Ich fasse zusammen: Nach meiner Auffassung interagieren Geist und Körper, sie sind aber nicht zwei verschiedene Dinge, denn geistige Phänomene sind einfach Eigenschaften des Hirns. Eine Art, diese Position zu kennzeichnen, ist folgende: mit ihr wird sowohl der Physikalismus als auch der Mentalismus vertreten. Definieren wir einmal den »naiven Physikalismus« als die Auffassung, nach der in der Welt ausschließlich Materie-Teilchen mit ihren Eigenschaften und Beziehungen existieren. Das physikalische Modell der Realität ist so machtvoll, daß schwer zu sehen ist, wie wir den naiven Physikalismus ernsthaft in Frage stellen können. – Definieren wir den »naiven Mentalismus« einmal als die Auffassung, daß gewisse geistige Phänomene wirklich existieren. Es gibt Geisteszustände wirklich; einige davon sind bewußt; viele sind intentional; alle sind subjektiv; und viele haben kausalen Einfluß auf materielle Ereignisse in der Welt. Die These dieses ersten Kapitels läßt sich nun ganz einfach formulieren. Der naive Mentalismus und der naive Physikalismus sind vollkommen miteinander verträglich. In der Tat, soweit wir überhaupt etwas darüber wissen, wie die Welt funktioniert: Sie sind nicht nur miteinander verträglich, sie sind beide wahr.

II
Können Computer denken?

Im vorigen Kapitel habe ich eine Lösung des sog. »Körper/Geist-Problems« zumindest in Umrissen vorgestellt. Auch wenn wir nicht im einzelnen wissen, wie das Hirn funktioniert, so wissen wir doch genug, um von den allgemeinen Beziehungen zwischen Hirnvorgängen und geistigen Vorgängen eine Vorstellung zu haben. Geistige Vorgänge sind vom Verhalten der Bestandteile des Hirns verursacht. Zugleich sind sie in der Struktur realisiert, die aus diesen Bestandteilen besteht. Diese Antwort verträgt sich m. E. damit, wie biologische Phänomene in der Biologie üblicherweise angegangen werden. Ja, angesichts dessen, was wir über das Funktionieren der Welt wissen, ist es so etwas wie eine Antwort des gesunden Menschenverstands auf die Frage. Trotzdem ist es sicher eine Mindermeinung. Die in der Philosophie, Psychologie und Künstlichen Intelligenz herrschende Auffassung betont die Analogien zwischen dem Funktionieren des menschlichen Hirns und dem Funktionieren digitaler Computer. Gemäß der extremsten Version dieser Auffassung ist das Hirn einfach ein digitaler Computer und der Geist einfach ein Computerprogramm. Diese Auffassung nenne ich »starke Künstliche Intelligenz« (oder »starke KI«); sie läßt sich so zusammenfassen: Der Geist verhält sich zum Hirn wie das Programm zur Hardware des Computers.

Eine Konsequenz dieser Auffassung ist, daß am Geist nichts wesentlich Biologisches ist. Die Programme, die menschliche Intelligenz ausmachen, könnten von Computern unbestimmt vieler verschiedener Hardware-Typen realisiert werden, und das Hirn ist zufällig ein Computer mit so einer Hardware. Jedes beliebige materielle System mit dem richtigen Programm, den richtigen Inputs und Outputs hätte nach dieser Auffassung in genau demselben Sinn einen Geist, in dem Sie und ich einen Geist haben. Würde man also beispielsweise einen Computer aus alten Bierdosen mit Windmühlen-Antrieb bauen, dann hätte dieser Computer einen Geist, wenn er das richtige Programm hätte. Und es geht hier nicht darum, daß dieser Computer nach allem, was wir wissen, vielleicht Gedanken und Gefühle hat, sondern

vielmehr darum, daß er Gedanken und Gefühle haben muß. Denn hinter Gedanken und Gefühlen steckt nichts weiter dahinter; sie zu haben, heißt einfach: das richtige Programm zu realisieren.

Die meisten Vertreter dieser Auffassung meinen, daß wir bisher noch kein Programm entwickelt haben, das ein Geist ist. Aber fast alle von ihnen halten es bloß für eine Frage der Zeit, bis die Fachleute aus der Computerwissenschaft und der Künstlichen Intelligenz schließlich Hardware und Programme entwickeln, die einem menschlichen Hirn und Geist entsprechen. Das werden dann künstliche Hirne und Geister sein, die den menschlichen in jeder Hinsicht gleichwertig sind.

Wer nicht mit der Künstlichen Intelligenz vertraut ist, ist wohl sehr erstaunt, wenn er erfährt, daß überhaupt irgendjemand solch eine Auffassung für wahr hält. Bevor ich mich der Kritik zuwende, möchte ich Ihnen deshalb ein paar Beispiele dafür geben, was für Behauptungen von Vertretern dieses Gebiets tatsächlich aufgestellt worden sind. Herbert Simon von der Carnegie-Mellon Universität sagt, daß wir bereits Maschinen haben, die buchstäblich denken können. Man muß gar nicht erst auf eine künftige Maschine warten, denn die vorhandenen digitalen Computer haben bereits in genau demselben Sinn Gedanken, in dem Sie und ich welche haben. Nun, das muß man sich einmal vorstellen! Philosophen haben sich über Jahrhunderte damit herumgeplagt, ob eine Maschine nun denken könne oder nicht – und nun entdecken wir, daß an der Carnegie-Mellon Universität schon solche Maschinen stehen. Alan Newell, ein Kollege von Simon, behauptet, wir hätten jetzt entdeckt (und man beachte, daß er »entdeckt« sagt, und nicht »die Hypothese aufgestellt« oder »die Möglichkeit erwogen«; nein, wir haben *entdeckt*), daß Intelligenz bloß eine Sache materieller Symbol-Manipulation ist; daß sie keinen wesentlichen Zusammenhang mit irgendeiner speziellen Art der biologischen oder materiellen Realisierung hat, sei die nun naß oder hart. Vielmehr sei jedes beliebige System, das materielle Symbole auf die richtige Weise manipulieren kann, in demselben wörtlichen Sinne intelligenzbegabt, in dem Menschen es sind. Simon und Newell betonen dankenswerterweise, daß an diesen Behauptungen nichts Metaphorisches ist; sie meinen das alles ganz wörtlich. Freeman Dyson wird mit der Bemerkung zitiert, daß Computer uns gegenüber einen evolutionären Vorteil

haben. Bewußtsein ist ja bloß eine Sache formaler Vorgänge, und Computer lassen es zu, daß diese Vorgänge in Substanzen ablaufen, die in einem immer kälter werdenden Universum viel besser überleben können als Lebewesen, die wie wir aus nassem und anfälligem Material bestehen. Marvin Minsky vom MIT sagt, die nächste Computer-Generation werde so intelligent sein, daß wir »Glück haben, wenn sie uns noch als Haustiere halten«. Meine absolute Lieblingsübertreibung über den digitalen Computer stammt von John McCarthy, dem Erfinder des Ausdrucks »artificial intelligence«. McCarthy sagt, sogar »von so einfachen Maschinen wie Thermostaten kann man sagen, daß sie Überzeugungen haben«. Laut McCarthy kann man in der Tat fast jeder Maschine, die Probleme lösen kann, Überzeugungen zuschreiben. Ich bewundere McCarthys Mut. Einmal habe ich ihn gefragt: »Was für Überzeugungen hat Ihr Thermostat?« Und er sagte: »Mein Thermostat hat drei Überzeugungen: Es ist zu warm hier drinnen, es ist zu kalt hier drinnen, und die Temperatur hier drinnen ist gerade richtig.« Als Philosoph gefallen mir all diese Behauptungen aus einem einfachen Grund. Im Gegensatz zu den meisten philosophischen Thesen sind sie ziemlich klar, und sie lassen sich einfach und endgültig widerlegen. Diese Widerlegung werde ich im vorliegenden Kapitel unternehmen.

Die Widerlegung hat überhaupt nichts damit zu tun, auf welcher Stufe sich die Computertechnik befindet. Es ist wichtig, dies hervorzuheben, denn es besteht immer die Verlockung anzunehmen, die Lösung unserer Probleme müsse warten, bis irgendein bislang noch nicht geschaffenes technisches Wunderwerk endlich da ist. Doch tatsächlich ist die Widerlegung völlig unabhängig davon, was auch immer der Stand der Technik ist. Sie hat mit der Definition eines digitalen Computers zu tun – damit, was ein digitaler Computer ist.

Für unser Verständnis vom digitalen Computer ist es wesentlich, daß sich seine Operationen rein formal angeben lassen; d. h. wir geben Operationsschritte des Computers mit abstrakten Symbolen an – beispielsweise mit Folgen von Nullen und Einsen, die auf ein Blatt gedruckt sind. Eine typische Computer-»Regel« legt fest, daß eine Maschine, wenn sie sich in einem gewissen Zustand befindet und auf ihrem Band ein bestimmtes Symbol hat, dann eine gewisse Operation ausführt (z. B. das Symbol löscht oder ein anderes druckt oder das Band ein Feld weiter nach links laufen

läßt) und anschließend in einen andern Zustand übergeht. Doch die Symbole haben keine Bedeutung; sie haben keinen semantischen Gehalt; sie beziehen sich auf nichts. Sie müssen ausschließlich mittels ihrer formalen oder syntaktischen Struktur spezifiziert werden. Die Nullen und Einsen beispielsweise sind bloß Ziffern; sie stehen nicht einmal für Zahlen. Tatsächlich verdanken digitale Computer ihre Stärke gerade dieser Eigenschaft. Hardware ein und desselben Typs kann – wenn sie passend konstruiert ist – unbestimmt viele verschiedene Programme zum Laufen bringen. Und ein und dasselbe Programm kann in unbestimmt vielen Typen von Hardware laufen.

Doch an diesem Merkmal von Programmen – rein formal oder syntaktisch definiert zu sein – scheitert die Auffassung, geistige Vorgänge und Programmvorgänge seien identisch. Und der Grund dafür läßt sich ganz einfach angeben. Es gehört mehr dazu, einen Geist zu haben, als formale oder syntaktische Prozesse zu durchlaufen. Unsere Geisteszustände haben kraft Definition verschiedene Inhalte. Wenn ich an Kansas City denke oder wenn ich wünsche, es wäre noch ein kaltes Bier für mich da, oder wenn ich mir überlege, ob der Zinssatz gesenkt wird, dann hat mein Geisteszustand jedesmal zusätzlich zu seinen formalen Eigenschaften (welche auch immer das sein mögen), auch noch einen gewissen geistigen Gehalt. Das heißt: selbst wenn meine Gedanken in mir als Symbolketten auftreten, so muß der Gedanke mehr sein als bloß eine abstrakte Folge, denn eine Folge kann aus eigener Kraft keine Bedeutung haben. Wenn meine Gedanken *von etwas handeln,* dann müssen die Folgen eine *Bedeutung* haben, dank derer die Gedanken von diesen Dingen handeln. In einem Wort: Der Geist hat mehr als nur eine Syntax, er hat eine Semantik. Kein Computerprogramm kann jemals ein Geist sein, und zwar aus dem einfachen Grunde, weil ein Computerprogramm bloß syntaktisch und der Geist mehr als bloß syntaktisch ist. Der Geist ist semantisch – semantisch in dem Sinne, daß er mehr hat als eine formale Struktur: er hat einen Gehalt.

Um dies zu veranschaulichen, habe ich ein gewisses Gedankenexperiment entwickelt. Man stelle sich vor, ein paar Computerprogrammierer hätten ein Programm geschrieben, das einen Computer in die Lage versetzt, das Verständnis des Chinesischen zu simulieren. Wenn dem Computer also beispielsweise eine Frage

auf Chinesisch vorgelegt wird, wird er die Frage mit seinem Gedächtnis (seiner Datenbasis) konfrontieren und passende Antworten auf Chinesisch produzieren. Nehmen wir der Argumentation halber einmal an, daß die Antworten des Computers genauso gut sind wie die eines chinesischen Muttersprachlers. Versteht der Computer dann allein deshalb schon Chinesisch? Versteht er wirklich Chinesisch – in dem wörtlichen Sinne, in dem Chinesen Chinesisch verstehen? Nun, stellen Sie sich vor, Sie wären in ein Zimmer eingesperrt, in dem mehrere Körbe mit chinesischen Symbolen stehen. Und stellen Sie sich vor, daß sie (wie ich) kein Wort Chinesisch verstehen, daß Ihnen allerdings ein auf Deutsch abgefaßtes Regelwerk für die Handhabung dieser chinesischen Symbole gegeben worden wäre. Die Regeln geben rein formal – nur mit Rückgriff auf die Syntax und nicht auf die Semantik der Symbole – an, was mit den Symbolen gemacht werden soll. Eine solche Regel mag lauten: »Nimm ein Kritzel-Kratzel-Zeichen aus Korb 1 und lege es neben ein Schnörkel-Schnarkel-Zeichen aus Korb 2«. Nehmen wir nun an, daß irgendwelche andern chinesischen Symbole in das Zimmer gereicht werden, und daß Ihnen noch zusätzliche Regeln dafür gegeben werden, welche chinesischen Symbole jeweils aus dem Zimmer herauszureichen sind. Die hereingereichten Symbole werden von den Leuten draußen »Fragen« genannt, und die Symbole, die Sie dann aus dem Zimmer herausreichen, »Antworten« – aber dies geschieht ohne Ihr Wissen. Nehmen wir außerdem an, daß die Programme so trefflich und Ihre Ausführung so brav sind, daß Ihre Antworten sich schon bald nicht mehr von denen eines chinesischen Muttersprachlers unterscheiden lassen. Da sind Sie nun also in Ihrem Zimmer eingesperrt und stellen Ihre chinesischen Symbole zusammen; Ihnen werden chinesische Symbole hereingereicht und daraufhin reichen Sie chinesische Symbole heraus. In so einer Lage, wie ich sie gerade beschrieben habe, könnten Sie einfach dadurch, was Sie mit den formalen Symbolen anstellen, kein bißchen Chinesisch lernen.

Der Witz der Geschichte ist nun schlicht folgender: weil Sie ein formales Computerprogramm ausführen, verhalten Sie sich aus der Sicht eines Außenstehenden genauso, als verstünden Sie Chinesisch – und dennoch verstehen Sie nicht ein Wort Chinesisch. Wenn aber die Ausführung eines passenden Computerprogramms *in Ihrem Fall* nicht ausreicht, um Chinesisch zu verste-

hen, dann reicht das auch bei *keinem andern digitalen Computer* aus. Auch hierfür läßt sich der Grund ganz einfach angeben. Wenn Sie kein Chinesisch verstehen, dann könnte auch kein anderer Computer Chinesisch verstehen; denn kraft seiner Ausführung eines Programmes hat kein digitaler Computer irgend etwas, das Sie nicht haben. Der Computer hat – genau wie Sie – nichts außer einem formalen Programm für die Handhabung uninterpretierter chinesischer Symbole. Um es zu wiederholen: Ein Computer hat eine Syntax, aber keine Semantik. Der ganze Witz der Parabel mit dem Chinesisch-Zimmer besteht darin, uns an etwas zu erinnern, das wir die ganze Zeit über gewußt haben. Eine Sprache verstehen oder überhaupt sich in gewissen Geisteszuständen befinden, dazu gehört mehr als daß man bloß ein paar formale Symbole hat. Dazu gehört, daß diese Symbole eine Interpretation (oder eine Bedeutung) haben. Und ein digitaler Computer kann laut Definition einfach nur formale Symbole haben, denn die Funktionsweise des Computers ist, wie bereits gesagt, dadurch definiert, daß er gewisse Programme ausführen kann. Und diese Programme lassen sich rein formal angeben – d. h. sie haben keinen semantischen Gehalt.

Was diese Argumentation leistet, können wir sehen, wenn wir zweierlei einander gegenüberstellen: wie es einerseits ist, etwas auf Deutsch gefragt zu werden und auf Deutsch zu beantworten, und wie es andererseits ist, in einer Sprache, wo wir die Bedeutung keines einzigen Wortes kennen, etwas gefragt zu werden und zu antworten. Stellen Sie sich vor, Ihnen würden im Chinesisch-Zimmer auch Fragen auf Deutsch gestellt – etwa über Ihr Alter und Ihre Lebensgeschichte – und Sie würden diese Fragen beantworten. Worin besteht der Unterschied zwischen dem Fall, wo Chinesisch benutzt wird, und dem, wo Deutsch benutzt wird? Nun, der Unterschied liegt wiederum auf der Hand, wenn Sie (wie ich) kein Chinesisch verstehen, wohl aber Deutsch. Sie verstehen die Fragen auf Deutsch, weil Sie in Symbolen ausgedrückt sind, deren Bedeutung Sie kennen. Entsprechend erzeugen Sie – wenn Sie auf Deutsch antworten – Symbole, die für Sie eine Bedeutung haben. Doch wenn das Chinesische benutzt wird, ist nichts von alledem der Fall. Im Falle des Chinesischen manipulieren Sie einfach formale Symbole in Übereinstimmung mit einem Computerprogramm und Sie verbinden keine Bedeutung mit irgendeinem dieser Symbole.

Auf diese Argumentation gab es mehrere Erwiderungen – von Fachleuten aus der Künstlichen Intelligenz, der Psychologie, auch aus der Philosophie. All diesen Antworten ist etwas gemeinsam: sie sind alle inadäquat. Und es liegt auf der Hand, warum sie indäquat sein müssen. Denn die Argumentation beruht ja auf einer einfachen logischen Wahrheit, und zwar: Syntax allein reicht nicht hin für Semantik, und digitale Computer haben, als Computer, kraft Definition bloß eine Syntax.

Ich möchte dies durch die Betrachtung von ein paar Argumenten klarmachen, die oft gegen mich vorgebracht werden.

Manchmal wird versucht, dem Beispiel mit dem Chinesisch-Zimmer mit dem Hinweis beizukommen, das gesamte System verstehe Chinesisch. Die Idee dabei ist folgende: Auch wenn ich – die Person, die im Zimmer mit den Symbolen zu tun hat – kein Chinesisch verstehe, so bin ich doch bloß die Zentraleinheit des Computersystems. Der Einwand besagt nun, es sei das ganze System – als eine Gesamtheit aufgefaßt –, das Chinesisch versteht: dazu gehören das Zimmer, die Körbe mit den Symbolen, die Register mit den Programmen und vielleicht noch anderes mehr. Doch dagegen spricht wiederum genau derselbe Einwand. Das System kann einfach nicht von der Syntax zur Semantik gelangen. Ich, die Zentraleinheit, kann einfach nicht hinter die Bedeutung eines einzigen Symbols kommen; und das kann dann auch das Gesamtsystem nicht.

Häufig wird auch damit reagiert, daß man sich vorstellt, wir bauten das Chinesisch-Programm in einen Roboter ein. Wenn dieser Roboter nun herumliefe, auf die Welt einwirkte, und sie auf ihn, wäre das dann keine Garantie dafür, daß er Chinesisch versteht? Die Unterscheidung zwischen Syntax und Semantik bleibt unumstößlich, und auch dieses Manöver kommt nicht gegen sie an. Solange wir unterstellen, daß der Roboter als Hirn nur einen Computer hat, könnte er einfach nicht von der Syntax zur Semantik gelangen, auch wenn er sich genauso benehmen mag, als verstünde er Chinesisch. Dies läßt sich einsehen, wenn man sich vorstellt, ich wäre der Computer. In einem Zimmer im Schädel des Roboters schiebe ich Symbole herum; ich weiß aber nicht, daß einige dieser Symbole von Fernsehkameras zu mir kommen, die am Kopf des Roboters angebracht sind, und daß andere hinausgeschickt werden, um die Arme und Beine des Roboters in Bewegung zu setzen. Solange ich nichts weiter als ein

formales Computerprogramm habe, kann ich mit keinem Symbol eine Bedeutung verbinden. Der Roboter steht zwar in kausaler Interaktion mit der Außenwelt, aber das hilft mir nicht dabei, mit den Symbolen eine Bedeutung zu verbinden, solange ich das nicht irgendwie herausfinde. Angenommen, der Roboter nimmt einen Hamburger in die Hand, und dadurch wird das Hamburger-Symbol auf seinen Weg in mein Zimmer gebracht. Solange ich nur das Symbol habe und nicht weiß, wodurch es verursacht wird und wie es ins Zimmer gelangt ist, kann ich nicht wissen, was es bedeutet. Die kausale Interaktion zwischen dem Roboter und der Welt ist unerheblich, solange sie nicht in dem einen oder anderen Geist repräsentiert ist. Doch das geht nicht, falls der sog. Geist nur aus rein formalen, syntaktischen Operationen besteht.

Es ist wichtig, genau zu sehen, was mit meiner Argumentation behauptet wird und was nicht. Betrachten wir die eingangs erwähnte Frage »Könnte eine Maschine denken?« Nun, in einem gewissen Sinn sind wir natürlich alle Maschinen. Den Inhalt unseres Hirns können wir als eine Maschine aus Fleisch auffassen. Und natürlich können wir alle denken. In wenigstens einem Sinn von »Maschine« – und zwar in dem Sinn, in dem eine Maschine einfach ein materielles System ist, das gewisse Arten von Operationen ausführen kann – sind wir alle Maschinen, und wir können denken. Mithin gibt es trivialerweise Maschinen, die denken können. Aber das war nicht die Frage, die uns geplagt hat. Probieren wir eine andere Formulierung aus. Könnte ein Artefakt denken? Könnte eine Maschine von Menschenhand denken? Nun, das hängt wiederum davon ab, um was für eine Art Artefakt es sich handelt. Angenommen wir entwickeln eine Maschine, die Molekül für Molekül einem Menschen gleicht. Nun, wenn man die Ursachen kopieren kann, dann kann man ja wohl auch die Wirkungen kopieren. Also lautet die Antwort auch auf diese Frage, zumindest im Prinzip, wiederum trivialerweise: Ja. Könnte man eine Maschine bauen, die die gleiche Struktur hätte wie ein Mensch, dann könnte diese Maschine vermutlich denken. Ja, sie wäre ein Ersatz-Mensch. Versuchen wir's also noch einmal. Die Frage lautet nicht: »Kann eine Maschine denken?« oder »Kann ein Artefakt denken?« Die Frage lautet: »Kann ein digitaler Computer denken?« Bei der Deutung dieser Frage müssen wir allerdings wiederum sehr behutsam sein. Aus mathematischer Sicht läßt sich überhaupt alles so beschreiben, *als ob* es ein

digitaler Computer wäre. Denn alles läßt sich als Realisierung oder Ausführung eines Computerprogramms beschreiben. Der Federhalter auf dem Schreibtisch vor mir läßt sich – in einem völlig trivialen Sinn – als digitaler Computer beschreiben. Er hat halt ein sehr langweiliges Computerprogramm. Das Programm besagt: »Bleib, wo du bist.« In diesem Sinn ist also einfach alles ein digitaler Computer, weil einfach alles sich so beschreiben läßt, als führe es ein Computerprogramm aus. Deshalb gibt es auf unsere Frage wiederum eine triviale Antwort. Unsere Hirne sind natürlich digitale Computer, denn sie führen ja jede Menge Computerprogramme aus. Und natürlich können unsere Hirne denken. Also hat unsere Frage dann wiederum eine triviale Antwort. Doch das war ja gar nicht die Frage, die wir eigentlich stellen wollten. Was wir wissen wollen, ist: »Kann ein digitaler Computer – so wie digitale Computer definiert sind – denken?« Und das heißt: »Ist die Realisierung oder Ausführung des richtigen Computerprogramms mit den richtigen Inputs und Outputs hinreichend oder konstitutiv für Denken?« Und auf diese Frage – anders als auf die früheren Fragen – lautet die Antwort offensichtlich »Nein«, und zwar aus dem erläuterten Grund: Das Computerprogramm ist rein syntaktisch definiert. Aber Denken ist mehr als bloß ein Manipulieren bedeutungsloser Symbole; zum Denken gehört semantischer Gehalt mit einer Bedeutung. Wenn wir von »Bedeutung« sprechen, dann meinen wir diese semantischen Gehalte.

Es ist wichtig, noch einmal hervorzuheben, daß hier nicht von einem speziellen Stand der Computertechnik die Rede ist. Die Argumentation hat nichts damit zu tun, welche erstaunlichen Fortschritte der Computerwissenschaft bevorstehen. Sie hat nichts mit der Unterscheidung zwischen seriellen und parallelen Prozessen zu tun, nichts mit dem Umfang von Programmen, nichts mit der Geschwindigkeit von Computeroperationen, nichts mit Computern, die mit ihrer Umwelt kausal interagieren können, und nicht einmal etwas mit der Erfindung von Robotern. Technischer Fortschritt wird immer maßlos übertrieben, aber selbst wenn man die Übertreibungen einmal abzieht, war die Entwicklung von Computern sehr bemerkenswert, und wir haben Grund anzunehmen, daß in der Zukunft noch bemerkenswertere Fortschritte gemacht werden. Wir werden menschliches Verhalten zweifelsohne noch viel besser simulieren können als

bisher. Worauf es mir hier ankommt, ist folgendes: Wenn wir davon reden, daß wir Geisteszustände bzw. einen Geist haben, dann sind all diese Simulationen einfach unerheblich. Die technische Qualität oder die Geschwindigkeit der Rechenvorgänge eines Computers spielen keine Rolle. Wenn es wirklich ein Computer ist, müssen seine Operationen syntaktisch definiert sein; beim Bewußtsein, den Gedanken, Empfindungen, Gefühlen und allem, was sonst noch dazugehört, ist hingegen mehr als bloß eine Syntax im Spiel. Kraft Definition kann der Computer kein *Duplikat* von diesen Merkmalen abgeben, auch wenn er eine noch so leistungsfähige *Simulation* von ihnen ist. Die Schlüssel-Unterscheidung ist hier die zwischen Duplikation und Simulation. Und keine Simulation stellt jemals aus eigener Kraft eine Duplikation dar.

Bis zu diesem Punkt habe ich nur unserem Eindruck eine Grundlage gegeben, daß die Zitate, mit denen ich diese Vorlesung begonnen habe, wirklich so grotesk sind, wie sie ausschauen. Es gibt allerdings eine verwirrende Frage in dieser Diskussion, und zwar: »Warum sollte jemand jemals auf den Gedanken gekommen sein, daß Computer denken können oder Empfindungen, Gefühle und dergleichen haben?« Schließlich können wir ja jeden beliebigen Vorgang, der sich formal beschreiben läßt, mit einem Computer simulieren. So können wir mit einem Computer simulieren, wie das Geld sich in der britischen Wirtschaft bewegt, oder nach welchem Muster die Macht in der Labour-Partei verteilt ist. Wir können Regenstürme in den Grafschaften mit einem Computer simulieren oder auch Warenhausbrände im Osten Londons. Nun nimmt niemand in diesen Fällen an, die Computersimulation sei tatsächlich die Sache selbst; keiner nimmt an, daß die Computersimulation eines Sturmes uns naß macht, oder daß die Computersimulation eines Feuers vermutlich das Haus zerstört. Warum in aller Welt sollte jemand, der bei Trost ist, annehmen, eine Computersimulation geistiger Vorgänge hätte tatsächlich geistige Vorgänge? Die Antwort darauf kenne ich wirklich nicht, denn die Idee kommt mir, offen gesagt, von Anfang an verrückt vor. Ich kann aber ein paar Spekulationen anstellen.

Erstens einmal fühlen sich viele Leute, wenn es um den Geist geht, immer noch zu irgendeiner Form des Behaviourismus hingezogen. Sie meinen, daß, wenn sich ein System so benimmt, als verstünde es Chinesisch, es dann auch wirklich Chinesisch verstehen müsse. Diese Form des Behaviourismus haben wir bereits durch das Argument mit dem Chinesisch-Zimmer widerlegt. Gemäß einer andern, vielerseits gemachten Annahme, gehört der Geist nicht zur biologischen Welt, zur Welt der Natur. Die Auffassung der starken Künstlichen Intelligenz beruht auf der Vorstellung, daß der Geist etwas rein Formales sei, daß er irgendwarum nicht als ein konkretes Ergebnis biologischer Vorgänge wie jedes andere derartige Ergebnis behandelt werden könne. In diesen Diskussionen gibt es also, kurz gesagt, eine Art Rest-Dualismus. Parteigänger der Künstlichen Intelligenz sind der Ansicht, der Geist sei mehr als nur ein Teil der natürlichen biologischen Welt; sie glauben, der Geist lasse sich rein formal charakterisieren. Daran ist paradox, daß in der Literatur der Künstlichen Intelligenz so viel gegen einen sog. »Dualismus« gewettert wird, während doch die gesamte These der starken KI auf einer bestimmten Form von Dualismus beruht. Sie beruht auf der Ablehnung der Idee, daß der Geist einfach ein biologisches Phänomen wie jedes andere in der Welt ist.

Ich möchte dieses Kapitel damit beschließen, daß ich die Thesen der beiden bisherigen Kapitel zusammenstelle. Diese Thesen lassen sich sehr einfach formulieren. Ich werde sie sogar derart einfach formulieren, daß es vielleicht ein bißchen zu viel des Guten ist. Wenn wir sie aber zusammennehmen, dann bekommen wir m. E. eine sehr leistungsfähige Vorstellung von den Beziehungen zwischen dem Geist, dem Hirn und dem Computer. Und die Argumentation hat eine ganz einfache logische Struktur, so daß Sie selbst beurteilen können, ob sie zwingend ist oder nicht. Die erste Prämisse lautet:

1. Hirn verursacht Geist.

Das ist nun natürlich wirklich zu ungenau. Damit ist gemeint, daß die geistigen Vorgänge, die unseres Erachtens einen Geist ausmachen, von Vorgängen im Hirn verursacht – und zwar vollständig verursacht – sind. Aber seien wir einmal ungenau, kürzen wir das mit drei Worten ab: Hirn verursacht Geist. Und damit wird einfach etwas Wahres über die Welt gesagt. Nehmen wir nun Feststellung Nummer zwei:

2. *Syntax reicht nicht für Semantik aus.*

Das ist eine begriffliche Wahrheit. Damit wird nur unsere Unterscheidung zwischen dem Begriff von etwas rein Formalem und dem Begriff von etwas mit einem Inhalt zum Ausdruck gebracht. Nehmen wir nun zu diesen beiden Feststellungen eine dritte und vierte hinzu:

3. *Computerprogramme sind vollständig durch ihre formale (oder syntaktische) Struktur definiert.*

Ich unterstelle, daß dies kraft Definition stimmt; es gehört zu unserem Begriff eines Computerprogrammes.

4. *Ein Geist hat geistige Inhalte, und zwar semantische Inhalte.*

Ich unterstelle, daß es einfach auf der Hand liegt, daß unser Geist so funktioniert. Meine Gedanken, Überzeugungen und Wünsche handeln von etwas, sie beziehen sich auf etwas, sie betreffen Zustände und Sachverhalte in der Welt; und das tun sie, weil ihr Inhalt sie auf diese Sachverhalte in der Welt richtet. Wir können nun aus diesen vier Prämissen unseren ersten Schluß ziehen; er folgt offensichtlich aus den Prämissen 2, 3 und 4:

SCHLUSSFOLGERUNG 1
Kein Computerprogramm kann aus eigener Kraft einem System einen Geist geben. Ein Programm ist, kurz gesagt, kein Geist und reicht – für sich selbst genommen – nicht hin, um einen Geist zu haben.

Dies ist nun ein sehr starkes Ergebnis, denn es besagt, daß das Projekt der Geist-Erschaffung durch Programmeschreiben von vornherein zum Scheitern verurteilt ist. Und ich möchte noch einmal hervorheben, daß dies nichts mit dem Stand der Technik oder der Komplexität des Programmes zu tun hat. Dies ist eine rein formale oder logische Konsequenz einiger Axiome, denen alle (oder fast alle) an der Diskussion Beteiligten zustimmen. Selbst die meisten eingefleischten Enthusiasten für Künstliche Intelligenz halten es für eine biologische Tatsache, daß Hirnvorgänge Geisteszustände verursachen, und sie stimmen darin überein, daß Programme rein formal definiert sind. Wenn man aber diese Schlußfolgerung mit gewissen anderen wohlbekannten Tatsachen zusammennimmt, so folgt daraus unmittelbar, daß das Projekt der starken KI undurchführbar ist.

Wie dem auch sei, wo wir diese Axiome schon einmal haben, wollen wir doch einmal schauen, was wir sonst noch aus ihnen ableiten können. Hier ist eine zweite Schlußfolgerung:

SCHLUSSFOLGERUNG 2
Hirnfunktionen können Geist nicht ausschließlich dadurch verursachen, daß sie ein Computerprogramm realisieren.

Diese zweite Schlußfolgerung folgt aus der ersten Prämisse im Verbund mit unserer ersten Schlußfolgerung. Daraus, daß das Hirn den Geist verursacht, und Programme das nicht schaffen, folgt, daß ein Hirn nicht ausschließlich dadurch einen Geist verursachen kann, daß es ein Computerprogramm realisiert. Auch das halte ich für ein wichtiges Ergebnis, denn daraus ergibt sich, daß das Hirn kein – oder zumindest nicht bloß ein – digitaler Computer ist. Wir haben oben gesehen, daß sich alles trivialerweise so beschreiben läßt, als wäre es ein digitaler Computer, und Hirne sind da keine Ausnahme. Die Bedeutung dieser Schlußfolgerung liegt jedoch darin, daß die durch einen Computer erfaßbaren Eigenschaften des Hirns einfach nicht ausreichen, um zu erklären, wie es Geisteszustände bewirkt. Und dies sollte uns allerdings ohnehin wie ein vernünftiges wissenschaftliches Ergebnis vorkommen, denn es erinnert uns ja nur daran, daß ein Hirn eine biologische Maschine ist; seine Biologie ist von Belang. Auch wenn von seiten der Künstlichen Intelligenz oft das Gegenteil behauptet wurde: Es ist für den Geist nicht irrelevant, in menschlichen Hirnen realisiert zu sein.

Aus unserer ersten Prämisse könnten wir nun auch noch eine dritte Schlußfolgerung herleiten.

SCHLUSSFOLGERUNG 3
Was auch sonst noch einen Geist verursachen mag, müßte Kausalkräfte besitzen, die wenigstens denen des Hirns gleichkommen.

Diese dritte Schlußfolgerung ist eine triviale Konsequenz aus unserer ersten Prämisse. Das ist ein bißchen so, wie wenn man

sagt: Falls mein benzinbetriebenes Auto 120 km/h schafft, dann müßte jedes Dieselauto, das dies kann, eine Leistung bringen, die der meines Benzin-Autos wenigstens gleichkommt. – Natürlich könnte irgendein anderes System geistige Vorgänge bewirken und dabei völlig andere chemische oder biochemische Eigenschaften verwenden als das Hirn. Es könnte sich herausstellen, daß es auf andern Planeten oder in andern Sonnensystemen Lebewesen mit Geisteszuständen gibt, die eine ganz andere Biochemie haben als wir. Angenommen, Marsmenschen kämen zur Erde, und wir gelangten zu dem Ergebnis, daß sie Geisteszustände haben. Es stellte sich dann allerdings heraus, daß sie in ihren Köpfen nur grünen Schleim haben. Trotzdem müßte der grüne Schleim, falls er in ihnen tatsächlich Bewußtsein und das übrige geistige Leben bewirkt, Kausalkräfte besitzen, die denen des menschlichen Hirnes gleichkommen. Doch nun ergibt sich aus unserer ersten Schlußfolgerung (daß Programme nicht ausreichen) und unserer dritten Schlußfolgerung (daß jedes andere System dem Hirn äquivalente Kausalkräfte haben müßte) unmittelbar:

SCHLUSSFOLGERUNG 4
Bei jedem Artefakt mit Geisteszuständen, die denen eines Menschen gleichkommen, würde es nicht allein ausreichen, daß es ein Computerprogramm ausführt. Vielmehr müßte das Artefakt über Kräfte verfügen, die denen des menschlichen Hirnes gleichkommen.

Es ist, wie ich glaube, das Ergebnis dieser Erörterung, daß wir an etwas erinnert werden, was wir die ganze Zeit über schon wußten: nämlich daß Geisteszustände biologische Phänomene sind. Bewußtsein, Intentionalität, Subjektivität und geistige Verursachung gehören allesamt zu unserer biologischen Lebensgeschichte, genau wie Wachstum, Fortpflanzung, die Absonderung von Galle und die Verdauung.

III
Kognitive Wissenschaft

Wir sagen mit völliger Selbstverständlichkeit so etwas wie »Basil hat für die Konservativen gestimmt, weil es ihm gefallen hat, wie Frau Thatcher mit der Falkland-Krise zu Rande gekommen ist«. Aber wir haben keine Ahnung, was wir mit einer Äußerung wie der folgenden anfangen sollten: »Basil hat wegen seiner Hypothalamus-Beschwerden für die Konservativen gestimmt«. D. h. wir haben Alltagserklärungen für das Verhalten von Menschen, in denen wir auf Geistiges zurückgreifen, auf ihre Sehnsüchte, Wünsche, Hoffnungen und so weiter. Und wir nehmen an, daß es außerdem neurophysiologische Erklärungen für das Verhalten von Menschen geben muß, mit denen das Verhalten durch Vorgänge im Hirn erklärt wird. Die Schwierigkeit ist: Erklärungen der ersten Art funktionieren zwar gut in der Praxis, sind aber nicht wissenschaftlich; hingegen sind Erklärungen der zweiten Art zwar gewiß wissenschaftlich, aber wir haben keine Ahnung, was wir tun müssen, damit sie in der Praxis funktionieren.
Es scheint mithin so, als gäbe es hier eine Lücke – eine Lücke zwischen dem Hirn und dem Geist. Und einige der bedeutendsten intellektuellen Bemühungen des 20. Jahrhunderts waren Versuche, diese Lücke zu füllen, eine Wissenschaft des menschlichen Verhaltens zu entwickeln, die nicht einfach die Allerweltspsychologie des gesunden Menschenverstandes, aber auch keine Neurophysiologie wäre. Bis heute sind alle diese Bemühungen, die Lücke zu schließen, fehlgeschlagen. Der Behaviourismus war der spektakulärste Fehlschlag, aber ich habe miterlebt, wie überzogene Ansprüche auch für eine Reihe von andern Ansätzen geltend gemacht und schließlich nicht eingelöst worden sind; dazu gehören die Spieltheorie, die Kybernetik, die Informationstheorie, der Strukturalismus, die Soziobiologie und noch ein paar andere. Ich werde – um ein bißchen vorwegzugreifen – die Behauptung vertreten, daß all diese Bemühungen, die Lücke zu schließen, fehlschlagen, weil es gar keine Lücke gibt, die da zu schließen wäre.
Die neuesten Bemühungen dieser Art beruhen auf Analogien zwischen Menschen und digitalen Computern. Gemäß der ex-