderes als über ein Programm zu verfügen. Die ›schwache KI‹ behaupte nur, daß geistige Prozesse, wie alles andere auch, auf einem Computer simuliert werden können. – Da mir nicht klar ist, was genau die Simulationsthese besagt, ziehe ich es vor, die Unterscheidung in der Frage auszudrücken, ob mentale Prädikate wörtlich auf Zustände eines Computers zutreffen können oder nicht (bzw. aus der umgekehrten Perspektive: ob mentale Prozesse wörtlich Rechenprozesse sind oder nicht). Ein prototypischer Vertreter der ›starken KI‹ im Sinne der Wörtlichkeitsthese ist Pylyshyn, *Computation and Cognition*.

nämlich die Frage auftritt, ob unser Selbstverständnis als denkende und erkennende Wesen noch mit (durch den technischen Erfolg der KI nahegelegten) Auffassungen darüber vereinbar ist, wie Kognitionsprozesse *auch* instantiiert sein können. Dazu wiederum Tetens:

»Der mögliche Erfolg der Künstlichen Intelligenz könnte uns [...] mit der Frage konfrontieren: Können wir angesichts von Maschinen, die sich nachweisbar aufgrund rein physikalischer Mechanismen genauso verhalten wie wir, an den psychologischen und alltagspsychologischen Beschreibungen und Theorien über uns selbst festhalten, oder müssen wir diese Theorien revidieren, ja überhaupt aufgeben?«[28]

Hier kommt es mir auf die Formulierung an, daß die Maschinen sich »genauso verhalten wie wir«. Woran bemißt sich, ob eine Maschine sich ›genauso verhält‹ wie wir? Diese Nachfrage scheint sehr begriffsstutzig zu sein. Sie wird auch in Tetens' elaborierter Fassung der Feld-, Wald- und Wiesendefinition nicht zum Problem. Die KI läßt die Maschine etwas tun, das, würde es vom Menschen getan usw. – *Was* wird denn sowohl von der Maschine als auch vom Menschen getan? Im rückverweisenden Pronomen ›es‹ steckt eine Identitätsunterstellung, für die es einer Begründung bedarf. Was ein Mensch tut, wenn er seinen Autoschlüssel sucht, eine Suppe abschmeckt oder jemanden herbeiwinkt, davon haben wir eine Vorstellung, denn eben auf menschliche Aktivitäten sind solche Beschreibungen zugeschnitten. Aber woran bemißt sich, ob eine Maschine sich ›genauso‹ verhält?

28 Tetens, »Maschinen, die ›Geist haben‹«, S. 12. – Es wäre zu überlegen, ob man den Bezug auf den ›möglichen Erfolg der KI‹ hier nicht ersatzlos streichen kann. Tetens' Überlegung benötigt die KI gar nicht als Illustration, denn sie betrifft schon das Geist-Körper-Problem. Schließlich dürfte doch auch unser eigenes Verhalten auf ›rein physikalischen Mechanismen‹ beruhen, ohne daß wir deshalb darauf verzichten würden, es psychologisch zu beschreiben. Davidson hat in diesem Sinne dafür argumentiert, daß durch einen perfekten Roboter »nichts bewiesen [wird], was nicht ebenso gut durch die Annahme geleistet würde, wir hätten eine ebenso umfassende Kenntnis des physikalischen Aufbaus eines Menschen, wie wir sie im Falle [des Roboters] fingiert haben« (Davidson, »Der materielle Geist«, S. 351). Und Searle sagt gelegentlich, daß es in einem trivialen Sinne denkende Maschinen gebe, nämlich uns (vgl. Searle, *Geist, Hirn und Wissenschaft*, S. 34).

Fairerweise sollte man zunächst Verhaltensleistungen betrachten, die nicht ohnehin außerhalb der Reichweite existierender Maschinen liegen. Im Falle von Leistungen wie ›einen kleinen roten Klotz auf einen großen blauen Klotz legen‹ kann man sich der Suggestion des ›genauso‹ kaum entziehen. Es drängt sich auf, die wahrnehmbaren Operationen eines Roboters mit denselben Worten zu beschreiben wie die Handlung eines Menschen, die dasselbe Resultat hat. Zumindest aus einer bestimmten Perspektive ist dies die nächstliegende Beschreibung dessen, was geschieht, nämlich aus der Perspektive desjenigen, der den Roboter zur Erreichung eines bestimmten Resultats konstruiert hat oder ihn dazu benutzt. Freilich ist es nicht der Roboter selbst, der die Bewegungen seiner ›Effektoren‹ als die fragliche ›Leistung‹ spezifiziert, denn der dazu erforderliche Bezug auf ein Handlungsziel wird ihm von einem Beobachter zugeschrieben. Offen ist deshalb die Frage, ob durch die Erreichbarkeit solcher Resultate mit Hilfe eines Roboters dessen *Handlungsfähigkeit* bewiesen wird, und zwar in einer solchen Weise, die schließlich die Zuschreibung mentaler Zustände legitimiert. Reicht es für Handlungsfähigkeit schon aus, wenn der Greifarm eines Roboters in SHRUDLU-Manier mit Gegenständen in seiner Umgebung hantiert? Kann man die behauptete Äquivalenz der Leistungen überhaupt allein am *Resultat* der Operation messen, oder kommt es auch auf die Art des *Vollzugs* an? Müssen auch Teilhandlungen äquivalent sein? Müßte es vielleicht verschiedene Wege geben können, das gleiche Ziel zu erreichen? Und mit Wittgenstein: Was ist es, was davon, daß der Roboter seinen Greifarm hebt, übrigbleibt, wenn wir die Tatsache abziehen, daß sein Greifarm sich hebt?

Man sollte sich, was den Handlungsbegriff betrifft, nicht unwissender stellen als nötig. Die philosophische Handlungstheorie beschäftigt sich seit Jahrzehnten ausführlich mit der Frage, welche Bedingungen erfüllt sein müssen, damit wir eine Bewegung unseres Körpers eine Handlung nennen. Der Mehrheitsmeinung zufolge, die auch dem *common sense* entspricht, muß eine Körperbewegung *absichtlich* vollzogen werden, um eine Handlung zu sein. Zwar ist notorisch umstritten, worin diese Absichtlichkeit besteht und ob sie noch auf elementare Bedingungen zurückgeführt werden kann. Einigkeit kann man aber über Beispiele erzielen: Der Kniescheibenreflex, das Wachsen der

Fingernägel oder das Nasenbluten sind keine Handlungen, denn sie sind Bewegungen oder Veränderungen unseres Körpers, die wir nicht absichtlich vollziehen. Interessanterweise wird diese Bedingung auch von Protagonisten der *kausalen* Handlungstheorie akzeptiert, namentlich von Goldman und von Davidson. Wenn Davidson sagt: »Wir sollten herauszufinden versuchen, ob wir ein Kennzeichen des Handelns entdecken, das sich nicht auf den Begriff der Absicht stützt«[29], und in der Folge auf die kausale Rolle der Wünsche und Überzeugungen des Akteurs zurückgreift, so entspringt dieses Verfahren nicht einer Ablehnung des Absichtlichkeitskriteriums. Die kausale Handlungstheorie zeichnet sich vielmehr dadurch aus, daß sie den Begriff der Absichtlichkeit nicht als unanalysierbaren Grundbegriff akzeptiert. Die Wege der Kausalisten und der Intentionalisten trennen sich also erst jenseits des Kriteriums der Absichtlichkeit.

Ein Handelnder steht nicht bloß in einer kausalen Beziehung zum hervorgebrachten Effekt, sondern zusätzlich in einer intentionalen. Würden wir diese Bedingung aufgeben, könnten wir überhaupt nicht unterscheiden zwischen dem, was wir tun, und dem, was uns widerfährt. Möglichst theoriearm kann man diesen Unterschied, parallel zu dem über Fehlwahrnehmung Gesagten, auch so ausdrücken: Handlungen sind etwas, was scheitern kann. Unbeabsichtigte Veränderungen in, an und mit unserem Körper können das nicht.[30]

Nun können die Roboterfreunde leicht zugeben, daß der philosophischen Handlungstheorie an dieser Unterscheidung gelegen sein mag. Was aber kümmert es die KI? Ist es nicht gerade die Pointe der Roboterantwort, die Bedingung der Intentionalität erst einmal aus dem Spiel zu lassen, von der ja überdies unklar sei, worin sie überhaupt besteht? Schließlich sollen doch die eindrucksvollen Leistungen des Roboters die Intentionalitätszuschreibung (hier: den Spezialfall der Absichtlichkeit) erst legitimieren. Wäre es nicht geradezu eine *petitio principii*, vom Roboter mentale Zustände wie Absichten zu fordern, um ihm Handlungsfähigkeit zuzuerkennen?

Hier möchte ich zunächst festhalten, daß auch die Vertreter der

29 Davidson, »Handeln«, S. 79.

30 Vgl. auch Janichs dreiteilige Bedingung: »Zu Handlungen kann man auffordern, man kann sie vollziehen oder unterlassen, und sie können ge- oder mißlingen.« Janich, *Grenzen der Naturwissenschaft*, S. 15.

Roboterantwort auf eine Unterscheidung zwischen den Kausalketten angewiesen sind, die den Roboter bloß durchlaufen, und denjenigen, an denen er aktiv beteiligt ist. Ohne diesen Unterschied könnte man nicht angeben, was einen Roboter von einer Maschine unterscheiden soll, die keine ›Sensoren‹ und ›Effektoren‹ besitzt. *Kausale Interaktionen* zwischen System und Umwelt gibt es bei allen Maschinen, selbst beim Taschenrechner. Jeder konventionelle Rechner mit von-Neumann-Architektur steht über seine Eingabe- und Ausgabeeinheit in kausalem Kontakt mit seiner Umwelt. Ohne kausalen Input würde er nicht zu arbeiten beginnen, und sein Outputverhalten hat Veränderungen in der Körperwelt zur Folge, und sei es auf dem Wege der kinetisch wenig energiereichen Strahlenemission des Bildschirms. Die Bedingung der kausalen Einbettung kann also nicht ausreichen; es muß noch etwas dazukommen, damit eine Maschine ein Roboter ist. Grob gesprochen, müßte der Roboter aktiv etwas *tun*, statt bloß in *irgendeinem* kausalen Kontakt mit seiner Umwelt zu stehen.

Einige Philosophen haben zu diesem Zweck zwischen *innen-induzierten* und *außeninduzierten* Bewegungen unterschieden. Dretske hat seine Unterscheidung zwischen »internal« und »external causes« sogar auf Pflanzen angewandt: »Plants behave for the same reason animals behave: some of the changes occurring to them are brought about from within«.[31] – Ich kann die Pointe von Dretskes Unterscheidung nicht nachvollziehen. Naturgegenstände werden von zahllosen Kausalketten durchlaufen, und es ist diesen Kausalketten völlig gleichgültig, oder weniger als das, welche Stationen auf ihrem Wege liegen. Welchen Grund kann Dretske dafür anführen, eine Kausalkette innerhalb einer Pflanze oder einer Maschine beginnen zu lassen? Kausalketten *beginnen* überhaupt nicht, es sei denn, man greift auf ein teleologisches ›Prinzip der Bewegung in sich selbst‹ zurück. Kausaltheoretisch gibt es keinen Unterschied zwischen ›innen-‹ und ›außeninduzierten‹ Bewegungen.

Tetens spricht, wie zitiert, von ›zielgerichtet-umweltbezogenem Verhalten‹, und er nennt die entsprechenden Komponenten des Roboters ›Effektoren‹. Die erste, teleologische Formulierung präsupponiert Absichtlichkeit, solange keine unabhängige nichtintentionale Rekonstruktion beigebracht wird. Aber auch die Titulierung einer Hardwarekomponente als ›Effektor‹ ist heikel, denn sie unterstellt ein Moment von *Aktivität*. Auch dieses wäre erst einmal ohne Rückgriff auf intentionale Bestimmungen auszuweisen. Der Verweis auf das asymmetrische Verhältnis von ›Ursache‹ und ›Wirkung‹ würde hier nicht genügen, denn die Asymmetrie der Kausalrelation ist unter nachhumeschen Bedingungen, soweit sie über die bloße zeitliche Sukzession hinausgeht, ihrerseits problematisch geworden. Der Unterschied zwischen einem ›aktiven‹ und einem ›passiven‹ Pol einer kausalen Transaktion läßt sich im Rahmen einer Regularitätsauffassung der Kausalität nicht erläutern. Der Rede vom ›Effektor‹, soll sie nicht eine bloße *façon de parler* sein, liegt die Vorstellung eines bewirkenden Agens zugrunde, die Russell und andere zum Überbleibsel einer animistischen Metaphysik erklärt haben.[32]

Die Vertreter der Roboterantwort können sich nicht auf das bloße kausale Eingebettetsein einer Maschine in die ›wirkliche Welt‹ berufen, und seien die kausalen Kontakte noch so vielfältig. Um einem System darüber hinaus Handlungsfähigkeiten zuzuerkennen, muß man erläutern können, was es heißt, daß das System *aktiv* in seine Umwelt eingreift, daß es *von sich aus etwas tut*. Dieses zur Kausalverbindung Hinzukommende ist nach herkömmlicher Auffassung eine *intentionale* Verbindung, nämlich die der Absichtlichkeit. Etwas absichtlich zu tun impliziert aber, mentale Zustände haben zu können. Deshalb erscheint das Insistieren auf Absichtlichkeit aus der Perspektive der Roboterfreunde eine *petitio*, denn sie hätten es gerne andersherum: mentale Zustände, *weil* handlungsfähig. Eine Maschine sollte genau dann über mentale Repräsentationen verfügen, Informationen verarbeiten, eine Sprache verstehen etc., wenn sie wahrnehmen und handeln kann. Nun hatte ich eingangs angedeutet, daß ich *diese* Auffassung für richtig halte. Handlungsfähigkeit ist tatsächlich eine Bedingung für das Haben von mentalen Zuständen. Wer mentale Zustände haben können soll, der muß auch handeln können.[33] (Körperlose Engel hätten demnach keine menta-

31 Dretske, *Explaining Behavior*, S. 9. Eine analoge Unterscheidung wendet Hauser auf Roboter an; vgl. Hauser, »Acting, Intending and Artificial Intelligence«, S. 24.

32 Vgl. dazu Keil, »Zu Russells These vom Absterben des Kausalbegriffs in den Wissenschaften«.

33 Handlungsfähigkeit ist meines Erachtens sogar eine Bedingung, ohne die wir nicht einmal einen *Begriff* von der vielbeschworenen ›Gerich-

len Zustände.) Die Bedingung scheint mir nicht nur notwendig, sondern auch hinreichend zu sein: Wir sollten keinem Wesen, dem wir Handlungsfähigkeit zuerkennen, mentale Zustände absprechen. Die komplementäre Auffassung ist aber auch richtig: Nur wer Überzeugungen, Wünsche und Absichten hat, kann handeln. Das heißt aber: *Handelnkönnen und Geisthaben setzen einander wechselseitig voraus*, und dies nicht temporal oder kausal, sondern *begrifflich*. Einem System Handlungsfähigkeit zuzuerkennen impliziert *begrifflich*, ihm mentale Zustände zuzuerkennen – und vice versa.[34]

Und deshalb ist die Roboterantwort so suggestiv: Sie kommt der Wahrheit so nahe. Handelnkönnen und Geisthaben sind zwei Seiten einer Medaille. Nur dürfen wir uns eben deshalb nicht mit einem zu anspruchslosen Verständnis des Handelns zufriedengeben. Ebenso wie die elektronischen Vorgänge innerhalb des Computers nicht von sich aus Repräsentationsgehalte haben, muß auch das, was eine Maschine ›tut‹, erst einmal vom intentionalen Standpunkt aus interpretiert werden, damit wir es Handeln nennen können. *Wenn* also die Legitimität der Intentionalitätszuschreibung noch nicht gesichert sein sollte, dann können Roboter aus denselben Gründen nicht handeln, aus dem Rechner keine intentionalen Zustände haben. Handelnkönnen und Geisthaben sind nicht hinreichend unabhängig voneinander, als daß man das eine als *Rechtfertigung* für die Zuerkennung des anderen verwenden könnte.

Die Vertreter der Roboterantwort unterstellen offenbar, daß das wahrnehmbare ›Verhalten‹ eines Roboters weniger interpretationsbedürftig ist als die Zustände eines Rechners. Damit wiederholen sie einen Irrtum, der schon im Behaviorismus endemisch war. Behavioristen haben stets reklamiert, bloß am beobachtbaren Verhalten interessiert zu sein und für den Begriff des absichtlichen Handelns keine Verwendung zu haben. Der Behaviorismus mußte sich aber darauf hinweisen lassen, daß sich schon elementare Verhaltenstypen überhaupt nicht als solche *identifizieren* lassen, ohne auf intentionale Charakterisierungen zurückzugreifen.[35] Schließlich arbeitet der Behaviorismus nicht mit kinematischen Beschreibungen von Bewegungen unserer Körperglieder. Auf diese Weise könnte er auch seine explanatorischen Ziele nicht erreichen, weil von dort kein Weg zur Klassifizierung einer Körperbewegung als Fall von ›eine Tür öffnen‹ oder ›einem Schlag ausweichen‹ führt. Eine Tür kann man mit Hilfe verschiedener möglicher Körperbewegungen öffnen. Umgekehrt lassen sich mit einer Körperbewegung ganz verschiedenartige Handlungen vollziehen. Carnap glaubte noch, die Klasse der Armbewegungen, mit denen man die Geste des Herbeiwinkens vollziehen kann, allein aufgrund ihrer »physikalischen Beschaffenheit« [36] auszeichnen zu können – wenn er diese Aufgabe auch für »gegenwärtig noch nicht gelöst«[37] hielt. Diese Hoffnung war verwegen. Die Handlungstheorie strotzt vor Beispielen dafür, daß die mit einer bestimmten Körperbewegung vollzogene Handlung vom Kontext und von der Intention des Handelnden abhängt.[38] – Über sein parasitäres Verhältnis zum intentionalen Begriffsrahmen hat sich der Behaviorismus stets hinweggetäuscht. Er litt an einem *anti-mentalistischen Selbstmißverständnis*, an dem er, im Zuge der kognitiven Wende in den Humanwissenschaften, schließlich auch zugrunde gegangen ist.

Searle hat diese Selbsttäuschung des Behaviorismus so beschrieben:

»In den behaviouristischen Standardanalysen geistiger Zustände wird der Begriff des intentionalen Verhaltens einfach so gebraucht, als sei er irgendwie weniger mentalistisch als die anderen Begriffe des Geistigen; aber wenn man von jemandem sagt, er gehe zu dem Laden oder nehme eine Mahlzeit zu sich, so schreibt man ihm nicht minder geistige Zustände zu, als wenn man von ihm sagt, er wolle in den Laden gelangen oder glaube, daß das Zeug auf dem Teller Nahrung ist. Wir sitzen der Illusion

tetheit‹ des Mentalen hätten. Allein über den Handlungsbegriff kann das Moment von Aktivität eingeführt werden, von dem die Metapher des Sich-auf-etwas-Richtens lebt; vgl. Keil, *Kritik des Naturalismus*, S. 357 f., 383.

34 »The claim that men have minds is, in other words, nothing but the claim that they are capable of doing all sorts of things [...] which machines and other inanimate things cannot do. [...] [I]t is no real *explanation* of how men are able to do such things, to say that they have minds. It is only a strange way of saying the same thing again.« Taylor, *Action and Purpose*, S. 247 f.

35 Vgl. Taylor, *The Explanation of Behavior*. Ähnliche Kritik am Behaviorismus haben Dennett und Stoutland geübt.

36 Carnap, »Psychologie in physikalischer Sprache«, S. 126.

37 Ebd.

38 Vgl. Danto, *Analytical Philosophy of Action*, S. ix.

auf, Verhalten sei nichts Geistiges, weil wir Körperbewegungen beobachten können; aber die Körperbewegungen stellen menschliches Handeln nur unter der Annahme dar, daß die geeigneten Absichten und Überzeugungen vorliegen.«[39]

Die Annahme der Roboterfreunde, daß das, was jemand *tut*, wenn er sich in seiner Umwelt bewegt, eo ipso weniger interpretationsbedürftig sei als seine mentalen Zustände, ist also unzutreffend. Im Rahmen einer Diskussion, in der gerade die Legitimität der Einnahme des intentionalen Standpunkts auf dem Spiel steht, kommt Tetens' Rede von ›Maschinen, die sich genauso verhalten wie wir‹, ihrerseits einer *petitio principii* gleich. Woran sich das ›genauso‹ bemißt und ob diese Parallelisierung gerechtfertigt ist, ist in Abwesenheit einer unabhängigen nichtmentalistischen Rekonstruktion der intentionalen Anteile von Verhaltensbeschreibungen alles andere als klar. *Wenn* das bislang von der KI Geleistete noch nicht ausreichen sollte, Maschinen vorbehaltlos mentale Zustände zuzuschreiben, dann ist die Zuschreibung von Handlungskompetenzen im Rahmen der Roboterantwort nicht minder problematisch. Der Umstand, daß Vertreter der Roboterantwort Maschinen Wahrnehmungs- und Handlungskompetenzen *zuschreiben*, bietet nicht mehr und nicht weniger Gewähr dafür, daß der Roboter diese Kompetenzen auch wirklich hat, als im Falle der Zuschreibung mentaler Zustände.

Den Roboterfreunden ist darin zuzustimmen, daß sie mentale Zustände und Handlungskompetenz so eng aneinander binden. Wir sollten tatsächlich keinem Wesen, dem wir Handlungsfähigkeit zuerkennen, mentale Zustände absprechen. Vielmehr sollten wir Robotern beides absprechen.

39 Searle, »Intentionalität und der Gebrauch der Sprache«, S. 169. Und weiter: »Somit wird entweder im Analysans der behaviouristischen Analyse Intentionalität vorausgesetzt – dann handelt es sich bei ihr um einen weiteren Fall des intentionalen Zirkels – oder dem ist nicht so – dann ist die Analyse inadäquat, denn dann geht es in der Analyse überhaupt nicht um Verhalten im Sinne von menschlichem Handeln.« Ebd.

Die vorgetragenen Überlegungen waren vornehmlich destruktiv. Ich habe nichts darüber gesagt, welche von einem Artefakt erfüllbaren Bedingungen denn überhaupt als hinreichend für Handlungsfähigkeit angesehen werden können. Was das Verhältnis von Handlungskompetenz und mentalen Zuständen betrifft, so habe ich angenommen, daß beide nur ko-instantiiert auftreten. Darüber hinaus habe ich eine wechselseitige *begriffliche* Abhängigkeit behauptet, der zufolge der Handlungsbegriff sich nicht ohne Verwendung des intentionalen Idioms analysieren läßt und vice versa (letzteres habe ich allerdings nicht begründet). Eine solche Strategie zieht sich natürlich den Vorwurf eines Explikationszirkels zu. Diesem Vorwurf möchte ich mit der Behauptung begegnen, daß die Forderung nach logisch voneinander unabhängigen Begriffsanalysen und -explikationen im Bereich der Grundbegriffe unserer deskriptiven Metaphysik oft nicht erfüllbar ist.[40] Dies gilt besonders für *intentionale* Grundbegriffe, deren Verweisungsstruktur in der Philosophie des Geistes als ›Zirkel der Intentionalität‹ beschrieben wird. Wenn solche Explikationszirkel unvermeidlich sind, kann ein Vorwurf nur demjenigen gemacht werden, der einige der begrifflichen Implikationen übersieht und die restlichen Abhängigkeiten als Fundierungsbeziehungen darstellt oder als solche der genetischen Voraussetzung.

Kehren wir zu der offenen Frage zurück, welcher Grad von Ähnlichkeit zwischen einem Artefakt und einer Person ausreichen würde, um ersterem sowohl mentale Zustände als auch Handlungskompetenz zuzuschreiben. Artenchauvinisten haben darauf eine *Apriori*-Antwort, und sie lautet: ›keiner‹. Sie müssen dann die These vertreten, daß eine Maschine *qua Artefakt* immer bloß zugeschriebene oder abgeleitete Intentionalität aufweisen kann, wobei die zuschreibenden Instanzen, nämlich die Konstrukteure und Benutzer der Maschine, die primären Intentionalitätsträger bleiben. Für diese Antwortstrategie braucht man natürlich sehr gute Argumente, also bessere, als etwa Searle sie hat. Wer jede neue maschinale Leistung, mit der die Roboterkonstrukteure aufwarten, aus dem Lehnstuhl mit der gelassenen Be-

40 Man denke zum Beispiel an Aristoteles' Einführung des Begriffspaars ›Zeit‹ und ›Bewegung‹.

merkung kommentiert, daß diese Leistung ›not the real thing‹ sei, setzt sich dem Verdacht eines begrifflichen Konservativismus aus, für den oft Wittgensteins Bemerkung als abschreckendes Beispiel angeführt wird: »Aber eine Maschine kann doch nicht denken! – Ist das ein Erfahrungssatz? Nein. Wir sagen nur vom Menschen, und was ihm ähnlich ist, es denke.«[41] Dem halten die Anwälte der KI entgegen, daß sich die Anwendungsbedingungen unserer Begriffe angesichts der Erfindung von Apparaten, von denen vergangene Sprechergemeinschaften nicht einmal träumten, schließlich *ändern* könnten. Auch Searles Zurückweisung der Roboterantwort erscheint vielen seiner Kritiker deshalb so wenig befriedigend, weil sie eingestandenermaßen a priori erfolgt. Searle betont, seine Argumentation (Syntax ja, Semantik nein) habe »nichts damit zu tun, welche erstaunlichen Fortschritte der Computerwissenschaft bevorstehen«.[42]

Es scheint also, als ob die defensive ›Not-the-real-thing‹-Attitüde gegenüber der KI an irgendeiner Stelle in den Dogmatismus umschlüge. Aber an welcher? Wenn man annimmt, daß es *kein* Apriori-Argument für den Artenchauvinismus gibt, erscheint der Vorschlag plausibel, Geistbegabtheit als eine Sache des Grades aufzufassen. So sieht es – ungeachtet seines grundsätzlichen Instrumentalismus – Dennett, so sieht es auch Davidson: »How much like a person an object must be to be intelligible – to have thoughts – is unclear; indeed, it makes the most sense to think of

41 Wittgenstein, *Philosophische Untersuchungen*, § 360.

42 Searle, *Geist, Hirn und Wissenschaft*, S. 35. Für Searle ist die echte, nicht bloß zugeschriebene Intentionalität in einer nicht näher erläuterbaren Weise untrennbar mit der biologischen ›wetware‹ des Homo sapiens verbunden. Diese biologisch begründete Version des Artenchauvinismus kritisiert Dennett treffend als die ›Wonder-tissue‹-Auffassung. Allerdings ist die These, daß wir einem Roboter qua Artefakt bloß abgeleitete Intentionalität zuschreiben können, nicht auf das Argument des falschen *Materials* verpflichtet. Es könnte sich auch um die falsche *Genese* handeln. – Übrigens ist Searles weitergehende Behauptung, Intentionalität sei eine *biologische Eigenschaft* des Gehirns, ein *non sequitur*. Nicht jede Eigenschaft, die nur in Wesen mit einer bestimmten biologischen Ausstattung realisiert ist, wird dadurch zu einer biologischen Eigenschaft. Andernfalls wäre es eine biologische Eigenschaft, einen Roman schreiben zu können. (Es sind auch nicht Gehirne die Träger intentionaler Zustände, sondern Personen.)

thoughtfulness as a matter of degree, as it surely is with a developing child.«[43] Soviel leuchtet ein: Neugeborene haben wahrscheinlich keine Gedanken und Überzeugungen, Kinder haben sie irgendwann, allerdings nicht von heute auf morgen. Es muß also Zwischenstadien geben, und es wäre zu begrüßen, wenn sich diese Abstufungen in unserer Verwendung des mentalistischen Idioms spiegeln würden. Das Zugeständnis der Gradualität von Geistbegabtheit würde ein neues Licht auf meine oben verwendeten holistischen Argumente werfen, und zwar ein ungünstiges. Daß wir ein Wesen erst dann als geistbegabt ansehen, wenn wir ihm eine große Menge von Überzeugungen zuerkennen können, die zudem »großenteils widerspruchsfrei und nach unseren eigenen Maßstäben wahr«[44] sein müssen, sollte ja nicht dazu führen, daß die Standards selbst von zahllosen unserer Artgenossen nicht mehr erfüllt werden. Wenn beispielsweise Kinder eines gewissen Alters geistbegabte Wesen sind, dann ist eine etwas kleinere Menge wahrer und kohärenter Überzeugungen als die eines erwachsenen Europäers mit mittlerer Schulbildung eben auch genug. Aber *wieviel* weniger? Dies ist die Crux aller holistischen Argumente: Wörtlich genommen sind sie zu stark, andererseits ist unklar, wieviel man von ihnen abstreichen darf.

Die KI kann diese Analogie zur Ontogenese aber nicht für sich verbuchen. Die ontogenetische Evolution des Geisthabens verstehen wir im Lichte ihres schon bekannten Ergebnisses. Könnten wir das nicht tun, wäre die Gradualisierung viel weniger überzeugend. Genau hier bricht die Analogie aber zusammen, denn für technische Fortschritte in der Konstruktion von Artefakten gilt dies nicht: Wir können sie nicht im Lichte eines bekannten Ergebnisses verstehen, welches den paradigmatischen Anwendungsfall unseres mentalistischen Idioms darstellt. Der ontogenetisch begründete Gradualismus liefert nicht schon ein Argument dafür, daß wir die eingangs beschriebene *kategoriale* Auszeichnung der Kandidaten für mentale Zuschreibungen überspringen dürfen.

Nun suchen die Roboterfreunde den mit der Zuschreibung mentaler Prädikate verbundenen holistischen *constraints* dadurch zu entgehen, daß sie zunächst nur eine Äquivalenz von

43 Davidson, »Turing's Test«, S. 8; Vgl. Rheinwald, »Können Maschinen eine Sprache sprechen?«, S. 44 f.

44 Davidson, »Radikale Interpretation«, S. 199.

›Verhaltensleistungen‹ behaupten. Dagegen hatte ich ins Feld geführt, daß die Rede von ›Maschinen, die sich genauso verhalten wie wir‹ nicht schon durch den Verweis auf erreichte *Resultate* gerechtfertigt wird. An keinem Endzustand eines Prozesses läßt sich ablesen, ob er auf diejenige Art zustande gekommen ist, die die Zuschreibung bestimmter Verhaltens- und Handlungsprädikate erlaubt. Selbst wenn wir von der Bedingung der Absichtlichkeit einen Moment absehen, wird die Abgrenzung von Handlungsresultaten gegenüber Handlungsvollzügen spätestens angesichts von Handlungen brüchig, die an unsere artspezifische körperliche Organisation gebunden sind: seine Stirn in Falten legen, sich hinsetzen, Schmetterlingsschwimmen.[45] Überdies können wir beliebige Handlungen derart in Teilhandlungen zerlegen, daß diese nur von Wesen ausgeführt werden können, deren körperliche Organisation der unseren mehr gleicht als die aller bekannten Roboter.

Diese Hinweise mögen dem Roboterfreund als irrelevant erscheinen. Für die Zuerkennung von Handlungsfähigkeit müßte es doch ausreichen, *irgendwelche* Handlungen vollziehen zu können. Diese Entgegnung übersieht die Pointe des holistischen Arguments. Mit Davidson hatten wir argumentiert, daß ein Wesen, das nur eine einzige Überzeugung hätte, nicht einmal diese hätte. Angesichts des intentionalitätspräsupponierenden Charakters von Handlungsbeschreibungen können wir nun sagen, daß Entsprechendes für Handlungen gilt. Ein System, das nur eine einzige oder sehr wenige Bewegungen ausführen könnte, beweist damit nicht diejenige Fähigkeit, die wir einem Wesen zuerkennen, bei dem behavioral davon ununterscheidbare Bewegungen in ein größeres Repertoire eingebettet sind. Das cartesische Argument aus der Universalität der menschlichen Vernunft läßt sich auch auf den Roboter anwenden: Selbst wenn dieser etwas »ebensogut oder vielleicht besser als einer von uns«[46] machte, zeigte sich daran, daß er »unausbleiblich in einigen anderen fehlen«[47] würde, daß er das wenige, was er kann, »nicht nach Einsicht«[48] tut. Die Interpretierbarkeit einer Bewegung als Handlung hängt also unter anderem vom *Repertoire* ab, aus dem die Bewegung aktualisiert wird. Ein Repertoire, das keine oder zu wenige Überschneidungen mit dem unsrigen hätte, könnten wir nicht als *Handlungs*repertoire identifizieren. Handlungen werden nicht allein über ihre Substrate in der Körperwelt identifiziert, sondern auch über ihre Beziehungen zu Überzeugungen, Absichten und weiteren Handlungen des Akteurs, genauer: über die inferentiellen Beziehungen, die zwischen der Handlung als Konklusion eines praktischen Schlusses (›X tut H‹) und den entsprechenden Prämissen bestehen.[49] Diese Einführung ›semantischer‹ Identitätsbedingungen für Handlungen hat zur Folge, daß die holistischen Zuschreibungsbedingungen mentaler Prädikate unmittelbar auf Handlungsprädikate übergreifen. Und die Frage, ob es nicht geistbegabte Wesen mit einem Handlungsrepertoire geben könnte, das sich mit dem unsrigen in keinem einzigen Element überschneidet, können wir mit den gleichen guten Gründen verneinen, mit denen Davidson die Frage verneint, ob es eine in die unsere unübersetzbare Sprache geben könnte.[50]

Diese Überlegungen sprechen dafür, daß ein Wesen, dem wir Handlungsfähigkeit zugestehen, einen humanoiden Körper haben muß. Leider hilft uns diese Feststellung nicht viel weiter. Zum einen mögen die Roboterfreunde leichten Herzens zugeben, daß ihre starke KI-These nicht schon durch Roboter von der Art heutiger Industrieroboter gestützt wird. Interessanterweise bewegt sich die Robotik ohnehin in diese Richtung: Sie orientiert sich – wenn auch aus nichtphilosophischen Gründen – bei ihren Problemlösungen an biologischen Vorbildern (Stichwort ›Bionik‹). Zum anderen kann vernünftigerweise keine vollständige

45 In Debatten über den Begriff der Person ist in diesem Sinne darauf hingewiesen worden, daß Prädikate wie »›is sitting down‹ and ›is coiling a rope‹ [...] do not free us from the necessity of positing that the entity to which they are ascribed be humanoid in biology.« McCall, *Concepts of Person*, S. 104. Unsere Möglichkeiten, überhaupt Handlungen zu vollziehen, hängen von zahllosen kontingenten ›Normalbedingungen‹ sowohl der Außenwelt als auch unserer körperlichen Organisation ab, von denen die analytische Handlungstheorie meist abstrahiert. Vgl. dazu Buekens, »Normal Circumstances«.

46 Descartes, *Abhandlung über die Methode des richtigen Vernunftgebrauchs*, S. 53.

47 Ebd.

48 Ebd.

49 Vgl. v. Wright, *Erklären und Verstehen*, S. 102 f.

50 Vgl. Davidson, »Was ist eigentlich ein Begriffsschema?«

Übereinstimmung der Handlungsrepertoires gefordert werden (dies wäre genauso unsinnig wie im Falle der Überzeugungen). Es war immer nur von ›genügend großen Überschneidungen‹ die Rede. Damit ist aber die Frage, *wie ähnlich* eine Maschine uns sein müßte, damit wir den Artenchauvinismus aufgeben, immer noch offen. Der Gradualismus scheint das letzte Wort zu behalten.

Es trifft sich daher gut, daß die genannte Frage in unserem Zusammenhang überhaupt nicht beantwortet werden muß. Dies zeigt eine Rückbesinnung auf das ursprüngliche Theorieziel der KI. Die KI hat sich auf die funktionalistische These der multiplen Realisierbarkeit des Mentalen verpflichtet. Damit es so etwas wie Künstliche Intelligenz geben kann, muß Intelligenz aus ihrer Verstrickung mit kontingenten physischen Eigenschaften menschlicher Intelligenzträger herausgelöst werden können.[51] *Mit der Roboterantwort ist der Boden dieses Forschungsprogramms verlassen.* Wenn nämlich die Roboterantwort zutreffen soll, wird völlig unklar, welche Merkmale des Menschen wir überhaupt noch als kontingent ansehen können. Sind die dem Roboter zugeschriebenen Fähigkeiten nicht nur an mentale, sondern auch an bestimmte physische Voraussetzungen gebunden, dann kehren auf diese Weise Restriktionen zurück, von denen die GOFAI gerade abstrahieren zu dürfen glaubte. Durch den Übergang von mentalen zu körperlichen Operationen haben die Vertreter der Roboterantwort *nolens volens* die These der multiplen Realisierbarkeit entscheidend geschwächt. Was die GOFAI anstrebte, war ein künstlicher *Geist*, nicht ein künstlicher Körper oder ein künstliches Gehirn. Wenn wir nun menschenähnliche Roboter oder gar biologisch-technische Hybriden einführen müssen, um die Zuschreibung mentaler Zustände zu retten, läuft dies auf eine Abkehr vom ursprünglichen Theorieziel der KI hinaus.[52] Dann sind wir längst auf dem Wege von der künstlichen Intelligenz zum künstlichen *Menschen*. Je mehr ein System können, je menschenähnlicher es sein muß, um als geistbegabt zu gelten, desto weiter bewegt sich die KI auf ihre Kritiker zu. Die Roboterantwort stellt also, aller Rhetorik zum Trotz, eine *Rückzugsposition* der KI dar. Alan Turing war auf der Suche nach »a fairly sharp line between the physical and the intellectual capacities of man«.[53] Vieles spricht dafür, daß es diese Grenze nicht gibt.

51 Ich habe an anderer Stelle argumentiert, daß der Funktionalismus in der Philosophie des Geistes aus zwei Teilthesen besteht, von denen die eine richtig, die andere falsch ist. Dafür müssen sie natürlich voneinander unabhängig sein. Es ist nicht zu sehen, warum aus der These von der ontologischen Abstraktheit des Mentalen, die ich teile, die der multiplen Realisierbarkeit folgen sollte. Vgl. Keil, »Die zwei Teilthesen des Funktionalismus«.

52 Diese Überlegung ist derjenigen Peter Golds verwandt, daß jede Verringerung der Modelldistanz zwischen Mensch und Maschine den Erkenntniswert der KI vermindert; vgl. Golds Beitrag in diesem Band. – Übrigens habe ich die Roboterfreunde sehr geschont: Neben den Handlungen, die auf bestimmten physischen Voraussetzungen beruhen, enthält ein menschliches Handlungsrepertoire eine Menge *institutionell* und *sozial* voraussetzungsreicher Handlungen (zum Beispiel sein Testament machen, auf Baisse spekulieren), für deren Vollzug ein Wesen entsprechend lebensweltlich sozialisiert sein müßte.

53 Turing, »Computing Machinery and Intelligence«, S. 434.

Literatur

Beckermann, A., »Semantische Maschinen«, in: Forum für Philosophie Bad Homburg (Hg.), *Intentionalität und Verstehen*, Frankfurt am Main 1990, S. 196-211.

Boden, M. A., *Artificial Intelligence and Natural Man*, Hassocks, Sussex 1977.

Buekens, F., »Normal Circumstances«, in: F. Buekens und H. Parret (Hg.), *Inquiries into Davidson's Philosophy of Mind and Language* (im Erscheinen).

Carnap, R., »Psychologie in physikalischer Sprache«, in: *Erkenntnis* 3 (1932), S. 107-142.

Danto, A. C., *Analytical Philosophy of Action*, Cambridge 1973.

Davidson, D., »Der materielle Geist« (1973), in: ders., *Handlung und Ereignis*, S. 343-362.

–, »Handeln« (1971), in: ders., *Handlung und Ereignis*, S. 73-98.

–, *Handlung und Ereignis*, Frankfurt am Main 1985.

–, »Radikale Interpretation« (1973), in: ders., *Wahrheit und Interpretation*, S. 183-203.

–, »Turing's Test«, in: K. A. Mohyeldin Said u. a. (Hg.), *Modelling the Mind*, Oxford 1990, S. 1-11.

–, *Wahrheit und Interpretation*, Frankfurt am Main 1986.

–, »Was ist eigentlich ein Begriffsschema?« (1974), in: ders., *Wahrheit und Interpretation*, S. 261-282.
Dennett, D. C., »Science, Philosophy, and Interpretation«, in: *Behavioral and Brain Sciences* 11 (1988), S. 535-546.
–, »True Believers« (1981), in: ders., *The Intentional Stance*, Cambridge, Mass. 1987, S. 14-35.
Descartes, R., *Abhandlung über die Methode des richtigen Vernunftgebrauchs* (1637), übersetzt von Kuno Fischer, Stuttgart 1961.
Dretske, F. I., *Explaining Behavior. Reasons in a World of Causes*, Cambridge, Mass. 1988.
Dreyfus, H. L., und St. Dreyfus, »Coping With Change: Why People Can and Computers Can't«, in: L. Nagl und R. Heinrich (Hg.), *Wo steht die Analytische Philosophie heute?*, Wien/München 1986, S. 150 bis 170.
Hauser, L., »Acting, Intending and Artificial Intelligence«, in: *Behavior and Philosophy* 22 (1994), S. 23-28.
Janich, P., *Grenzen der Naturwissenschaft*, München 1992.
Keil, G., »Die zwei Teilthesen des Funktionalismus«, in: P. Schefe u. a. (Hg.), *Informatik und Philosophie*, Mannheim/Leipzig/Wien 1993, S. 195-209.
–, *Kritik des Naturalismus*, Berlin/New York 1993.
–, »Zu Russells These vom Absterben des Kausalbegriffs in den Wissenschaften«, in: Ch. Hubig und H. Poser (Hg.), *Cognitio humana – Dynamik des Wissens und der Werte. XVII. Deutscher Kongreß für Philosophie*, Leipzig 1996, S. 522-529.
McCall, C., *Concepts of Person. An Analysis of Concepts of Person, Self and Human Being*, Aldershot 1990.
Pylyshyn, Z. W., *Computation and Cognition*, Cambridge, Mass. 1984.
– (Hg.), *The Robot's Dilemma. The Frame Problem in Artificial Intelligence*, Norwood, N. J. 1987.
Rheinwald, R., »Können Maschinen eine Sprache sprechen? Sind Computerprogramme syntaktisch oder semantisch?«, in: *Kognitionswissenschaften* 2 (1991), S. 37-49.
Schnädelbach, H., »Rationalität und Normativität« (1990), in: ders., *Zur Rehabilitierung des ›animal rationale‹*, Frankfurt am Main 1992, S. 79-103.
Searle, J. R., *Geist, Hirn und Wissenschaft*, Frankfurt am Main 1986.
–, »Intentionalität und der Gebrauch der Sprache« (1978), in: G. Grewendorf (Hg.), *Sprechakttheorie und Semantik*, Frankfurt am Main 1979.
Taylor, Ch., *The Explanation of Behaviour*, London 1964.
Taylor, R., *Action and Purpose*, Englewood Cliffs, N. J. 1966.
Tetens, H., *Geist, Gehirn, Maschine*, Stuttgart 1994.
–, »Maschinen, die ›Geist haben‹. Überlegungen zur Künstlichen Intelligenz im Anschluß an John Searle«. Vortrag auf dem XV. Deutschen Kongreß für Philosophie 1990 in Hamburg, Typoskript.
Turing, A., »Computing Machinery and Intelligence«, in: *Mind* 59 (1950), S. 433-460; deutsch: »Rechenmaschinen und Intelligenz«, in: *Intelligence Service. Schriften*, Berlin 1987, S. 147-182.
Wittgenstein, L., *Philosophische Untersuchungen*, Frankfurt am Main 1960.
Wright, G. H. v., *Erklären und Verstehen*, Frankfurt am Main 1974.