Was ist Cognitive Science?

Worum geht es also in der Cognitive Science? Die Cognitive Science ist eine theoretische Disziplin. Im Unterschied zu ihren beiden Nachbardisziplinen, der Kognitionspsychologie, die kognitive Prozesse unter empirischen Bedingungen untersucht, und der Künstlichen- Intelligenz-Forschung als einer technischen Disziplin geht es der Cognitive Science um eine rein abstrakte Beschreibung von Kognition.
Bei der Einführung der CS hat man sich denn auch überlegt statt CS den Namen „Intellektik" zu verwenden einfach weil man eben sagt es geht um eine allgemeine Lehre von der Intelligenz (was sich aber nicht durchgesetzt hat)
Dahinter steht allerdings eine ganz massive inhaltliche Grundannahme über Kognition. Kognitive Prozesse seien, so meint jedenfalls die Cognitive Science, gleichzusetzen mit *formalen* Beschreibungen von kognitiven Prozessen. Die formale Beschreibung eines kognitiven Prozesses sei nicht nur ein Modell zur Erklärung wie Kognition funktioniert, sondern letztere sei von Hause aus von formaler Natur. Was dies genauer bedeutet, darauf komme ich später zurück.
Kognitive Prozesse seien daher auch unabhängig von ihrer jeweiligen physikalischen Realisierung erfassbar. Aus dieser Annahme folgt auch, dass Intelligenz auf beliebigen Hardwareträgern sich realisieren lassen müsste. Es ist diese Annahme, die erst die theoretischen Rahmenbedingungen schafft, um eine künstliche Intelligenz technisch erzeugen zu können.
Denn nur wenn es möglich ist die physikalische Realisierung auszuklammern hab ich die Möglichkeit jene kognitive Prozesse die ja zunächst an die biochemischen Prozesse in unserem Kopf gebunden sind auch auf einem ganz anderen Hardwareträger zum Laufen zu bringen.
Im Extremfall müsste es möglich sein durch eine detaillierte formale Beschreibung der Gedanken eines Menschen diese Gedanken so „abschöpfen" zu können, dass, selbst wenn sein Körper eines Tages nicht mehr vorhanden ist, seine Gedanken in einem Computer weiterleben. Denn was diese Gedanken essentiell ausmache, wäre ja nicht die „Neuronenhardware" auf der sie rein per Zufall ablaufen, sondern ihre formale Beschreibung. Das ist die Grundbotschaft sozusagen die Vision die dahinter steckt.

Worauf stützt sich die CS in einem solchen Anspruch und was ist daran zu kritisieren?

Bevor wir uns gleich auf eine Kritik an dieser Annahme einlassen, gilt es zuerst einmal die positiven Motive für eine solche Betrachtung von Kognition herauszuarbeiten. Und tatsächlich entsprang die Cognitive Science einer sachlichen Notwendigkeit. Aus der gleichen sachlichen Notwendigkeit entsprang im Übrigen das, was uns allen unter dem Titel „kognitive Wende in der Psychologie" geläufig ist.

Worum geht es in der kognitiven Wende in der Psychologie? Sie ist entstanden aus einem Erklärungsnotstand des Behaviorismus, der sich bei der Beschreibung intelligenten Verhaltens ausschließlich am Schema Reiz und Reaktion orientiert und dabei innere Verarbeitungsprozesse ausklammern möchte. Als Reaktion auf eine derart einseitige Sichtweise geht es in der kognitiven Wende um eine Rehabilitierung und Wiedereinführung der Sprache des Geistes in den wissenschaftlichen Diskurs. Es geht um die Suche nach einem geeigneten Modell

für die Beschreibung der inneren Verarbeitungsprozesse im Kopf eines Kognizierenden. Als besonders brauchbar hat sich hierbei die so genannte komputationale Theorie des Geistes erwiesen. Das Verhältnis von kognitiven Prozessen zu ihren physikalischen Trägerprozessen wird dabei in Analogie zum Verhältnis von Software und Hardware eines Computers bestimmt. Kognitive Prozesse seien, grob gesprochen, die „Software" des Gehirns. Bei dieser notwendigen Korrektur des Behaviorismus darf man allerdings nicht übersehen, dass letzterer seinerseits aus einem Erklärungsnotstand einer ganz bestimmten Denkströmung Anfang des 19 JHs hervorgegangen ist, nämlich des so genannten alten Mentalismus der Würzburger Schule in der Psychologie.

Will man die Hintergründe und sachlichen Notwendigkeiten der komputationalen Theorie des Geistes verstehen, so ist es unbedingt notwendig zuerst einmal zu zeigen, inwiefern sie sich gegenüber dem alten Mentalismus *und* dem Behaviorismus als echte Alternative, ja nachgerade als eine Revolution der Denkart herausstellt.

Dass diese kognitive Wende ihrerseits keine wirkliche Lösung der anstehenden Probleme gebracht hat, ist eine andere Angelegenheit und kann erst in einem weiteren Schritt behandelt werden. Zunächst einmal ist es wichtig zu verstehen, welche Motive bei der Herausbildung der zentralen Forschungsparadigmen der Cognitive Science im Vordergrund standen. Und auf diese Motive bzw. historischen Hintergründe gehe ich jetzt kurz ein.

Der alte Mentalismus

Die Würzburger Schule wurde um 1900 von Oswald Külpe begründet. Eine seiner Grundannahmen war, dass psychische Phänomene qualitativ verschieden sind von physikalischen Phänomenen und daher nicht auf diese reduziert werden können.
Da psychische Phänomene nicht übersetzbar sind in das von außen beobachtbare Verhalten, ist der einzige Zugang zu ihnen die Introspektion. Für die Beschreibung psychischer Phänomene hat die Würzburger Schule daher auch ausgewählte Versuchspersonen verwendet, die als hinreichend geschult angesehen wurden, um auf kompetente Art und Weise ihre eigenen psychischen Abläufe beschreiben zu können. Speziell geschulte Experten sollten genau angeben, wie sie beispielsweise eine bestimmte intellektuelle Aufgabe bewältigen.
Das Problem dieses Ansatzes war freilich, dass er keine Objektivität für sich in Anspruch nehmen konnte.
Dahinter stand im Grunde genommen das Weltbild des cartesianischen Rationalismus, der die Welt in zwei Seinsbereiche aufspaltete, in den Bereich des Geistigen, die so genannte res cogitans und in den Bereich des Körperlichen, die res extensa. Das Hauptproblem einer solchen dualistischen Sichtweise war das Problem der Interaktion.
Wie kann nämlich etwas psychisches, eine bestimmte Überlegung, ein Gedanke, überhaupt auf etwas physikalisches, eine Körperbewegung als Folge einer solchen Überlegung einwirken, wenn Gedanke und Körperbewegung zwei qualitativ getrennten Seinsbereichen angehören?

In der Geschichte der Philosophie finden sich dazu teilweise obstruse Lösungsvorschläge, so z.B. das Uhrengleichnis von Geulincx. Dieser hatte angenommen, dass Geistiges und Körperliches zwar nicht direkt miteinander interagieren, dass sie aber dank eines göttlichen Ratschlusses vom Anbeginn aller Zeiten an immer parallel ablaufen wie zwei Uhren, die aufeinander abgestimmt sind.

Der Behaviorismus

Eine ernsthafte und für die psychologische Forschung brauchbare Antwort auf eine solche dualistische Weltsicht kam dann allerdings erst mit dem Behaviorismus. Der Behaviorismus war ein Versuch wissenschaftlich objektive Methoden in der Psychologie zu etablieren. Zu unterscheiden sind der klassische Behaviorismus in der Psychologie, wie er beispielsweise um 1950 von Skinner vertreten wurde und der logische Behaviorismus in der analytischen Sprachphilosophie.
Grundsätzlich geht es darum menschliches Verhalten zu erklären im Schema von Reiz und Reaktion. Betrachten wir dazu ein ganz einfaches Beispiel: Jemand hat die Absicht die Hand zu heben und führt danach eine ganz bestimmte Körperbewegung aus. Statt nun, wie dies der Dualismus tut, eine mentale Ursache für das Handheben zu unterstellen, beschreibt der Behaviorismus die Handbewegung als Reaktion auf einen Reiz. Einfaches Beispiel: Ich wurde von der Sonne geblendet und als Reaktion darauf hebe ich die Hand, um mich gegen das Blenden zu schützen.
Enthält der klassische Behaviorismus eine Anleitung zu empirischem Forschen, so geht es dem logischen Behaviorismus um eine Analyse von Sätzen. Gemäß dem logischen Behaviorismus lassen sich mentale Prädikate ohne Informationsverlust übersetzen in Verhaltensdispositionen. Wenn jemand beispielsweise sagt ich bin traurig, so müsse dieser Satz durch ganz bestimmte Sätze ersetzt werden können, die Verhaltensdispositionen beschreiben. Dahinter steht die Skepsis gegenüber einer Verdinglichung mentaler Phänomene, wie dies bei Descartes, aber auch noch in der Würzburger Schule der Fall war. So warnt Gilbert Ryle in seinem Buch der Begriff des Geistes vor der Annahme eines Gespenstes in der Maschine, vor einer künstlichen Verdoppelung der Wirklichkeit.
Er nennt die Vorstellung von Descartes, derzufolge der Geist vom Körper ontologisch zu unterscheiden sei den Mythos vom „Gespenst in der Maschine“ und warnt vor einer Verdoppelung der Wirklichkeit. Der Grund für die Auffassung, geistige Vorgänge seien Ursachen von Körperbewegungen, läge an einer Kategorienverwechslung. Geistige Fähigkeiten seien keine okkulten Vorgänge, sondern Ursachen anderer Art als Körperbewegungen. Aussagen über den Geist gehören zu einem anderen logischen Typ als Aussagen über physikalische Vorgänge. Stattdessen sind Aussagen über geistige Fähigkeiten Aussagen über das Benehmen, es handelt sich um Beschreibungen von Verhaltensdispositionen. Und so wenig wie es Sinn macht, die Universität als ein zusätzliches Gebäude am Campus zu vermuten, ergäbe es wenig Sinn, würde man geistige Vorgänge als Angehörige einer okkulten Schattenwelt hinter den physikalischen Vorgängen interpretieren. Ein solcher Irrglaube entsteht erst dann, wenn man Geist und Körper als Ausdrücke der gleichen Kategorie behandelt. Ganz ähnlich hat auch Wittgenstein in den Philosophischen Untersuchungen argumentiert: „Dort wo wir einen Körper vermuten und keinen finden, dort sagen wir, sei ein Geist.“

Das Mehrebenen Modell der Intelligenz

Die Vorstellung, man könne sich bei der Erklärung menschlichen Verhaltens allein auf eine Beschreibung des von außen beobachtbare Verhalten beschränken, hat sich dann aber als fataler Fehler und Sackgasse herausgestellt. Gescheitert ist der Behavioismus hauptsächlich auf Grund von zwei Annahmen: Um das Verhalten eines Systems angemessen verstehen zu können, müssen wir dessen innere Zustände kennen. Und nicht alles ist erlernbar, schließlich gibt es auch ein angeborenes Vorwissen.
Der Behaviorismus, so wird jedenfalls von Vertretern der so genannte Folk psychology argumentiert, stehe im Widerspruch zu unserer Alltagspsychologie. In der Alltagspsychologie verwenden wir ständig psychologische Zuschreibungen, um das Verhalten eines Menschen erklären zu können. Wir unterstellen unserem Gegenüber intentionale Einstellungen mit dem Ziel, daraus auf sein Verhalten schließen zu können.
Die zentrale Frage ist nur, wie wir die Black box der Behavioristen öffnen können, ohne dabei das naturwissenschaftliche Weltbild zu verletzen.
In den Naturwissenschaften gehen wir von einem physikalisch kausal geschlosseneen Weltbild aus. Doch was bedeutet das? Betrachten wir dazu das folgende Beispiel: Es kommt zu einem Gewitter irgendwo schlägt der Blitz ein durch den Einschlag des Blitzes fängt ein Baum zu brennen an der brenndende Baum steht neben eienem haus der breendene Baum führt dazu dass auch das Haus angezündet wird usw. eine ganze Kette haben wir da von Ursache und Wirkung und die Idee im naturwissenschaftlichen Weltbild warum es auch physikalisch kausal geschlossen ist, dass in dieser ganzen Kette von Ereiegnissen immer nur phyiskalische Ursachen vorkommen dürfen es darf nicht sein dass in dieser kette ein Glied ist das physilkalisch nicht erklärbar ist. Die einzigen Ursachen die nach der Naturwissenschaft legitimerweise angenommen werden können sind physikalische Ursachen.
Wenn man das jetzt mit dem Leib Seele Dualsimus vergleicht so sieht man gleich wo das Problem liegt nämlich im Leibseele Dualismus gibt es ausser physikalischen Ursachen sehr wohl noch etwewas ansderes nämlich die ganze Ebene des Kognitiven des Emotionalen des Psychischen das ja als Ursache des Physikalischen angesheehn wird und damit wiedeerspricht der Dualsimus diesem naturwissenwaschaftlichen Weölt, das eben davon ausgeht dass alles physikalisch kausal geschlossen ist.
Nun der Behaviorismus wiederum hat dieses Problem nicht. der Behaviorimus verwendet zur Erklärung des Verhaltens nur Beobachtungsdaten insofern kann der Behaviorsimus mit Recht von sich in Anspruch nehmen dass er den Grundprinzipien der Naturwissenschaft entspricht.
Das Problem des Behaviorimus wiederum aber ist dass er zwar den Grundprinzipen der Naturwissenschaft entspricht dafür aber einen hohen Preis bezahlt und welcher Preis ist das?
Na der Preis ist ganz einfach dass er unser Verhalten nicht erklären kann.
Wenn wir uns an die Alltagapsychologie halten dann dürfen wir nicht wie dies der Behaviorimus tut psychische Phänomene als Ursachen für physikalische Phänomene von vornherein ausschalten
Nur wie können wir das tun? Wie ist es möglich psychische Phänomene zur Erklärung physikalischer Phänomene zu verwenden ohne dabei das naturwissenschaftliche Weltbild zu verletzen? Wie können wie entgegen dem Behaviorismus psychische Phänomene als Ursachen für physikalische Phänomene zulassen und zugleich aber die Fehler des Dualismus vermei-

den? Wie können wir die Black box öffnen ohne wiederum einen inneren Homunkulus anzunehmen?
Das ist im Grunde die Gretchenfrage. Ein solches Ansinnen erinnert fürs erste betrachtet an die Quadratur des Kreises.
Eine Antwort auf eine solche Frage kam von der Cognitive Science, u.zw. von einem so genannten Mehrebenenmodell von der Intelligenz. Psychische Zustände stehen nach diesem Modell zu physikalischen Zuständen in einem ähnlichen Verhältnis wie die Software eines Computers zu seiner Hardware. Dahinter steht die komputationale Theorie des Geistes (computational theory of mind).
Worum geht es bei diesem Mehrebenenmodell? Ein kurzer Blick auf die Architektur eines modernen Allzweckrechners kann dies verdeutlichen. Eine fundamentale Eigenschaft eines solchen Rechners ist: Wir können im Befehlssatz einer Maschine eine andere, eine virtuelle Maschine programmieren.

DOCH WAS HEIßT DAS???

Nehmen wir ein ganz einfaches Beispiel: In vielen statistischen Verfahren wird der Mittelwert als Basisbaustein für weitere Berechnungen benötigt. Jedes Mal, wenn Sie aber den Mittelwert benötigen, durchläuft ein Programm etwa die folgenden Anweisungsschritte:

```
        Sum = 0
        I = 0
5       LESE ZAHL
        WENN ZAHL = 999 GEHE NACH 6
        SUM = SUM + ZAHL
        I = I + 1
        GEHE NACH 5
6       MW = SUM / I
        DRUCKE MW
        STOP
```

Anmerkung: GEHE NACH 5 oder 6 sind Sprungbefehle und 999 markiert das Ende einer Zahlenreihe.

Statt nun aber in jedem Programm, das den Mittelwert benötigt, immer wieder die gleichen Anweisungen einzubauen, können wir auch ein kleines Programm – ein so genanntes Unterprogramm – schreiben und jedes Mal, wenn wir den Mittelwert benötigen, dieses Unterprogramm aus einer Programmbibliothek aufrufen. Auf diese Weise können wir ganz neue Befehle in einer Programmiersprache erzeugen, ja wir können in einer Programmiersprache sogar eine andere Programmiersprache programmiere. Man nennt dies auch „emulieren“.

(auf diese Weise entstehen verschiedene Schichten der Programmierung)

Dieses Prinzip erklärt, wie es möglich ist, in einer Maschinensprache eine höhere Programmiersprache zu erzeugen und in dieser höheren Programmiersprache wiederum ein Programm

zu schreiben, das Eingaben und Ausgaben enthält. Höhere Schichten der Programmierung (denken Sie nur an das uns allen bekannte SPSS) sind auf diese Weise in tieferen Schichten (Assemblersprachen, Maschinensprachen usw.) realisiert.
Wichtig dabei ist, dass höhere Schichten der Programmierung in tiefere Schichten *übersetzbar* sind. Höhere Schichten der Programmierung sind also nichts ontologisch eigenständiges, sie werden nur aus Gründen der Bequemlichkeit und besseren Übersichtlichkeit verwendet. Höhere Schichten der Programmierung sind daher nichts anderes als abstraktere Beschreibungsebenen der tieferen Schichten. Ein Beispiel aus dem militärischen Bereich kann verdeutlichen, was damit gemeint ist. Gibt jemand den kurzen Befehl „Habt acht", so meint er damit „stramm stehen", „gerade ausschauen" usw. Das Kürzel „Habt acht" ist daher nichts anderes als die mit ihm verknüpften Einzelbefehle, es ist nur eine bequeme Art und Weise, um in möglichst kurzer Zeit einen Befehl geben zu können.
Dieses Schichtenmodell wird von der Cognitive Science nun auf kognitive Prozesse angewandt, u.zw. auf folgende Art und Weise. Die Cognitive Science unterscheidet drei Schichten:
Die Ebene der intentionalen Einstellungen (die semantische Ebene)
Die Ebene der formalen Beschreibungen (die syntaktische Ebene) und schließlich
Die physikalische Ebene.

Besondere Bedeutung kommt in diesem Modell der syntaktischen Ebene zu. Die Syntax hat nämlich in der Cognitive Science eine Art Scharnierfunktion. Ohne die Zwischenschicht der Syntax kann die intentionale Beschreibungsebene nämlich nicht auf die physikalischen Trägerprozesse übertragen werden. Die Semantik wird zuerst formalisiert und die formalisierte Semantik mechanisiert. Die Formalisierung ist also die notwendige Vorbedingung und Vorstufe zur Mechanisierung.

Um nun zu verstehen, welche Alternative sich durch die komputationale Theorie des Geistes gegenüber dem Leib-Seele-Dualimus und dem Behaviorimus aufgetan hat, muss das Verhältnis, das die Schichten zueinander haben, geklärt werden. Dieses Verhältnis läßt sich nach zwei Hinsichten bestimmen:

1) Es ist die Übersetzbarkeit der höheren Beschreibungsebenen in tiefere Beschreibungsebenen, wodurch sich die Cognitive Science vom Leib-Seele-Dualismus unterscheidet.
2) Erst dadurch, dass sich höhere Schichten unabhängig von ihrer konkreten Realisierung in tiefere Schichten beschreiben lassen, werden kognitive Zustände als relativ eigenständiger Forschungsbereich thematisierbar. Dies ist zumindestens eine zentrale These des Funktionalismus. Erst durch diese These werden die theoretischen Rahmenbedingungen für die künstliche Intelligenz und die Kognitionswissenschaft abgesteckt.

Übersetzbarkeit

Beginnen wir mit dem Problem der Übersetzbarkeit. Nur in einem schwachen Sinne können wir von einem Realismus der intentionalen Einstellungen ausgehen. Intentionale Einstellun-

gen haben keine ontologische Eigenständigkeit, wie dies noch vom Dualismus angenommen wird, sondern sind nur eine abstrakte Beschreibungsebene der zugrunde liegenden physikalischen Prozesse. Alles, was wir auf der Ebene der intentionalen Einstellungen ausdrücken können, läßt sich genau so gut, d.h. ohne Informationsverlust auch auf der physikalischen Ebene beschreiben. Denn alle Abläufe auf einer höheren Beschreibungsebene sind letzten Endes durch Abläufe auf der physikalischen Ebene determiniert. Was das bedeutet, kann man sich anschaulich am folgenden ganz einfachen Beispiel vorstellen: Sie sitzen vor einem PC und tun nichts anderes als in einem Textverarbeitungsprogramm einen Text markieren und verschieben. Jetzt stellen wir uns einen zweiten PC vor, der vom exakt gleichen Typ sei wie Ihr PC und auf dem auch die exakt gleichen physikalischen Prozesse stattfinden wie auf Ihrem PC zum Zeitpunkt als sie gerade den Text verschieben. Was glauben Sie wohl wird auf dem Bildschirm des zweiten, geklonten PC's zu beobachten sein? Natürlich das gleiche wie auch auf Ihrem Monitor: Es wird ein Text verschoben. Denn die Softwareprozesse sind durch die Hardwareprozesse festgelegt.

Nun geht es in der Cognitive Science aber um etwas ganz anderes. Um den Leib-Seele-Dualismus zu vermeiden geht es ja nicht, wie noch in dem Beispiel der Textverarbeitung um die Übersetzbarkeit eines Softwareprozesses in einen physikalischen Prozess, es geht vielmehr um die Übersetzbarkeit der intentionalen Beschreibungsebene in die physikalische Ebene, was etwas ganz anderes und auch ungleich schwierigeres ist. Dass die Software eben durch die Hardware festgelegt ist, mag ja noch einleuchten, dass aber jene Beschreibungsebene, die gewissermaßen das Herzstück kognitiver Prozesse ausmacht, nämlich das inhaltliche Denken, gleichermaßen durch die zugrunde liegenden physikalischen Prozesse determiniert ist, ist zumindestens eine offene Frage. Was tut die Cognitive Science, um dieses Problem zu lösen? Sie verwendet einen Trick. Sie übersetzt zuerst die Ebene der intentionalen Einstellungen in die syntaktische Ebene, das heißt sie formalisiert zunächst die inhaltliche Ebene unseres Denkens und übersetzt dann erst in einem zweiten Schritt diese formalisierten Gedanken in physikalische Prozesse. Das eigentliche Problem hierbei ist die Übersetzbarkeit von Semantik in Syntax. Denn wenn die Semantik einmal formalisiert wurde, das heißt in eine Art „Software des Gehirns" umgewandelt wurde, ist es kein Problem mehr, diese Software dann auch auf einer entsprechenden Hardware zu implementieren.

Wollen wir die Cognitive Science einer kritischen Überprüfung unterziehen, so ist die Formalisierbarkeit von Semantik der geeignete Ansatzpunkt.

Aber zunächst ein paar terminologische Erläuterungen: Was ist eigentlich unter Semantik zu verstehen? Semantik ist die Lehre von den Bedeutungen sprachlicher Zeichen. Was ist aber unter der Bedeutung eines sprachlichen Zeichens zu verstehen? Nehmen wir ein beliebiges Wort, z.B. das Wort „Tisch". Um zu wissen wovon das Wort „Tisch" handelt, müssen wir seine Bedeutung kennen. Und das Wort „Tisch" handelt eben von Tischen. Dies ist sein Begriffsumfang bzw. seine Extension. Die Bedeutung ist aber nicht mit der Extension zu verwechseln. So meint einmal Wittgenstein, das Wort „Nothung", die Bezeichnung für ein Schwert bei den Germanen, verliere doch nicht seine Bedeutung, wenn das Schwert zerstört wurde. Dennoch hat die Bedeutung eines Wortes etwas mit seiner Extension zu tun. Die Bedeutung eines Wortes ist jenes Vehikel, wodurch wir uns mit dem Wort auf einen ganz bestimmten Gegenstandsbereich beziehen.

Was sind nun intentionale Einstellungen? Unter Intentionalität versteht man in der Philosophie nicht nur Absichten, die ein Mensch haben kann, ein Mensch hat eine intentionale Ein-

stellung, wenn er sich im weitesten Sinne des Wortes auf etwas bezieht und sich dabei in einem ganz bestimmten psychischen Zustand befindet, wenn er also etwas glaubt, hofft, befürchtet usw. Draußen rennt ein Hund vorbei und Sie befürchten, dass der Hund Sie beißen könnte. Um sich also mit unseren Wörtern, Sätzen, Äußerungen auf etwas beziehen zu können, müssen wir im Besitz von intentionalen Einstellungen sein.
Bezeichnen wir die Fähigkeit, Bedeutungen verstehen zu können, als Semantizität, so verfügen nur solche Lebewesen über Semantizität, die auch intentionale Einstellungen haben können. Wie soll nun aber die Übersetzung dieser Semantizität in Syntax von statten gehen?
Betrachten wir dazu ein Beispiel: Ein Arzt untersucht einen Patienten, diagnostiziert eine bakteriologische Infektionskrankheit und verordnet daruhin ein bestimmtest Antibiotikum. Die Diagnose des Arztes beruht auf Grund ganz bestimmter inhaltlicher Überlegungen. Der Arzt muß sowohl *über* Krankheiten als auch *über* deren Therapie Bescheid wissen. Dies setzt aber ein *semantisches* und eben nicht nur ein formales Wissen voraus.
Um nun aber zu verstehen, wie die Übersetzbarkeit von semantischem Wissen in die syntaktische Verarbeitung von Symbolen zumindestens im Prinzip funktiuonieren könnte, dazu hat die Cognitive Science ein ganz bestimmtes Vorbild, nämlich die so genannte Logische Beweistheorie.
In der logischen Beweistheorie wird nur ein ganz bestimmtes Wissen formalisiert, nämlich mathematisches Wissen.
Auch ein Mathematiker verwendet semantisches Wissen, wenn er eine wahre Aussage machen will. Denken wir an den Satz „7 ist eine Primzahl“. Dass diese Aussage wahr ist, dazu muß unser Mathematiker einiges über Zahlen wissen. Unter anderem muß er wissen, dass eine Primzahl eine Zahl ist, die nur durch sich selber und durch 1 teilbar ist. Er muß natürlich auch wissen, was teilbarkeit *bedeutet* usw.
In der logischen Beweistheorie wird nun der Versuch unternommen, jedem wahren Satz in der Mathematik einen syntaktisch korrekt abgeleiteten Satz in einem formalen System zuzuordnen. Haben wir also einen syntaktisch korrekt abgeleiteten Satz in einem formalen System, so entspricht diesem Satz, wenn wir ihn interpretieren, eine wahre Aussage in der Mathematik.
Jerry Fodor, ein bekannter Kognitionswissenschaftler drückt dies so aus: „(..) certain of the semantic relations among symbols can be, as it were, ‚mimicked’ by their syntactical relations; that, when seen from a greater distance, is what proof-theory is about.“ (Fodor, 1987, 19)
So hat beispielsweise Bertrand Russell gemeinsam mit Alfred Whithehead in den Principia Mathematica den Versuch unternommen, sämtliche mathematische Wahrheiten durch eine Menge von Axiomen und Schlußregeln rein formal zu rekonstruieren.
Um zumindestens im Prinzip nachvollziehen zu können, wie dies funktioniern kann, dazu müssen wir eine ganze Menge mehr über formale Systeme wissen. Darauf komme ich etwas später zurück.
Betrachten wir zunächst die folgende Geschichte: Srinivasa Aiyangar Ramanujan war ein hochbegabter indischer Mathematiker, der schon sehr früh als „Wunderkind“ gegolten hat und bereits im frühen Alter neue Theoreme in der Mathematik entdeckte. Zu den neu entdeckten Theoremen fehlten allerdings die Beweise. Auf Einladung des Trinity College kam er nach England.
Sein Vorstellungsbrief, dem dann die Einladung folgte, begann mit den Worten:

Sehr geehrter Herr,
ich bitte darum mich Ihnen vorstellen zu dürfen als Angestellter der Buchhaltung in der Hafenverwaltung von Madras mit einem Jahreseinkommen von £ 20. Ich bin jetzt 26 Jahre alt. Ich habe keine abgeschlossene Universitätsausbildung, habe aber den üblichen Unterricht absolviert. […] Ich bitte Sie, die beigelegten Papiere durchzusehen. Da ich arm bin, möchte ich gerne meine Sätze veröffentlichen, falls Sie überzeugt sind, dass sie einen Wert haben.

Da die Beweise fehlten, wurde er gefragt, wie er denn diese neuen Theoreme entdeckt habe. Seine Antwort war: Die Göttin Namagiri habe ihm dies im Traum eingegeben (vgl. Hofstadter, 1985, 201). Eine Gleichung hat für mich keinen Sinn, es sei denn, sie drückt einen Gedanken Gottes aus..

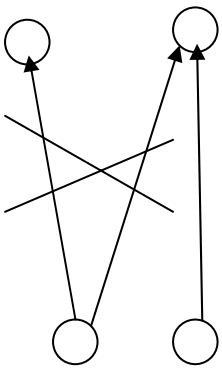
Dies ist ein Beispiel für Intuition in der Mathematik. Für die logische Beweistheorie ist diese Vorgehensweise eine harte Nuß. Denn das Ziel ist ja, jede wahre mathematische Aussage in einem formalen System so rekonstruieren zu können, dass sie aus diesem allein unter Nutzung der Axiome und der Schlußregeln deduziert werden kann. Was auf diese Weise formalisiert wurde, kann in einem weiteren Schritt auch von einem Computerprogramm bewältigt werden. Die Idee, die hinter der Beweistheorie steckt, ist also, grob gesprochen, die: Wir wollen beispielsweise wissen, ob 7 eine primzahl ist oder nicht. Dazu müssen wir nur ein kleines Programm schreiben, das am Monitor dann die Aussage ausgibt: „7 ist eine Primzahl“. Dies kann aber nur für solche Aussagen funktionieren, für die wir auch die nötigen Beweise haben. Ramanujans Theoreme gehören nicht dazu.
Und die zentrale Frage ist, ob es zu jedem wahren Satz in der Mathematik auch ein Äquivalent gibt in der logischen Beweistheorie, nämlich einen syntaktisch korrekt abgeleiteten Satz.
Nur wenn dies der Fall ist, können wir sagen, dass „semantic relations can be mimicked by syntactic relations“. Wir könnten auf diese Weise für jeden Satz in der Mathematik überprüfen, ob er wahr oder falsch ist. Wahr ist er nur dann, wenn er ein syntaktisch korrektes Äquivalent in der logischen Beweistheorie hat. Falsch wäre er dann, wenn ein solches Äquivalent fehlt.
Es geht mir im Augenblick nicht darum, dieses Unternehmen der Cognitive Science zu kritisieren, es geht mir vielmehr darum, seinen Anspruch klar herauszuarbeiten.
Fodors Behauptung „semantic relations can be mimicked by syntactic relations“ läßt sich auch mit anderen Worten so ausdrücken. Syntaktische Beziehungen *spiegeln* semantische Beziehungen. Diese Unterstellung wird in der Cognitive Science auch als so genanntes Formalistenmotto bezeichnet: syntax mirrors semantics.
Das Formalistenmotto ist eine moderne Neuerzählung von Geulincx Uhrengleichnis, denn auch die semantische und die syntaktische Ebene sollen ja paralell laufen. Der Unterschied zu Geulincx ist freilich der, dass aufgrund der Übersetzbarkeit von Semantik in Syntax die semantische Ebene keine ontologische Eigenständigkeit hat.

Multiple Instanziierung

Nun zum zweiten Punkt: Warum sind höhere Schichten relativ unabhängig von den niedereren Schichten? Wir haben doch soeben unter dem Punkt „Übersetzbarkeit“ festgestellt, dass

letzten Endes alle höheren Schichten durch die zugrundeliegende physikalische Schicht determiniert sind. Das heißt doch, dass letzten Endes alles auf die materielle Ebene reduziert werden kann. Was soll dann aber noch die relative Eigenständigkeit höherer Schichten bedeuten?
Die Antwort darauf lautet: Betrachten wir das Verhältnis der Schichten von unten nach oben, so sind Ereignisse auf einer höheren Beschreibungsebene durch Ereignisse auf einer niedereren Ebene zwar genau festgelegt (ein Ereignis auf einer niederen Ebene ist *genau einem* Ereignis auf einer höheren Ebene zuordenbar), das umgekehrte ist jedoch nicht der Fall: Ein Ereignis auf einer höheren Ebene kann durch verschiedene Ereignisse auf einer tieferen Ebene realisiert werden. Dies ist das Prinzip der multiplen Instanziierung. Man kann sich das Verhältnis der Schichten zueinander grafisch auch wie folgt vorstellen. Erlaubt ist das folgende (die runden Kreise symbolisieren Ereignisse):

Verboten ist dahingegen (da von unten nach oben betrachtet immer nur je ein Pfeil zu einem Ereignis führt):



Bei der Beschreibung eines Ereignisses auf einer höheren Ebene spielt es daher im Grunde genommen auch gar keine Rolle, wie dieses Ereignis auf einer tieferen Ebene realisiert ist. Ob also mentale Ereignisse in einer der unseren äquivalenten „brainware“ realisiert sind oder etwa gar in gasförmigen Lebensformen eines fernen Planeten, muß bei der Beschreibung der mentalen Ereignisse nicht berücksichtigt werden. Der Funktionalismus muß daher auch überhaupt keine Annahmen machen über die ontologische Natur der physikalischen Prozesse, in denen die kognitiven Abläufe realisiert sind.

Die Sprache formaler Systeme

Wie bereits gesagt wurde, hat die Formalisierung der Semantik eine wichtige Mittelfunktion in der Cognitive Science. Denn die Formalisierung des inhaltlichen Denkens ist die Vorbedingung für die sich unmittelbar daran anschließende Mechanisierung der formalisierten Gedanken. Man suche, grob gesprochen, die „Software“ in unserem Kopf, denn, haben wir einmal die richtige Software gefunden, so ist es kein großes Problem mehr, diese Software auf einem geeigneten Hardwareträger zu implementieren. Ein Computer ist, so betrachtet, die Instanziierung eines formalen Systems.
Was ist aber unter einem formalen System zu verstehen? Es handelt sich hierbei um eine Sprache, die aus vorgegebenen Zeichenketten nach bestimmten Regeln weitere Zeichenketten erzeugt. Die Produktionsregeln zur Erzeugung neuer Zeichenketten bedienen sich ausschließlich der äußeren Gestalt der Zeichenketten ohne Rücksicht darauf, was diese bedeuten könnten (falls sie überhaupt etwas bedeuten).
Damit in diesem Zusammenhang überhaupt von einer Sprache die Rede sein kann, benötigen wir dreierlei: Erstens ein Alphabet, also einen ganz bestimmten Zeichenvorrat, zweitens eine Grammatik, die uns sagt, wie wir die Buchstaben des Alphabets zu sinnvollen Wörtern und Sätzen zusammensetzen können und schließlich und drittens einer Entscheidungshilfe, die uns sagt, welche der grammatikalisch sinnvoll gebildeten Sätze in dem System auch gültige Sätze sind.

Das Alphabet des Computers: der binäre Code

Der binäre Code hat einen auf zwei Zeichen eingeschränkten eingeschränkten Zeichenvorrat. Wir stellen uns diese zwei Zeichen gewöhnlich als die Ziffern 0 und 1 vor, was freilich, wie wir gleich sehen werden, nicht zwangsläufig so sein muß. Nicht jedes formale System hat einen auf zwei Zeichen beschränkten Zeichenvorrat. Ich verwende den binären Code als prototypisches Beispiel für die in formalen Systemen verwendeten Zeichen hautsächlich aus dem folgenden Grund. Ein Computer ist, wie ich ja schon erwähnt habe, die Instanziierung eines formalen Systems, so dass sich auch die Eigenschaften des binären Codes auf alle in formalen Systemen verwendeten Zeichenketten übertragen lassen.
Ich möchte den binären Code in drei Hinsichten betrachten: 1) Es handelt sich hierbei um eine *strukturelle* Abbildung und 2) handelt es sich um einen rein *quantitativen* Informationsbegriff und schließlich 3) hat der Code die Eigenschaft *transitiv* zu sein.

Was ist eine strukturelle Abbildung?

Der Ausdruck „Code“ kommt aus der Nachrichtentechnik und meint eine Verschlüsselungsvorschrift für die Übertragung von Informationen. Man kann eine solche Vorschrift im weitesten Sinne des Wortes als eine Abbildung interpretieren. Doch welche Abbildung ist im Zusammenhang mit dem binären Code gemeint? Und inwiefern soll es sich hier um eine strukturelle Abbildung handeln?
Betrachten zunächst den folgenden ganz einfachen Fall: Sie schauen aus dem Fenster und sehen einen Berg. Sie fertigen eine mehr oder weniger naturalistische Zeichnung dieses Berges auf einem Stück Papier an und reichen den Zettel an eine Studentin, die vielleicht nicht in unmittelbarer Nähe des Fensters sitzt. Sie wollen Ihrer Kollegin mitteilen, dass Sie gerade einen Berg sehen. Eine solche Abbildung ist allerdings keine strukturelle Abbildung, denn sie beruht auf einer gewissen naturalistischen Ähnlichkeit von Bild und Abgebildeten.
Bei einem strukturellen Abbildungsbegriff besteht zwischen den Zeichen und der zu vermittelnden Information lediglich eine rein per Konvention festgelegte Zuordnung. Aber selbst diese Einschränkung erklärt noch immer nicht hinreichend, was unter einer strukturellen Abbildung zu verstehen ist. Eine strukturelle Abbildung hat nämlich noch die folgende wichtige Eigenschaft: dass nämlich der materielle Informationsträger beliebig austauschbar ist.
Was damit gemeint ist, dazu bedarf es einer etwas längeren Ausführung. Betrachten wir einmal die Zeichenfolge „10“. Um diese Zeichenfolge überhaupt interpretieren zu können, müssen wir einen ganz bestimmten Code zu Grunde legen. Im dekadischen Zahlensystem bedeutet „10“ die Zahl zehn, im binären Code dahingegen die Zahl zwei.
Was benötigen wir nun aber, um „10“ einmal als zehn und ein ander mal als zwei zu lesen? Erst wenn wir auf diese Frage die richtige Antwort finden, so können wir auch verstehen, warum bei einer strukturellen Abbildung der materielle Informationsträger beliebig ausgetauscht werden kann.
Betrachten wir dazu ein ganz einfaches Beispiel: Nehmen wir an, Sie hätten ein Zahlenschloß vor sich, wie es bei Fahrrädern verwendet wird, und das Schloß bestünde aus nur zwei Ringen, wobei an jeweils jedem Ring die Ziffern 0 bis 9 einstellbar sind. Jeder Ring hätte also zehn verschiedene Einstellmöglichkeiten, da wir ja bei 0 zu zählen anfangen.
Um nun die Zahl Null darzustellen, gehen wir wie folgt vor: Den linken Zahlenring belassen wir auf 0 und den rechten Zahlenring drehen wir ebenfalls auf 0. Die Zahl Eins wiederum stellen wir so dar: Den linken Ring belassen wir auf 0 und den rechten Ring drehen wir auf auf 1. Dies ist allerdings nur eine Vereinbarung, also eine Konvention, denn wir könnten genauso gut in umgekehrter Reihenfolge verfahren (also zur Darstellung der Zahl eins den linken Ring auf 1 stellen und den rechten Ring auf 0).
Nach dem gleichen Schema können wir bis einschließlich der Zahl neun verfahren. Nach neun wird’s spannend: Wie stellen wir die Zahl zehn dar? Versuchen wir nach der gleichen Methode vorzugehen wie bisher, das heißt belassen wir den linken Ring auf 0 und versuchen den rechten Ring so lange zu drehen, bis wir zur Ziffer zehn kommen, so können wir drehen, so lange wir wollen. Die Ziffer zehn werden wir am rechten Ring niemals finden!
Was ist also im dekadischen Zahlensystem zu tun? Man macht einen Übertrag: Man dreht den linken Ring auf 1 und den rechten auf 0. Sobald man also zu einer Zahl kommt, die auf dem rechten Ring nicht als Ziffer dargestellt werden kann, verwendet man eben eine Kombination zweier Ringe.

Ab der Zahl zehn verfährt man dann auf die gleiche Weise wie wir ursprünglich begonnen haben. Um die Zahl elf darzustellen, belassen wir den linken Ring auf 1 und drehen den rechten Ring auf 1. Diese Vorgehensweise lässt sich bis zur Zahl neunzehn fortsetzen. Ab der Zahl zwanzig benötigen wir wieder einen Übertrag: Wir drehen den linken Ring auf 2 und den rechten auf 0. Mit genau zwei Ringen können wir auf diese Art und Weise die Zahlen von null bis neunundneunzig darstellen.
Wie schaut die Sache nun aber beim binären Zahlensystem aus? Am besten stellen wir uns wiederum ein Zahlenschloß vor mit zwei Ringen, wobei in diesem Falle pro Ring jedoch nur zwei Einstellmöglichkeiten vorkommen.
Überlegen wir uns dazu, wie wir beim dekadischen Zahlensystem vorgegangen sind. Zur Darstellung der Zahl null gehen wir ganz analog vor wie im dekadischen Zahlensystem. Wir belassen den linken Ring auf 0 und drehen den rechten ebenfalls auf 0. Für die Darstellung der Zahl eins gehen wir ebenfalls analog zum dekadischen System vor: linken Ring auf 0 einstellen und rechten Ring auf die Ziffer 1. Im binären Zahlensystem wird's aber bereits bei der Darstellung der Zahl zwei spannend. Versuchen wir auf dem rechten Ring die Ziffer 2 einzustellen, so bekommen wir im binären Zahlensystem ähnliche Probleme wie wir sie beim dekadischen bei der Darstellung von zehn bekamen: Nirgendwo auf dem rechten Ring lässt sich die Ziffer 2 ausmachen! Was ist zu tun? Wir verwenden die gleiche Methode wie beim dekadischen Zahlensystem, das heißt, so bald wir zu einer Zahl kommen, die auf dem rechten Ring nicht als Ziffer dargestellt werden kann, verwenden wir eine Kombination von zwei Ringen. Zur Darstellung der Zahl zwei verwenden wir also einen Übertrag. Wie stellen den linken Ring auf die Ziffer 1 und den rechten Ring auf 0. 10 bedeutet daher im binären Zahlensystem die Zahl zwei. Um die Zahl drei im binären Zahlenstystem darzustellen, belassen wir den linken Ring auf 1 und drehen den rechten ebenfalls auf 1. Mit genau zwei Ringen können wir auf diese Weise Zahlen von null bis drei darstellen.
Was bedeutet das nun aber? Woran *erkennen* wir, dass 11 beispielsweise drei bedeutet? Über welche Informationen müssen wir verfügen, um überhaupt 11 als drei lesen zu können? An der Beantwortung dieser Frage lässt sich ablesen, warum es sich hier um einen strukturellen Abbildungsbegriff handelt und warum der materielle Informationsträger beim binären Code beliebig ausgetauscht werden kann.
Es sind zwei Informationen, die wir benötigen, um 11 als drei lesen zu können: Wir müssen erstens die Menge des verfügbaren Zeichenvorrats kennen. Wir müssen also wissen, um in dem Beispiel zu bleiben, wie viele Einstellmöglichkeiten unser Zahlenschloß pro Ring hat. Denn im Falle zweier Einstellmöglichkeiten bedeutet 11 etwas ganz anderes als im Falle von zehn Einstellmöglichkeiten.
Als zweites müssen wir wissen, wie die einzelnen Stellen zu interpretieren sind. Eine zweistellige Zahl im dekadischen Zahlensystem ist im Grunde genommen ein Polynom, also eine Summe von Vielfachen von Zehnerpotenzen. So bedeutet 12 beispielsweise einmal zehn hoch eins plus zweimal zehn hoch null. Dies ist allerdings nur eine Konvention, die wir jedoch mit der Zeit bereits so verinnerlicht haben, dass wir uns ihrer überhaupt nicht mehr bewusst sind. Es könnte allerdings, worauf ich schon hingewiesen habe, genauso gut eine Konvention geben, derzufolge die Einser und die Zehner Stellen vertauscht sind. In diesem Falle würde 12 statt zwölf eben einundzwanzig bedeuten.
Warum es sich hier um einen strukturellen Abbildungsbegriff handelt, kann man sich vor dem Hintergrund dieser Überlegungen so vorstellen. Nehmen wir einmal an, wir würden statt der

Ziffern 0 und 1 die Ziffern A und B verwenden, wobei A 0 bedeuten soll und B 1. Statt beispielsweise 10 würden wir also BA schreiben. BA würde also in diesem neuen Notationssystem zwei bedeuten.
Da wir zur Interpretation von BA als die Zahl zwei nur den zur Verfügung stehenden Zeichenvorrat und die Bedeutung der einzelnen Stellen kennen müssen, spielt es auch gar keine Rolle, wie die verwendeten Zeichen materiell realisiert sind. Zwei diskret unterscheidbare Zeichen lassen sich auf die unterschiedlichste Art und Weise physikalisch realisieren, wobei der Phantasie keine Grenzen gesetzt sind: Polarität der Magnetisierung (im Falle eines Ferritkernspeichers), die Ziffern 0 und 1 oder auch Streichhölzer, die geknickt wurden oder nicht. Somit lässt sich auch auf einem Computer, in dem sich zwei physikalisch verschiedene Zustände unterscheiden lassen, dieser Binärcode realisieren.
Denken wir an ein Musikstück im MP3-Format, abgespeichert auf einer CD-ROM. Da auch hier digitale Daten gespeichert sind, haben wir auf der physikalischen Ebene eine Folge zweier diskret voneinander unterscheidbarer Zustände. Da der materielle Träger der Information austauschbar ist, könnte man auf den Gedanken kommen, das gleiche Musikstück durch eine Folge von geknickten bzw. nicht geknickten Zündhölzern zu codieren.
Natürlich wird sich der Musikgenuß bei Betrachtung der geknickten und nicht geknickten Streichhölzer in Grenzen halten.Um aus einer Folge von zwei physikalisch verschiedenen Zuständen die Musik herauszuhören, benötigen wir eine technische Vorrichtung, die die im binären Code enthaltene Information wiederum in Musik rückübersetzt.
Die Informationen, die eine solche technische Vorrichtung für die Umsetzung der digitalen Daten in hörbare Musik benötigt, sind indes nur formale Strukturen. Was beim binären Code an Infomationen ausgetauscht wird, sind Strukturen und eben nichts Materielles.
Schreiben Sie beispielsweise eine E-Mail auf Ihrem Bildschirm, so ist es ja nicht so, dass ein Postbote kommt, den Speicher aus Ihrem Computer ausbaut, um ihn dann in rasender Eile nach den USA zu transportieren und ihn dort in einen Rechner einzubauen, sondern was übertragen wird, sind nur Strukturen und gerade deshalb erscheint in Sekundenschnelle am Bildschirm Ihres Adressaten in den USA der genau gleiche Brief auf, so wie Sie ihn auf Ihrem Computer getippt hatten. Die Gleichheit, von der hier die Rede ist, ist eine Strukturgleichheit.

Ein quantitativer Informationsbegriff

Kennen Sie La Gomera? Es handelt sich um eine kleine Insel im Atlantischen Ozean. Sie gehört zur Gruppe der Kanarischen Inseln und ist trotz ihrer geringen Größe von zahlreichen tiefen Schluchten durchfurcht. Was sie für uns hier interessant macht, ist eine spezielle Pfeifsprache, genannt El Silbo, die nur auf La Gomera verwendet wird. Mit Hilfe dieser Sprache war es den Einwohnern möglich, über die stark zerklüftete Landschaft Botschaften auszutauschen.
Nehmen wir an, die einzigen Informationen, die man austauschen wollte, wären nur zwei: Gefahr bzw. keine Gefahr. In diesem Falle würden zwei verschiedene Pfiffarten (unterscheidbar durch Tonhöhe oder –länge vollkommen ausreichen.
Stellen wir uns nun vor, es ginge darum, insgesamt fünf Informationen auszutauschen. Dies ließe sich auf verschiedene Art und Weise realisieren: Entweder verwenden wir für jede der

fünf zu übertragenden Informationen eine eigene Pfiffart (beispielsweise in einer bestimmten Tonhöhe). In diesem Falle würde es genügen, einmal zu pfeifen.
Nehmen wir statt dessen aber an, dass unser „Pfiffalphabet“ aus nur zwei klar unterscheidbaren Pfiffarten bestünde (vielleicht aus der Sorge, zu viele verschiedene Pfiffarten könnten bei der Übertragung verfälscht werden), so genügt es nicht, einmal zu pfeifen, wenn wir fünf verschiedene Informationen übertragen wollen.
Fragen wir also: Wie oft müssen wir hintereinander pfeifen, wenn wir fünf Informationen übertragen wollen und wenn uns nur zwei verschiedene Pfiffarten zur Verfügung stehen? Diese Frage ist identisch mit der Frage, wie viele Zahlenringe wir für ein Zahlenschloß mit zwei Einstellmöglichkeiten benötigen, wenn wir fünf Informationen darstellen wollen. Sie ist auch identisch mit der Frage, wie viele Stellen im binären Zahlensystem wir benötigen, um die Zahl fünf darzustellen.
Diese Anzahl der Stellen lässt sich exakt nach der Formel berechnen:

$$2^x = 5$$

Womit müssen wir die Zahl 2 potenzieren, um 5 zu erhalten? Dieses x ist der Logarithmus von 5 zur Basis 2. Er gibt uns die Anzahl der Stellen an, die wir zur Darstellung einer bestimmten Anzahl von Informationen benötigen. Statt von einer Anzahl von Stellen zu sprechen, können wir auch fragen, wie viele Bits wir benötigen, um eine Information zu übertragen. Ein Bit ist die kleinste Informationseinheit im binären Code.

Quantitativ ist dieser Informationsbegriff nun insofern als die einzige Frage, die sich zur Speicherung irgendeiner Information stellt, eben die Frage ist, wie viele Bits wir zu ihrer Codierung benötigen. Jede gespeicherte Information wird hier gleich behandelt, ein Gemälde von Michelangelo, eine Sonate von Mozart, gefragt wird in jedem Falle nach der Anzahl der Bits, die ich zu ihrer binären Darstellung benötige. Die Bedeutung des jeweiligen Kunstwerks, etwa seine ästhetische Wirkung auf den Rezipienten, spielt beim quantitativen Informationsbegriff keine Rolle.

Transitivität

Dekadische und binäre Verschlüsselungen sind nur zwei Beispiele von strukturellen Abbildungen. Daneben gibt es beliebig viele andere, beispielsweise den hexadezimalen Code, dessen Zeichenvorrat aus 16 verschiedenen Zeichen besteht (16 Einstellmöglichkeiten an einem Ring unseres Zahlenschlosses).
Wegen der Transitivität des Abbildungsbegriffes gilt nun das folgende: Verschickt eine Person A eine Nachricht im hexadezimalen Code an eine Person B und übermittelt diese Person B nach der Dechiffrierung diese Nachricht im binären Code an eine dritte Person C, so erhält C die gleiche Nachricht wie im Falle einer direkten Übertragung von A nach C im binären Code.
Wegen dieser Transititivität des Codes ist es möglich, sich auf einen einheitlichen Standardcode zu einigen. Es spielt daher auch keine Rolle, welchen Code wir tatsächlich verwenden, da sich jeder Code prinzipiell in einen solchen Standardcode übersetzen lässt. Aus diesem

Grunde konnte sich der binäre Code als allgemein gültige Universalsprache des Computers etablieren.
Wir haben im Binärcode ein universelles Alphabet gefunden, mit dem jede Art von Information (da ja nur quantitativ betrachtet) ausgedrückt werden kann. Ganz unabhängig davon, ob wir etwas mit einer WebCam betrachten, ob wir einen Ton, ein Bild, einen Text, eine mathematische Formel übertragen, es gilt Total Digital (Negroponte) - alles ist übersetzbar in den digitalen Code, denn dieser ist ein struktureller Code, verwendet einen quantitativen Informationsbegriff und ist als solcher zugleich transitiv.
Der Siegeszug des digitalen Codes ist allgegenwärtig, sei es dass wir mit einem Handy, einem MP3-Player, einem digitalen Fotoapparat oder womit auch immer umgehen.

Die Gültigkeit von Sätzen

Wir haben jetzt ein Alphabet einer formalen Sprache gefunden. Was uns aber noch gänzlich fehlt, ist, wie wir aus den Buchstaben eines solchen Alphabets eine Sprache generieren können. Dazu gehört zweierlei: Zum einen eine Grammatik, die uns sagt, wie wir die Buchstaben zu sinnvollen Aussagen (so genannten wohl geformte Formeln) zusammensetzen können und welche dieser grammatikalisch sinnvollen Aussagen auch gültig sind. Für beides benötigen wir Regeln, wobei für uns hier jene Regeln von besonderer Bedeutung sind, die die Gültigkeit der Aussagen garantieren (die grammatikalischen Regeln lasse ich hier beiseite).
Diese Gewichtung hängt mit der Zielsetzung des Formalistenmottos zusammen. Syntaktische Beziehungen sollen semantische Beziehungen widerspiegeln, es geht also darum, wahre Aussagen durch gültige Aussagen in einem formalen System nachzuahmen.
Was ist unter gültigen Aussagen in einem formalen System zu verstehen?
Douglas Hofstadter hat in den achtziger Jahren des vorigen Jahrhunderts ein Buch über formale Systeme geschrieben (Gödel, Escher, Bach), das in kürzester Zeit zu einem Kultbuch avancierte. Ich bringe hier zwei Beispiele für formale Systeme aus diesem Buch: das MIU-System und das pg-System. Am Beispiel dieser beiden Systeme soll schrittweise erläutert werden, was die Gültigkeit von Sätzen in einem formalen System überhaupt bedeutet und inwiefern sich in dieser Gültigkeit auch die Wahrheit von Aussagen widerspiegeln kann.

Das MIU-System

Fragen sie nicht, was MIU bedeutet. Es handelt sich hier nur um ein Spiel, bei dem aus Zeichenketten, die uns am Beginn des Spiels zur Verfügung gestellt werden, nach ganz bestimmten Regeln weitere Zeichenketten produziert werden können. Das MIU-System ist ein uninterpretiertes formales System, d.h. die in diesem Spiel verwendeten Zeichenketten haben keine Bedeutung und sinnvoll gebildete Sätze in dem System sind weder wahr noch falsch, da sie ja von nichts handeln.
Die einzige Zeichenkette, die uns am Beginn des Spiels zur Verfügung gestellt wird, ist die Zeichenkette MI. Sie ist sozusagen unser Spielkapital. Dazu bekommen Sie vier Regeln, um, ausgehend von Ihrem Spielkapital, weitere Zeichenketten produzieren zu können.

Regel I: Besitzen wir eine Zeichenkette, an deren Ende ein I ist, so können wir am Schluß ein U anfügen.

Jetzt haben wir eine Möglichkeit, unser Spielkapital zu erweitern. Da wir ja vom Anfang an mit der Zeichenkette MI ausgestattet sind und deren letzter Buchstabe ein I ist, können wir die neue Zeichenkette MIU erzeugen.

Regel II: Angenommen wir haben ein Kette Mx, dann können wir eine weitere Kette Mxx produzieren (x ist eine beliebige Zeichenkette)

Haben wir beispielsweise unter Anwendung von Regel I MIU, erzeugt, so können wir jetzt mit Hilfe der Regel II die Zeichenkette MIUIU produzieren,

Mit Hilfe der zweiten Regel sind wir schon ganz schön flexibel. Wir können jetzt schon eine ganze Menge verschiedener Zeichenketten erzeugen. So könnten wir beispielsweise auch aus MI MII produzieren.

Regel III: Wenn in einer Kette III vorkommt (also dreimal I), so können wir diese Kette durch ein U ersetzen.

Dazu das nachstehende Beispiel: Wir haben eine Zeichenkette MIIIIU (enstanden durch zweimalige Anwendung von Regel II auf unser Startkapital MI und der anschließenden Anwendung von Regel I), so können wir daraus MUIU produzieren.

Regel IV: Kommt in einer Kette UU vor, so kann man es streichen.

Beispiel: Aus MUUUIII lässt sich die Kette MUIII erzeugen.

Unser Startkapital – die Zeichenkette MI – bezeichnen wir nun als ein *Axiom*. Bei einem Axiom steht von vornherein fest, dass es ein gültiger Satz in dem formalen System ist. Axiome sind sozusagen unsere Spielvorgaben bei der Produktion von Zeichenketten.
Die typographischen Regeln zur Erzeugung weiterer Zeichenketten sind die Ableitungsregeln in dem formalen System. Alle Zeichenketten, die sich aus den Axiomen unter Nutzung dieser Regeln erzeugen lassen, sind gleichfalls gültige Sätze in dem formalen System.
Da wir es hier mit einem *uninterpretierten* System zu tun haben, entsprechen diesen gültigen Sätzen keine wahren Aussagen (Gültigkeit meint in diesem Zusammenhang eine ausschließlich syntaktische, durch die formalen Spielregeln festgelegte Eigenschaft) und die Ableitung eines Satzes im MIU-System ist daher auch kein Beweis für den Satz (denn ein Beweis ist eine *semantische* Eigenschaft).

Das pg-System

Betrachten wir nun das folgende formale System, das auf den ersten Blick ganz analog zum MIU-System aus einer Folge von bedeutungslosen Zeichen besteht, die durch typographische Erzeugungsregeln generiert werden können. Das pg-System hat eine unendliche Anzahl von Axiomen, die sich in einer Schablone beschreiben lassen, und eine einzige Ableitungsregel.

Die *Schablone* zur Erzeugung eines Axioms lässt sich so darstellen:

xp-gx- sei eine gültige Zeichenkette, wenn x nur aus einer bestimmten Anzahl von Bindestrichen besteht (man beachte: die Variable x enthält immer die *gleiche* Anzahl von Bindestrichen).

Folgende Zeichenketten sind dieser Schablone zufolge Axiome:

-p-g--

--p-g---

Die *Regel* lautet: Angenommen, x, y und z seien eine Folge von Bindestrichen. Nehmen wir ferner an *xpygz* sei ein gültiger Satz in unserem pg-System, so sei auch *xpy-gz-* ein gültiger Satz.

Eine Anwendung dieser Regel wäre etwa die folgende: Der Satz --p-g--- ist (man beachte die Schablone) ein Axiom im pg-System und somit *per definitionem* ein gültiger Satz. Folgen wir nun der einzigen Ableitungsregel, so gilt: Die Variable x besteht aus 2 Bindestrichen, y aus einem Bindestrich und z besteht aus 3 Bindestrichen. Wenden wir nun die Regel an, so erhalten wir den folgenden gültigen Satz:

--p--g----

Dies ist das pg-System. Mit Hilfe der Axiome und der einen Ableitungsregel können wir eine ganze Menge gültiger Sätze konstruieren. Das pg-System erzeugt beim Leser auf den ersten Blick trotz oder gerade wegen seiner Einfachheit ein gewisses Unverständnis und hinterläßt den Eindruck von Umständlichkeit. Dieser Eindruck verschwindet allerdings sofort, sobald wir die Zeichenketten zu interpretieren beginnen: Der Anzahl der Bindestriche entspreche eine natürliche Zahl (- bedeute die Zahl eins, -- die Zahl zwei usw.), p *bedeute* plus und g *bedeute* gleich.
Was wir hier also vor uns haben, ist ein ganz einfaches formales System, das die Addition formalisiert. Was allerdings das pg-System in diesem Zusammenhang so interessant macht, ist, dass wir an diesem schlichten Beispiel sehr schön demonstrieren können, was die Spiegelung von Semantik in Syntax bedeutet.
Betrachten wir hierzu nur den letzten gültigen Satz, den wir aus dem System abgeleitet haben: --p--g----. Auf der *einen* Seite wissen wir, dass wir es hier mit einem gültigen Satz im pg-System zu tun haben (gültig im Sinne von formal ableitbar). Auf der *anderen* Seite bedeutet

der Satz, wenn wir ihn interpretieren, zwei plus zwei gleich vier. Und dies ist eine wahre Aussage in der Mathematik. Ein Satz in einem interpretierten formalen System führt also ein Doppelleben: Immer dann, wenn wir einen gültigen Satz im System erzeugen, so ist dieser gültige Satz, wenn wir ihn interpretieren, *zugleich* auch ein wahrer Satz. Dies hat aber eine praktische Konsequenz: Wenn wir einen *interpretierten* Satz vor uns haben und wissen wollen, ob er wahr ist, müssen wir nur wissen, ob es sich um einen syntaktisch gültigen Satz handelt, denn die Wahrheit *folgt* der formalen Gültigkeit. Dies ist das Formalistenmotto: Syntax mirrors Semantics! Aufgrund der Spiegelung von Syntax und Semantik kann ich daher an einem formalen System ablesen, ob eine mathematische Aussage wahr oder falsch ist.
Dazu bräuchten wir allerdings ein viel mächtigeres formales System, das nicht nur die Addition, sondern auch komplexere Fragestellungen der Mathematik abbildet. Ein Beispiel für ein solches System haben wir bereits kennengelernt: So haben Russell und Whitehead in den *Principia Mathematica* den Versuch unternommen, die gesamte Mathematik zu formalisieren. Ohne hier auf Details einzugehen, so gilt im Wesentlichen auch für ein derart mächtiges formales System: Es enthält eine bestimmte Anzahl von Kritzelzeichen als Startkapital und Regeln, mit deren Hilfe wir aus diesen Kritzelzeichen (den Axiomen) weitere Kritzelzeichen generieren können.

Den Parallelismus von Syntax und Semantik kann man sich am folgenden freilich vereinfachenden Beispiel vergegenwärtigen: Eine bestimmte Anzahl der Zeichen ‚kritzel1' bedeute eine natürliche Zahl (kritzel1 bedeute 1; kritzel1, kritzel1 bedeute 2 usw.). Das Kritzel ‚kritzel2' bedeute ‚hat die Eigenschaft eine Primzahl zu sein'.

Nehmen wir nun an, unser mächtiges formales System hätte die Zeichenkette produziert: ‚kritzel1, kritzel1, kritzel1, kritzel2'.

Nun wissen wir, dass ein im System aufgrund der Ableitungsregeln generierter Satz ein *gültiger* Satz im System ist. Dazu müssen wir unsere Kritzelzeichen überhaupt nicht interpretieren, denn die Gültigkeit von Sätzen ist ja eine rein formale Eigenschaft der Sätze. Wenn wir nun aber unsere Kritzelzeichen interpretieren, so bedeutet der Satz: Drei ist eine Primzahl. Wir wissen von unseren Schulkenntnissen, dass dieser Satz wahr ist. Denn die Zahl drei ist nur durch sich selber und durch eins teilbar. Um aber die Wahrheit dieses Satzes festzustellen, benötigen wir diese Schulkenntnisse (unser semantisches Wissen von Zahlen) aber überhaupt nicht. Denn aufgrund des Doppellebens von semantischer Wahrheit und syntaktischer Gültigkeit gilt: Immer wenn wir einen gültigen Satz im formalen System interpretieren, so handelt es sich zugleich auch um einen wahren Satz.

Wir können das Formalistenmotto in abgewandelter Form auch so ausdrücken: Wenn man auf die typographischen Erzeugungsregeln achtet (die aus gültigen Zeichenketten weitere gültige Zeichenketten produzieren), so wird die Wahrheit der generierten Sätze schon selbst auf sich achten! Wir müssten uns also selbst bei den komplexesten Aufgaben in der Mathematik nicht mehr den Kopf darüber zerbrechen, ob eine Aussage wahr oder falsch ist, wir könnten statt dessen unser formales System befragen: Entspricht unserer mathematischen Aussage ein gültiger Satz im System, so können wir mit Sicherheit davon ausgehen, dass die betreffende Aussage auch wahr ist. Wir hätten sozusagen ein mächtiges Werkzeug in den Händen, mit dessen Hilfe wir die Wahrheit mathematischer Aussagen überprüfen könnten.

Diese Idee hätte weitreichende, auch praktisch verwertbare Folgen, wenn es in einem weiteren Schritt gelänge, die einmal formalisierte Mathematik auch mechanisieren zu können, denn, wie ich schon erwähnt habe, ist die Formalisierung die Vorstufe zur Mechanisierung.

Schließlich ist das Ziel der *computational theory of mind* nicht nur die Semantik zu formalisieren, vielmehr geht es darum, diese formalisierte Semantik auch auf einem Computer implementieren zu können.

Alan Turing

Welche Bedingungen müssen also erfüllt sein, um die formalisierte Semantik automatisieren zu können?

1) Es darf kein semantisches Wissen (das ja außerhalb der Spielregeln angesiedelt ist) verwendet werden. Auch dürfen keine magischen Fähigkeiten zum Einsatz kommen. Was wir stattdessen benötigen, ist eine Anleitung, die in so detaillierten Einzelschritten gehalten ist, dass die einzelnen Verarbeitungsschritte nur noch aus simplen Operationen bestehen, die auch mechanisch ausgeführt werden können.
2) Die Anweisungen müssen garantiert in endlicher Zeit zu einem richtigen Ergebnis führen.

Was ich hier beschrieben habe, nennt man auch einen Algorithmus. Freilich muß erst erfragt werden, was denn nun eigentlich unter diesen „primitiven" Operationen zu verstehen ist. Solange dies nicht feststeht, haben wir nur eine vage intuitive Vorstellung von einem Algorithmus, aber keine brauchbare Definition. Alan Turing hat in den 30er Jahren des vorigen Jahrhunderts dafür eine klare Definition vorgelegt, die als Turings These bekannt geworden ist. Im Lichte der hier angestellten Überlegungen können wir diese These so formulieren: Für jeden Algorithmus gibt es eine formal äquivalente Turing-Maschine.

Algorithmen, die nicht mehr leisten können als eine Turing-Maschine, bezeichnet man als Turing-vollständig. Dies gilt insbesondere für die meisten Programmiersprachen: Alles, was man in einer höheren Programmiersprache ausdrücken kann, kann auch durch eine Turing-Maschine simuliert werden. Wie funktioniert eine solche Turingmaschine?
Eine Turingmaschine besteht im Wesentlichen aus zwei Teilen: Einem Lese/Schreibkopf und einem langen Band, das in einzelne Felder unterteilt ist. Auf jedem der Felder kann jeweils nur ein Zeichen stehen (aus einem vorher fesgelegten Alphabet, beispielsweise die Zeichen 0 und 1). So könnte z.B. auf dem Band die Zeichenfolge 000111000 stehen.
Die Aufgabe des Lesekopfs besteht darin, die einzelnen Felder abzutasten, ein Zeichen zu lesen, eines zu schreiben, sich nach links oder rechts zu bewegen und irgendwann einmal zu stoppen. Welche Operationen vom Kopf der Maschine durchzuführen sind, hängt von zwei Umständen ab: vom inneren Zustand des Kopfes und vom Inhalt des gerade abgetasteten Feldes. Über den Zustand des Kopfes ‚weiß' die Turingmaschine, welche Operationen der Kopf in Abhängigkeit vom abgetasteten Zeichen durchzuführen hat (über diese Metapher müssen wir uns noch genauer unterhalten). Dies ist mit den einfachen Operationen einer Turingmaschine gemeint. Die Operationen sind tatsächlich so einfach, dass man sich unschwer vorstellen kann, wie eine Maschine (ein Computer) diese Aufgabe bewältigen kann.
Um sich eine Vorstellung davon machen zu können, wie eine Turingmaschine im Prinzip ihre Arbeit verrichtet, betrachte man das folgende einfache Beispiel. Es handelt sich hierbei um

eine Maschine, die nichts anderes tut, als einer Folge von 1en eine weitere 1 anzufügen. Die Instruktionen dafür sind die folgenden:

oO -> oOR
o1 -> ı1R
ı1 –> ı1R
ıO -> o1STOP

Das Band enthalte die Folge nachstehender Zeichen:

…00001111….

Jede der vier Zeilen beschreibt eine Operation des Kopfes der Turingmaschine. Die klein geschriebenen Ziffern (o und ı) bedeuten den Zustand, in dem sich der Lese/Schreibkopf gerade befindet. Die großen Ziffern (O oder 1) bezeichnen die Zeichen auf dem Band. Links vom Pfeilzeichen werden der Zustand des Kopfes und der Inhalt des gerade abgetasteten Feldes beschrieben. Rechts werden die aufgrund dieser beiden Bedingungen festgelegten Tätigkeiten beschrieben. Und R bedeutet einfach, gehe einen Schritt nach rechts.
Unsere Maschine tut also das folgende: Sie startet im Zustand o. Liest der Kopf im Zustand o das Zeichen O, so bleibt er im Zustand o, schreibt das Zeichen O und bewegt sich ein Feld nach rechts (dies ist jedenfalls der Sinn der ersten Befehlszeile). Diese Tätigkeit wird so lange wiederholt, bis der Kopf das Zeichen 1 liest. Liest er das Zeichen 1, so geht er in den Zustand ı über, schreibt das Zeichen 1 und rückt ein Feld nach rechts (zweite Zeile). Diese Tätigkeit wird beibehalten (dritte Zeile), bis der Kopf das Zeichen O liest. Ist der Kopf im Zustand ı und liest eine O, so geht er in den Zustand o über, schreibt eine 1 und beendet seine Tätigkeit (vierte Zeile). Unser simples Programm hat somit nichts anderes getan, als an eine Folge von Einsern eine weitere Eins hinzuzufügen.
Neben diesem einfachen Programm gibt es auch Turingmaschinen für die Berechnung des größten gemeinsamen Teilers zweier Zahlen, das kleinste gemeinsame Vielfache usw., wobei all diese verschiedenen speziellen Turingmaschinen in der gleichen Notation geschrieben werden (die Anzahl der Befehlszeilen wäre freilich wesentlich größer als dies in unserem Beispiel der Fall ist). Selbst die Addition, Substraktion, Multiplikation und Division, also unsere Grundrechenarten, müssen erst in Turings Sprache übersetzt werden, sind also nicht „einfach“ genug, um mechanisch reproduziert werden zu können.
Was kann also mit Hilfe von Turings Notation alles berechnet werden? Mit dieser Frage nähern wir uns zuallererst dem Kern von Turings These. Turings Antwort darauf ist die folgende: Jede intuitiv berechenbare Funktion kann auch durch eine Turingmaschine berechnet werden. Immer dann also, wenn ein Mensch zu einem bestimmten Rechenergebnis kommt, kommt auch eine Turingmaschine zum gleichen Ergebnis. Dies ist Turings Variante des Formalistenmottos: Was Menschen berechnen können, kann auch durch eine Turingmaschine berechnet werden. Damit hat Turing mit seiner Maschine ein allgemeines theoretisches Konzept von Berechenbarkeit vorgelegt.
Doch – was bedeutet nun dieser Parallelismus von intuitiver Berechenbarkeit und Turings Maschinenkonzept? Es wurde öfters darauf hingewiesen (Born, Haugeland, Krämer), dass

Turings These kein beweisbares Theorem ist, sondern nur ein Notationssystem, das zuallerst definiert, was Berechenbarkeit bedeutet. Eine Turingmaschine ist kein interpretiertes formales System und die Produktionsregeln einer solcher Maschine sind Anweisungen wie die Maschine von einem Zustand in einen anderen übergehen kann. Sie sind aber keine Schlussregeln, durch die sich die Wahrheit von Sätzen beweisen lässt. (persönliche Mitteilung Hilary Putnams). Das einzige was wir also sagen können, ist lediglich das folgende: Immer dann, wenn eine Maschine einen bestimmten Output generiert, so entspricht diesem, wenn wir ihn interpretieren, eine wahre Aussage in der Mathematik.
Damit haben wir es hier mit einem *Modell* von Berechenbarkeit zu tun, an dem wir mathematische Wahrheiten *ablesen* können. Ähnlich wie wir bei einem Thermometer an der Höhe der Quecksilbersäule die Temperatur ablesen können. Dennoch ist diese Höhe der Quecksilbersäule nicht mit der Temperatur selber identisch.

Die Turingmaschine: ein Modell der menschlichen Rechentätigkeit

Fragen wir also, inwiefern eine Turingmaschine ein Modell der menschlichen Rechentätigkeit ist und betrachten wir hierzu die folgenden beiden Beispiele: Eine Computersimulation zur Vohersage von Wettervorgängen und die Berechnung von Planetenbewegungen mit der Hilfe von Differentialgleichungen.
Durch die Simulation eines Wettervorgangs können wir unter anderem voraussagen wann und wo am nächsten Tag ein Gewitter eintreten wird. Leuchtet beispielsweise ein roter Punkt am Monitor auf, so können wir an dem roten Punkt beispielsweise *ablesen*, wo am nächsten Tag der Blitz einschlagen wird. Das einfache Beispiel zeigt, inwiefern unsere Computersimulation ein Modell des Wettergeschehens ist: Immer dann, wenn an einem Ort der Blitz einschlagen wird, leuchtet am Monitor ein roter Punkt auf (und dort, wo kein roter Punkt aufleuchtet, ist auch nicht mit einem Gewitter zu rechnen). Wir haben es hier mit einem Parallelismus zwischen bestimmten Vorgängen in der Realität und bestimmten Vorgängen im Modell zu tun. Niemand würde jedoch auf die Idee kommen, das Modell mit der Realität gleichzusetzen. Niemand würde auf die Idee kommen, dass in den roten Punkten am Monitor der Blitz einschlägt.
Die gleiche Überlegung gilt auch im Falle der Berechnung von Planetenbahnen. Differentialgleichungen sind ein Modell für die Planetenbahnen. Daraus lässt sich aber nicht der Schluss ziehen, dass die Planeten, um sich um die Sonne drehen zu können. Differentialgleichungen ausführen müssen! Auch hier besteht nur ein Parallelismus (und keine Identität) zwischen dem Modell und der Realität.
Aus diesen Beispielen ist ersichtlich, unter welchen Bedingungen eine Turingmaschine als adäquates Modell der menschlichen Rechentätigkeit gelten kann. Dies ist dann der Fall, wenn ein menschlicher Rechner nur zu Ergenissen kommen kann, die auch an den von der Machine produzierten Zeichenketten abgelesen werden können. Es dürfte also nicht der unseres indischen Mathematikers eintreten, der mathematische Theoreme entdeckt, zu denen jedoch ein syntaktisch äquivalenter Output einer Turingmaschine fehlt. Aber selbst unter der Voraussezung, dass Turings Maschine tatsächlich ein geeignetes Modell der menschlichen Rechentätigkeit wäre, hätten wir noch lange nicht, und das möchte ich hervorheben, eine künstliche Intelligenz geschaffen. Denn der Parallelismus von menschlicher und maschineller Rechentä-

tigkeit besagt ja nur, dass immer dann, wenn der Mensch zu einem richtigen Ergebnis kommt, eine Maschine einen dazu passenden syntaktish korrekten Output erzeugt. Dass wir an diesem Output das Ergebnis des menschlichen Mathematikers kontrollieren können, ist ja nur dann der Fall, wenn wir diesen Output zu interpretieren beginnen! Diese Wenn-Bedingung ist aber ein massiver Einschub. Sie besagt eben, dass maschinelles Rechnen im besten Fall als Modell des menschlichen Rechnens verwendet werden kann, nicht aber mit letzterem gleichgesetzt werden darf.
KI-Forscher könnten diesem Einwand allerdings mit dem folgenden Argument begegnen: Klarerweise ist im Falle einer Wettersimulation das Modell etwas ganz anderes als das reale Wettergeschehen. Eine Simulation eines Wettervorgangs ist natürlich kein Wettervorgang, ganz anders verhalte es sich jedoch, so wird jedenfalls argumentiert, bei einem Denkvorgang. Schließlich sei Denken nichts anderes als die Verarbeitung abstrakter Symbole, unabhängig von deren physikalischer Realisierung. Wegen dieser Unabhängigkeit von einem speziellen materiellen Träger sei Intelligenz auch Spezies unabhängig. Es wäre allerdings etwas voreilig, dieses Argument nun dazu zu verwenden, um eine Gleichsetzung von Turings Modell der Berechenbarkeit mit dem menschlichen Rechnen zu rechtfertigen. Denn diese Gleichsetzung gilt ja nur unter der Vorausstezung, dass sich menschliches Denken (Rechnen) auf das Hantieren mit formalen Symbolen reduzieren lasse. Diese Gleichsetzung des Modells mit dem menschlichen Rechner wird auch als These der so genannten *starken Künstlichen Intelligenz* bezeichnet. Diese versucht durch Produkte der Künstlichen Intelligenz die natürliche Intelligenz zu *imitieren* im Unterschied zur so genannten schwachen Künstlichen Intelligenz, der es nur um eine *Simulation*, also um Erklärungsmodelle für die natürliche Intelligenz zu tun ist.

Die Turingmaschine – ein Ersatz menschlichen Rechnens

Was muß aber geschehen, damit man tatsächlich die Tätigkeit einer Rechenmaschine (das Produzieren irgendwelcher Zahlzeichen) als *Rechnen* im vollen Sinne des Wortes bezeichnen könnte? Dies wäre dann der Fall, wenn die Machine nicht *nur* ein Modell der menschlichen Rechentätigkeit wäre (der Parallelismus von maschineller und menschlicher Rechentätigkeit ist sozusagen eine notwendige aber eben nicht hinreichende Bedingung), sondern eben selbständig *rechnen* täte, einen menschlichen Rechner also ersetzen könnte. Und ersetzen könnte eine Turingmaschine einen menschlichen Rechner dann, wenn sie sich gleich *verhält* wie ein menschlicher Rechner. Ist also das Verhalten ein geeignetes Kriterium für die Bestimmung der Identotät von machinellem und menschlichem Rechnen? Betrachten wir hierzu das folgende Beispiel: Ein Vogel kann fliegen. Aber auch ein Flugzeug kann fliegen. Können wir aber tatsächlich vom gleichen Verhalten ausgehen? Ein Vogelflug ist doch mehr als eine bloße Fortbewegung von A nach B. Wenn wir vom Verhalten eines Vogels beim Fliegen sprechen, so können wir damit den Flug bei der Balz meinen, die Jagd nach Beute oder aber auch die Reise der Vögel während der Winterzeit von Norden nach Süden. Das Verhalten eines Vogels beim Fliegen lässt sich nicht reduzieren auf die physikalischen Bewegungen eines Körpers, sondern ist abhängig vom Kontext, Lebenserfahrung und Geschichte eines Organismus. Kurz und gut, die Redeweise vom Verhalten enthält auch eine *intentionale Dimension*.
Übertragen wir jetzt diese Überlegungen auf die Leistungen eines menschlichen Rechners, so ergibt sich das folgende Bild: Auch für einen menschlichen Rechner kann ein Rechenergebnis in unterschiedlichen Kontexten eine ganz verschiedene Bedeutung haben. Für einen Buchhal-

ter bedeutet eine Zahl unter Umständen etwas ganz anderes als für eine Hausfrau, die gerade einkaufen geht. Dies liegt daran, dass beide die Zahl in verschiedenen Kontexten verwenden und daher auch mit ihr etwas anderes anfangen können. Damit überhaupt davon die Rede sein kann, dass eine Turingmaschine das gleiche leistet wie ein menschlicher Rechner, dürften die von der Maschine produzierten Zahlzeichen nur auf eine einzig zulässige Weise interpretiert werden. Würde die Interpretation der Zahlzeichen eindeutig aus deren syntaktischer Produktion folgen, so könnte ein menschlicher Rechner mit den von der Maschine produzierten Zahlzeichen auch nichts anderes anfangen als die Maschine selbst.

Dies wäre aber nur dann der Fall, wenn ein Mensch sich beim Rechnen strikt an die Anweisungen einer Turingmaschine hält! So hat Turing selbst einen rechnenden Menschen, der mit Anweisungen, Bleistift, Papier und Radiergummi arbeitet, auch als *Papiermaschine* bezeichnet: „Es ist möglich, den Effekt einer Rechenmaschine zu erreichen, indem man eine Liste von Handlungsanweisungen niederschreibt und einen Menschen bittet, sie auszuführen. Eine derartige Kombination eines Menschen mit geschriebenen Instruktionen wird ‚Papermaschine' genannt (Turing, 1988, 91).

Jetzt müssen wir uns nur noch überlegen, an welcher Stelle bei einer solchen Papiermaschine denn überhaupt Semantik ins Spiel kommt. Nehmen wir mal an, die Aufgabe einer solchen Maschine bestünde darin, den größten gemeinsamen Teiler von zwei Zahlen zu berechnen. Der mit den geschriebenen Instruktionen arbeitende Mensch hat keine Ahnung, welchen Sinn seine einzelnen Arbeitsschritte in Wirklichkeit haben. Er hat auch keine Ahnung davon, dass das von ihm schlußendlich produzierte Zahlzeichen der größte gemeinsame Teiler zweier Zahlen ist. Das einzige, dessen er sich unmittelbar bewußt ist, ist das Symbol an einer bestimmten Stelle des Turingbandes. Dieser Umstand läßt sich nun mutatis mutandis auch auf eine Turingmschine übertragen. So betont Turing ausdrücklich, dass das abgetastete Symbol das einzige sei, dessen sich die Maschine „direkt bewußt" sei. Turing ist sich freilich des metaphorischen Sprachgebrauchs im klaren, schreibt er doch „direkt bewußt" unter Anführungszeichen. Nachdem aber dies die einzige Stelle ist, an der auch bei einer Papiermaschine Semantik hereinkommt, kann ein auf diese Weise rechnender Mensch mit seinem Rechenergebnis nichts anderes anfangen als die Turingmaschine selbst. Semantik spielt in diesem Zusammenhang also dehalb nur eine untergeordnete Rolle, weil im Konzept der Papiermaschine das menschliche Rechnen von vorne herein durch die Brille einer Turingmaschine betrachtet wird!

So soll Turing in einer Diskussion über Künstliche Intelligenz einmal das folgende gesagt haben: Sage mir *exakt*, wodurch sich die Leistung eines Menschen von jener eines Computers unterscheidet und ich werde eine Maschine bauen, die eben diesen Unterschied auch leisten kann. Turing führt hier seinen potentiellen Kontrahenten alllerdings aufs Glatteis. Denn mit seiner Forderung nach einer exakten Definition des Unterschiedes von Mensch und Maschine verlangt er eine Beschreibung dieses Unterschiedes in der Diktion eines *Algorithmus*, also gerade in jener Diktion, die gemäß der Turingschen These auch durch eine Turingmaschine ausgedrückt werden kann. Turing übersieht dabei die Möglichkeit, dass ein etwaiger Unterschied von Mensch und Maschine eben nicht exakt beschrieben werden kann. Schließlich könnte es ja auch mathematische Wahrheiten geben, die überhaupt nicht formalisierbar sind.

Um entscheiden zu können, ob eine Machine das gleiche leisten könne wie ein Mensch, gibt Turing nun ein Kriterium an, den so genannten Turing-Test. Er wollte damit die Frage klären, ob Maschinen denken können. Mit den Gütekriterien dieses Tests müssen wir uns noch ge-

nauer auseinandersetzen. Doch vorher sei der Test vorgestellt. Wir müssen uns dazu zwei Situationen vorstellen.
Situation I: Ein Mann und eine Frau befinden sich in einem Zimmer. Von ihnen räumlich getrennt sitzt ein Fragesteller in einem anderen Zimmer, dessen Aufgabe es ist, herauszufinden, welche der beiden Personen die Frau bzw. der Mann ist. Einzige Kommunikationsmöglichkeit zwischen den beiden Zimmer sein ein Fernschreiber, der Fragesteller erhält also nur Antworten in schriftlicher Form. Ziel des Mannes sei es, den Fragesteller zu täuschen, in seinen Antworten also das Verhalten einer Frau zu imiitieren. Turing bezeichnet daher diese Versuchsanordnung auch als Imitationsspiel.
Was bei Turing freilich nur ein Gedankenexperiment war, findet in der Gegenwart seine ganz konkrete Entsprechung in Chatrooms, in denen so genannte „fakes“ dem Gesprächspartner ein falsches Geschlecht vortäuschen.
Situation II: Man tausche den Mann durch eine Maschine aus. Sollte sich der Fragesteller auch durch die Maschine ebenso oft täuschen lassen wie in dem Falle, wenn das Spiel mit einem Mann und einer Frau gespielt wird, so habe die Maschine den Turing-Test bestanden.

Ein praktisches Beispiel für einen solchen Test war das von Joseph Weizenbaum entwickelte Programm Eliza, das ein Gespräch mit einem Psychiater simulierte. Im Wesentlichen arbeitete das Programm nach der Methode, aus Sätzen Schlüsselworte zu extrahieren, um dann aus einer Datenbank vorfabrizierte Fragen oder Antworten zu geben. Macht der Patient die Bemerkung, „Meine Mutter hat mir heute meine Lieblingsspeise gekocht“, so fragt Eliza beispielsweise zurück: „Erzählen Sie mir mehr von Ihrer Mutter“. Der Grund für den erstaunlichen Erfolg von Eliza lag aber nicht an seinem tatsächlichen Wissen über die reale Welt, sondern daran, dass im therapeutischen Gespräch oft eine standardisierte Konversation stattfindet. Sobald nämlich Eliza mit ungewohnten oder gar sinnlosen Fragen konfrontiert wurde, konnte es ohne größere Schwierigkeiten aufgrund seiner konfusen Antworten als Programm enttarnt werden.

So ist dem Turing-Test in der Praxis nur ein schmaler Erfolg beschieden gewesen. Ein seit 1991 ausgeschriebener und mit einem Preisgeld von 100.000 US Dollar dotierter Wettbewerb konnte bis heute noch von keinem Computerprogramm gewonnen werden. Dies liegt unter anderem wohl daran, dass Turings Test nur in solchen Bereichen erfolgreich war, in denen eher oberflächliche Gespräche – man denke an die Gespräche bei einer Cocktailparty – zu finden sind.

Grundsätzlich ist gegen diesen Test einzuwenden, dass selbst bei bestandenem Turingtest keineswegs der Nachweis erbracht ist, dass das Computerprogramm auch die von ihm verarbeiteten Symbole tatsächlich versteht. Denn wer sagt denn, dass sich Semantizität (die Fähigkeit, die Bedeutung von Zeichen zu verstehen) und die für einen Turingtest benötigte Intelligenz sich gegenseitig bedingen? Daran anschließend müsste auch noch geklärt werden, welche Rolle Semantizität bei intelligentem Verhalten zukommt. Ferner stellt sich die Frage, ob denn aus der Beobachterperspektive überhaupt entschieden werden kann, ob ein Programm, das formal Symbole verarbeitet, auch tatsächlich ein Verständnis für diese Symbole entwikkeln kann. Ich komme auf diese Fragen im Verlauf dieser Untersuchungen zurück.

Fürs erste möchte ich jedoch auf einen anderen, bislang noch unerwähnten Aspekt von Turings Maschine eingehen, der womöglich ein psychologisches Motiv dafür war, dass die Turingmaschine von einem Modell der menschlichen Rechentätigkeit zu einem Ersatz menschlichen Rechnens avancieren konnte.

Die universale Turingmaschine

Wir haben in den oben angebenen Beispielen immer nur spezielle Turingmaschinen kennengelernt. Jede dieser Maschinen ist dazu in der Lage, eine ganz bestimmte Aufgabe zu lösen, wobei die einzelnen Lösungsschritte durch die Anweisungen der Turingmaschine vorgegeben sind. Nun gibt es aber auch Turingmaschinen, die dazu in der Lage sind, andere Maschinen zu simulieren. Man kann sich dies am Beispiel jener einfachen Turingmaschine, die an einer Folge von 1en eine weiter 1 anhfügt, so vorstellen: Die Folge von 1en auf dem Band der speziellen Turingmaschine sind unsere *Daten*. Die Anweisungen dieser Maschine sind unser *Programm*. Wir bezeichnen jene Turingmaschine, die unsere spezielle Maschine simulieren soll, als den *Simulator*. Codieren wir jetzt das Programm der speziellen Turingmaschine und verwenden wir es als Eingabe auf einem bestimmten Abschnitt des Bandes unseres Simulators, so geschieht das folgende: Der Simulator liest das Programm ein, *interpretiert* dessen Anweisungen und führt sie dann aus.

Turing konnte den Beweis erbringen, dass es Maschinen gibt, die dazu in der Lage sind, *jede* spezielle Turingmaschine simulieren zu können. Mit einer solchen universalen Turingmaschine hat Turing die theoretischen Grundlagen gelegt für einen modernen Allzweckrechner.

Man kann sich diese fundamentale und revolutionäre Idee am besten an der Gegenüberstellung eines ganz einfachen (nicht programmierbaren) Taschenrechners mit einem PC verdeutlichen. Ein nicht programmierbarer Taschenrechner entspricht einer speziellen Turingmaschine. Er kann nur das ausführen, wozu er vom Hersteller programmiert wurde. Seine Anweisungen sind in seiner Hardware „fest verdrahtet“. Im Unterschied zu einem solchen Taschenrechner ist jeder moderne Personal Computer *offen* für neue Anwendungen. Er entspricht von seinen Möglichkeiten her einer universalen Turingmaschine. Immer dann, wenn Sie ein neues Programm haben, können Sie dieses Programm in den PC einlesen und es dann von einem *Compiler* (zu Deutsch: einem Übersetzer) in die Maschinensprache übersetzen lassen. Ein PC ist also dazu in der Lage, nicht nur Daten einzulesen, sondern auch Instruktionen zur Verarbeitung von Daten. Die Interpretation dieser Instruktionen (die Übersetzung in die Maschinensprache) wird von der Maschine selbst übernommen. Dadurch verwischt sich aber der Unterschied zwischen den Anweisungen zur Verarbeitung der Daten und den Daten selber. Die Anweisungen der allgemeinen Turingmaschine verarbeiten Anweisungen der speziellen Turingmaschine. Was hier verarbeitet wird und von wem diese Arbeit erledigt wird, läßt sich also in ein und demselben Code ausdrücken. Turings Code bezieht sich auf sich selbst und es entsteht der Anschein, als könne die Syntax ihre eigene Semantik verschlucken.

Es stellt sich im Anschluß an diese Überlegungen die Frage, inwiefern ein solches sich selbst reflektierendes System überhaupt dazu in der Lage ist, seinen Aufgabebereich – dies ist im Falle von Turings Code die Mathematik - widerspiegeln zu können. Enthält ein Code, der sich auf sich selbst bezieht, am Ende gar Wahrheiten, die sich in diesem Code zwar ausdrükken, aber nicht in ihm beweisen lassen? Inwiefern gilt also für einen selbstbezüglichen Code das Formalistenmotto: Syntax mirrors semantics?

Gödels Theoreme

Es war der österreichische Mathematiker Kurt Gödel, der in den 30er Jahren des vergangenen Jahrhunderts in einem bahnbrechenden Aufsatz *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I* eben diese Spiegelung in Frage gestellt hat. In seinen berühmten Unvollständigkeitstheoremen hat er die Möglichkeiten und Grenzen der Formalisierung arithmetischer Wahrheiten ausgelotet. Er konnte beweisen, dass die Wahrheit von Aussagen in der Mathematik nicht auf formale Beweisbarkeit reduziert werden kann. Der Begriff Beweisbarkeit ist sozusagen schwächer als der Begriff der Wahrheit.
Nachdem aber ein formales System auch in einem Computer realisiert bzw. „instanziiert" werden kann (siehe Turing), markieren die Grenzen der Formalisierung zugleich die Grenzen der Mechanisierung. Gödels Theoreme wurden daher von manchen Autoren (insbesondere von Lukas und Penrose) als Argumente dafür angeführt, dass der menschliche Geist einem Computer prinzipiell überlegen sei und daher auch die klassische Künstliche-Intelligenz-Forschung von vornherein zum Scheitern verurteilt sei. Diese Argumentation blieb allerdings nicht unwidersprochen. Was uns diese Argumente über die Natur des menschlichen Geistes verraten (falls sie überhaupt etwas verraten), darauf komme ich gleich zurück.

Zunächst aber eine knappe Zusammenfassung von Gödels Theoremen.

Das erste Unvollständigkeitstheorem besagt: Für jede konsistente formale Theorie, die grundlegende arithmetische Wahrheiten beweisen kann, kann in dem System eine Aussage konstruiert werden, die wahr ist, nicht aber in dem System abgeleitet werden kann. Jedes hinreichend mächtige konsistente formale System ist unvollständig (*hinreichend mächtig* heißt hier, dass das formale System zumindestens die elementare Zahlentheorie enthält).

Das zweite Unvollständigkeitstheorem ergibt sich unmittelbar aus dem ersten. Es besagt: In einem hinreichend mächtigen konsistenten formalen System kann eben diese Konsistenz nicht bewiesen werden.

Gödels Beweis für die beiden Theoreme ist technisch sehr aufwändig. Er lässt sich umrisshaft so beschreiben. Gödel transkribiert Aussagen des formalen Systems in zahlentheoretische Aussagen. Nachdem das formale System von der Zahlentheorie handelt, lassen sich auch Aussagen in dem formalen System über diese zahlentheoretischen Aussagen konstruieren. Dadurch wird das formale System „introspektiv". Es lassen sich Aussagen im formalen System über dieses System konstruieren, ja sogar Aussagen, die sich auf sich selber beziehen.
Gödel konstruiert einen Satz G, der von sich selber seine eigene Unbeweisbarkeit im System behauptet. Der Satz kann wahr oder falsch sein. Wäre er falsch, so müsste sich der Satz im System beweisen lassen. Nun sind aber in einem konsistenten formalen System beweisbare Aussagen *wahre* Aussagen. Lässt sich aus einem formalen System eine *falsche* Aussage ableiten, so ist das System inkonsistent.
Ist das System dahingegen konsistent, so kann der Satz nur wahr sein. Diese Wahrheit des Satzes G lässt sich aber nicht über eine Ableitung des Satzes aus dem System ermitteln, obwohl es sich hier um einen Satz handelt, der im System ausgedrückt werden kann.

Damit ist aber das System unvollständig. Man beachte, dass diese Unvollständigkeit nur formale Systeme betrifft, die etwas ausdrücken können (so ist etwa das weiter oben besprochene MIU-System viel zu ausdrucksschwach, um unvollständig sein zu können). Darüber hinaus bezieht sich die Unvollständigkeit nur auf das, was innerhalb des Systems ausgedrückt werden kann. So ist der Satz „Paris ist die Hauptstadt von Frankreich" zwar wahr, lässt sich aber nicht aus einem formalen System ableiten. Es handelt sich hierbei indes um keinen Satz, der im formalen System ausgedrückt werden kann. Unvollständigkeit besteht nur dann, wenn ein Satz, der sich im formalen System ausdrücken lässt, wahr ist, aber nicht aus dem System abgeleitet werden kann (und auch kein Axiom des Systems ist).

Gödel hat aber nicht *nur* bewiesen, dass das System der Principia Mathematica unvollständig ist, sein Beweis gilt stattdessen für alle formalen Systeme von hinreichender Mächtigkeit. Denn auch der Versuch, das unvollständige System um den unentscheidbaren Satz G als Axiom zu erweitern, um es dadurch vollständig zu machen (ein Axiom ist *per definitionem* wahr), schlägt fehlt. Auch in einem derart erweiterten System lässt sich nämlich erneut ein unentscheidbarer Satz konstruieren.

Gödels zweites Unvollständigkeitstheorem ist eine unmittelbare Folge des ersten. Im ersten Unvollständigkeitstheorem wurde der Beweis erbracht: Wenn ein System S konsistent ist, so folgt daraus, dass G wahr ist, kurz ausgedrückt: Aus der Konsistenz von S folgt G. Daraus ergibt sich: Wenn die Konsistenz von S bewiesen werden kann, so muss auch G bewiesen werden können (da G aus S folgt). Nun kann aber G in S nicht bewiesen werden. Damit kann aber auch die Konsistenz von S in S nicht bewiesen werden.

Mit diesen beiden Theoremen hat Gödel grundsätzlich die Grenzen der Formalisierung aufgezeigt. Was das aber für die Natur des menschlichen Geistes bedeutet, ist alles andere als klar und wurde auch im Anschluss an Gödel kontroversiell diskutiert. Ich möchte diese sich an Gödels Unvollständigkeitstheoreme anschließende Debatte am Beispiel einer Kontroverse zwischen Roger Penrose (der die Gleichsetzung des menschlichen Geistes mit einem Computer aufgrund von Gödels zweitem Unvollständigkeitstheorem bestreitet) und Hilary Putnam (der Penrose vehement widerspricht) nachzeichnen.

Die Putnam-Penrose-Kontroverse

Beginnen wir mit folgender Überlegung: Angenommen, es gäbe ein formales System, das vollständig die Fähigkeiten eines menschlichen Mathematikers simuliert. Das würde bedeuten, dass jede wahre Aussage des menschlichen Mathematikers zugleich auch ein beweisbares Theorem in dem formalen System wäre. Nehmen wir ferner an, unser menschlicher Mathematiker weiß, dass es dieses formale System gibt, das seine gesamten mathematischen Fähigkeiten simuliert. Dann müsste er aber auch dazu in der Lage sein, die Gültigkeit des formalen Systems erkennen zu können (Gültig ist ein System dann, wenn aus seinen Axiomen unter Nutzung der Ableitungsregeln nur wahre Sätze erzeugt werden können). Würde er aber die Gültigkeit eines formalen Systems erkennen, das seine *gesamten* mathematischen Fähigkeiten simuliert, so müsste sich die Aussage „das System ist gültig" auch als Theorem in diesem System ableiten lassen (da die Gültigkeit eines formalen Systems zu erkennen, ja ebenfalls eine mathematische Fähigkeit ist). Dies wäre aber nach dem zweiten Gödelschen Unvollstän-

digkeitstheorem nicht möglich. Die Gültigkeit eines solchen Systems kann nur „von außen" und nicht innerhalb des Systems festgestellt werden.
Nun ist ein Computer die Instanziierung eines formalen Systems. Haben wir ein bis ins kleinste Detail beschriebenes formales System (eine Menge von Axiomen und Ableitungsregeln), so können wir das System in einem Computerprogramm implementieren, das automatisch alle Theoreme dieses Systems generiert. Roger Penrose kommt aufgrund dieser Überlegungen zu dem Schluß, dass die menschliche Intelligenz nicht durch einen Computer simuliert werden könne. Die Fähigkeiten eines menschlichen Mathematikers seien dahingegen prinzipiell von nicht algorithmischer Natur. Denn ein menschlicher Mathematiker könne immer die Gültigkeit eines Computerprogramms erkennen, ohne dass sich diese Erkenntnis auch als Theorem in diesem Programm erzeugen lässt. Er geht daher von der Annahme aus, dass im menschlichen Gehirn nicht komputationale Prozesse stattfinden müssen, was er über die Quantenmechanik zu zeigen versucht. (Penrose Hypothese, nicht deterministische Prozesse im Gehirn über einen Kollaps der Wellenfunktion, ausgelöst durch Mikrotuboli in den Nervenzellen, erklären zu können, ist eine auch unter Neurobiologen umstrittene Spekulation. Ich gehe auf diesen Aspekt im folgenden nicht näher ein.)
In einer Rezension von Penrose' *Shadows of the Mind*, erschienen 1994 in der New York Times und in einem Vortrag gehalten in Wien anlässlich des 100ten Geburtstages von Gödel im Jahre 2006 mit dem Titel *The Gödel Theorem and Human Nature* widerspricht Putnam vehement der Auffassung, aus Gödels Theoremen auf die Existenz nichtkomputationaler Prozesse im menschlichen Gehirn schließen zu können. Putnams Argumentation ist äußerst subtil. Denn er verwendet das *gleiche* Argument, mit dem er Penrose kritisiert, um den Glauben an eine uneingeschränkte Anwendung des Funktionalismus in seine Schranken zu verweisen. Wie dies möglich sein kann, soll im folgenden gezeigt werden. Schauen wir uns dies im Detail an.
Zunächst einmal argumentiert er ganz ähnlich wie Penrose: Ein Programm, dessen Gültigkeit wir erkennen können, kann nicht unsere gesamten mathematischen Fähigkeiten simulieren. Putnam greift hier eine Argumentation auf, die er bereits vor mehreren Jahrzehnten gegenüber Chomsky vertreten hat. Nehmen wir an, es wäre möglich die gesamte wissenschaftliche Epistemologie durch eine spezielle Turingmaschine zu simulieren, so wäre es uns unmöglich, diese Tatsache zu erkennen. Es wäre allerdings ein Trugschluss, daraus schließen zu wollen, dass eine solche Turingmaschine *prinzipiell* nicht existieren könne (wie dies eben Penrose getan hat). Denn schließlich könnte es ein Programm geben, das so komplex ist, dass wir seine Gültigkeit eben nicht erkennen können. Doch was heißt *prinzipiell* existieren zu können? Am Leitfaden dieser Frage entwickelt Putnam einen Gedankengang, der schließlich zu einer Beschneidung der *faktischen* Möglichkeiten des Funktionalismus führt.
Zunächst gilt das folgende: Es kann die Gültigkeit nur solcher formaler Systeme erkannt werden, die eben *nicht* unsere gesamten mathematischen Fähigkeiten beschreiben. Daraus folgt ferner, dass die Vernunft über alles, was sie zu formalisieren *vermag*, hinausgehen kann (vgl. Putnam, RR, S. 208). Wichtig ist in diesem Zusammenhang der Nebensatz: was sie zu formalisieren vermag. Wir können nämlich nur das formalisieren, dessen Gültigkeit wir auch erkennen können. Ein formales System, das unsere gesamte mathematische Fähigkeit beschreibt, ist dahingegen nur ein Ideal der Vernunft, eine Normierung der menschlichen Rechentätigkeit, und darf nicht mit einer ontologischen Beschreibung der faktischen Rechenleistung eines menschlichen Mathematikers verwechselt werden. Es ist nun diese Normierung

der menschlichen Rechentätigkeit, die nicht auf ein Computerprogramm (dessen Gültigkeit wir erkennen können) reduziert werden kann. Man beachte: Es geht hier nicht um die faktische, sondern um die idealtypische Rechenleistung eines Mathematikers. Der Output eines idealtypischen Mathematikers ist nicht beschreibbar als der Output einer Maschine, deren Programm wir kennen können. Doch was folgt daraus?
Zunächst folgt daraus, dass *wir* unsere eigene mathematische Fähigkeit nicht vollständig formalisieren können, weil es eben zu dieser mathematischen Fähigkeit gehört, über alles hinausgehen zu können, was sie zu formalisieren vermag (vgl. Putnam, RR, S. 208). Hier zeigen sich die Grenzen des Funktionalismus. Daraus nun aber ableiten zu wollen, dass kein Computerprogramm *prinzipiell* dazu imstande wäre, unsere gesamten mathematischen Fähigkeiten zu simulieren, wäre ein unzulässiger Fehlschluß. Zu einem derartigen Fehlschluß kommen wir nur dann, wenn wir die faktischen Fähigkeiten eines menschlichen Mathematikers mit einer idealtypischen Beschreibung dieser Fähigkeiten verwechseln. Und nur im Falle einer solchen Verwechslung können wir auf die Idee verfallen, im menschlichen Gehirn irgendwelche mysteriöse nichtkomputationale Prozesse zu vermuten.
Mathematische Wahrheit ist sozusagen zumindestens prinzipiell nicht jenseits jeder Formalisierung, sie übersteigt nur jedes formale System, dessen Gültigkeit wir erkennen können.

Dass wir aus den faktischen Fähigkeiten eines Mathematikers nicht auf die Fähigkeiten eines idealtypischen Mathematikers schließen können, dafür führt Putnam das folgende Argument an. Wenn es ein Programm gibt, das unseren epistemischen Beweisbegriff formalisiert, so kann dieses Programm alle möglichen Beweise produzieren – eine potentiell unendliche Liste von Beweisen. Warum das so ist, zeigt nachstehendes Beispiel: Haben wir ein Programm, das für jede x-beliebige Zahl überprüft, ob sie eine Primzahl ist oder nicht, so bestimmt dieses Programm *alle* Primzahlen – eine potentiell unendlich lange Liste von Primzahlen. Fragen wir nun:
Welche Evidenz haben wir dafür, dass ein solches Programm existiert? Die Möglichkeit eine potentiell unendlich lange Liste von Primzahlen erzeugen zu können, gehört nicht zu dieser Evidenz. Das einzige, worauf wir uns stützen können, sind die Fähigkeiten eines real existierenden Mathematikers. Dessen Fähigkeiten, Primzahlen bestimmen zu können, sind jedoch physikalisch beschränkt.
Die einzige Evidenz dafür, dass es möglich ist, eine potentiell unendlich lange Liste von Primzahlen zu erzeugen, besteht vielmehr darin, dass ein Programm existiert, dessen Gültigkeit wir *erkennen* können. So konnte Chomsky zeigen, dass ein idealisierter Sprecher mit Hilfe der Transformationsgrammatik eine potentiell unendliche Liste von Sätzen generieren kann. Er konnte dies aber nur zeigen, *weil* seine idealisierte Grammatik eben einen genau definierten Algorithmus verwendet, dessen Gültigkeit wir auch erkennen können.
Versuchen wir nun aber den epistemischen Beweisbegriff zu formalisieren, so haben wir ein Problem: Via des Anti-Chomsky-Unvolltsändigkeitstheorems lässt sich nämlich zeigen, dass, falls ein Programm existiert, das den epistemischen Beweisbegriff vollständig repräsentiert (d.h. es existiert eine Turingmaschine, die alle Beweise eines menschlichen Mathematikers reproduzieren kann), wir dessen Gültigkeit nicht erkennen können.
Aus der Tatsache, dass ein menschlicher Mathematiker rekursiv eine „phantastisch“ große Menge an mathematischen Theoremen beweisen kann, folgt daher auch nicht, dass ein Algo-

rithmus existiert, mit dessen Hilfe ein idealtypischer Mathematiker eine potentiell unendliche Liste von Theoremen beweisen kann.

Das Chinesische-Zimmer-Gedankenexperiment

Eine kurze Zwischenbilanz darüber, welche Lehren nun tatsächlich aus Gödels Theoremen gezogen werden können, ergibt das folgende Resultat: Zumindestens *wir* als denkende Wesen können unser eigenes Denken nicht vollständig formalisieren. Dies ist aber – aufgepasst! – noch kein Beweis, dass das Programm der Künstlichen Intelligenz von vornherein zum Scheitern verurteilt ist.
Denn die meisten KI-Systeme bedienen sich heuristischer Regeln und verlangen daher auch gar keine vollständige Formalisierung eines bestimmten Wissensgebietes. Heuristische Regeln sind häufig verwendete Faustregeln eines menschlichen Experten, die in vielen Fällen zwar brauchbar sind, aber nicht immer zu einer garantiert richtigen Lösung führen müssen. Formale Systeme vom Typ der Principia Mathematica, auf die sich Gödels Theoreme anwenden lassen, verwenden dahingegen Ableitungsregeln, die nur korrekte Beweise erzeugen.
Der amerikanische Sprachphilosoph John Searle hat 1980 in einem Gedankenexperiment zu zeigen versucht, dass die natürliche Intelligenz (insbesondere das Verstehen von Symbolen) prinzipiell nicht durch Computerprogramme *ersetzt* werden könne. Seine Kritik richtet sich somit gegen die These der starken Künstlichen Intelligenz, also gegen die Annahme, natürliche Intelligenz durch KI-Systeme nachbauen bzw. duplizieren zu können. Davon unbetroffen ist die These der schwachen Künstlichen Intelligenz, derzufolge KI-Systeme nichts anderes sind als Erklärungsmodelle, also Simulationen von natürlicher Intelligenz, sofern man sich nur der Differenz von Modell und modellierter Realität bewusst ist. Searle möchte mit seinem Gedankenexperiment aufzeigen, dass ein Computerprogramm nicht dazu in der Lage ist, im buchstäblichen Sinne die Bedeutung der von ihm verarbeiteten Symbole zu verstehen.
Searle knüpft in seinem Gedankenexperiment an die Arbeit von Roger Schank an (vgl. Schank und Abelson, 1977), wobei dieser Bezug allerdings nur exemplarischen Charakter hat. Seine Kritik richtet sich prinzipiell gegen jeden Versuch, mit Hilfe formaler Manipulationen von Symbolen das Niveau menschlichen Verstehens erreichen zu können. 'Formal' bedeutet hier, dass Zeichenketten nur an Hand ihrer Form bzw. äußeren Gestalt und eben *nicht* aufgrund ihrer Bedeutung verarbeitet werden. Schank macht sich in seinen Programmen eine bestimmte Fähigkeit zunutze, die Menschen bei ihrem Alltagsverständnis von Geschichten einsetzen. Sie erfassen bei einer Geschichte wesentlich mehr, als in ihr erzählt wird. Sie können aus ihr auch Informationen herauslesen, die explizit in ihr gar nicht enthalten sind. Dies liegt daran, daß eine Geschichte vor dem Hintergrund eines bestimmten Vorverständnisses - einer Menge stereotyper Erwartungen und Kenntnissen - interpretiert wird. Derartige Vorkenntnisse bezeichnet Schank als Skripte. Der Ausdruck 'Skript' wird der Filmsprache entlehnt und bezeichnet die Szene eines Films. So verbinden wir beispielsweise mit einer alltäglichen Geschichte über einen Restaurantbesuch ein ganzes Set von Erwartungen. Denken wir an das folgende Beispiel: "John ging in ein Restaurant. Nach dem Essen half ihm der Kellner in den Mantel und wünschte ihm noch einen schönen Tag."
In der Geschichte wurde nicht erwähnt, dass John die Rechnung beglichen hat. Nun gehört zu unserem stereotypen Restaurantskript aber die Erwartung, dass in einer solchen Situation

das Essen auch bezahlt wurde. Schank macht sich diesen Umstand zunutze und gibt derartige Erwartungen in Form von 'Skripten' explizit in sein Programm ein. Liest das Programm eine derartige Geschichte, so wird es die Frage, ob John auch bezahlt hat, mit Ja beantworten. Anders als Weizenbaum's Eliza, das aufgrund bestimmter eingegebener Schlüsserwörter (z.B. Mutter oder Vater) vorgefertigte Sätze am Bildschirm ausgibt, ensteht so der Eindruck, als könne ein Programm unter Nutzung derartiger Skripts tatsächlich eine Geschichte verstehen. Von Vertretern der starken KI wird aufgrund eines solchen Ergebnisses sogar unterstellt, dass es sich hier nicht nur um eine Simulation von menschlichem Verstehen handle, sondern um ein Verstehen im buchstäblichen Sinne des Wortes. Diese Unterstellung ist aber schon allein aus softwaretechnischen Gründen (also nicht nur aus prinzipiellen philosophischen Erwägungen) problematisch. Zu jeder Geschichte finden sich nämlich Ausnahmesituationen, auf die zwar ein Mensch flexibel reagieren kann, nicht jedoch ein Programm. Man denke nur an den folgenden Fall. John sei der Sohn des Restaurantbesitzers und werde aus diesem Grunde keine Rechnung erhalten. Dieses Problem läßt sich zwar dadurch abschwächen, dass man dem stereotypen Restaurantskript weitere Skripts in Form von Ausnahmeregeln hinzufügt. Dennoch ist es praktisch unlösbar. Denn zu jeder Ausnahmesituation lassen sich weitere Ausnahmesituationen erfinden, die nicht alle in dem expliziten Textbuch eines Programms untergebracht werden können. Es kommt hier sehr schnell zur kombinatorischen Explosion der Ausnahmeregelungen. Darüber hinaus ist es völlig unklar, woher denn Schank's Programme überhaupt wissen können, welches Skript in welcher konkreten Situation zur Anwendung kommen soll. Um eine solche Entscheidung treffen zu können, müssten sie über eine Art Superskript, also ein generelles Interpretationsschema verfügen, das ihnen dabei hilft, das für die jeweilige Situation relevante Skript auswählen zu können.
Die Suche nach einem solchen allgemeinen Interpretationsschema führt indes zu einem unendlichen Regress. Die Frage, wie wir als verstehende Wesen aus einer sich ständig verändernden Umwelt die für uns relevanten Informationen extrahieren können, ist als das sogenannte Frame-Problem in die Geschichte der KI eingegangen. Ich komme später wieder darauf zurück.
Im Unterschied zu diesen technischen Überlegungen ist Searles Kritik noch viel radikaler. Ein Programm könne *prinzipiell* nicht die Bedeutung der von ihm verarbeiteten Symbole verstehen, da es nämlich nur nach syntaktischen Prinzipien arbeite und daher über keine inhärente Semantik verfüge. Nur wir als externe Nutzer geben den vom Programm verarbeiteten Symbolen eine Bedeutung. Die Bedeutung von Symbolen in einem Computerprogramm sei somit immer nur eine *zugeschriebene* Bedeutung, ähnlich den Buchstaben in einem Textbuch, die ja auch nur über einen Leser eine Interpretation erfahren. Selbst die Redeweise von *Symbolen* habe streng genommen lediglich metaphorischen Charakter, da ja die Zeichenketten in einem Computerprogramm nichts symbolisieren (vgl. Einsicht ins Ich, S. 352).
Um dies zu demonstrieren, überlegt sich Searle, wie es denn wäre, wenn ein Mensch tatsächlich nach der Arbeitsweise eines Computerprogramms vorginge. Nehmen wir an, wir seien tatsächlich in der Situation eines Programms von Roger Schank (das Programm dient, wie bereits gesagt, hier nur als anschauliches Beispiel). Auf dieser Annahme beruht sein Gedankenexperiment mit dem chinesischen Zimmer, das in facheinschlägigen Kreisen ein enormes Echo hervorgerufen hat. Searle stellt mit diesem Gedankenexperiment die Angemessenheit des in der behavioristischen Tradition stehenden Turing-Tests in Frage. Allein

auf der Grundlage des beobachtbaren Verhaltens könne nicht geprüft werden, ob ein System, das die richtigen Antworten gibt, auch tatsächlich denken kann. Prinzipiell seien zwei Systeme denkbar, die beide den Turing-Test bestehen, von denen jedoch nur eines tatsächlich Chinesisch versteht (vgl. Einsicht ins Ich, S. 344). Eine Person, die über die gleiche Ein- und Ausgabe als ein des Chinesischen mächtigen Muttersprachlers und zudem über ein Programm verfügt, das die richtige Verknüpfung zwischen Ein- und Ausgabe herstellt, verstehe keineswegs Chinesisch, selbst wenn sie in ihrem beobachtbaren Verhalten nicht von einem chinesischen Muttersprachler unterscheidbar sei. Nach Searles Auffassung sei es ausschließlich diese Person selber, die sich in einer solchen Situation befindet, die darüber entscheiden kann, ob sie nun Chinesisch verstehe oder nicht.

Nehmen wir also an, wir befänden uns in einem abgeschlossenen Raum und bekämen von außen einen Stapel mit chinesischer Schrift hereingereicht. Zu betonen ist in diesem Zusammenhang, dass die hereingereichte Schrift für den im Zimmer Eingeschlossenen nicht als solche erkennbar ist, sondern lediglich ein sinnloses Gekritzel darstellt. Streng genommen weiß er nicht einmal, dass es sich bei den chinesischen Schriftzeichen um *Symbole* handelt, die irgendwelche Dinge in der Welt repräsentieren. Nehmen wir ferner an, wir erhielten anschließend einen zweiten Packen mit chinesischer Schrift gemeinsam mit Anweisungen in unserer Muttersprache, aufgrund derer wir den ersten Stapel von chinesischer Schrift mit dem zweiten in Verbindung bringen. Diese Verbindung erfolgt nur auf Grundlage formaler Kriterien, etwa wie folgt: "Wenn Du im zweiten Packen ein bestimmtes Kritzel-Zeichen findest und dies auch im ersten Packen vorkommt, so füge dem zweiten Packen ein bestimmtes Kritzel-Zeichen des ersten Packens hinzu." Schließlich nehmen wir noch an, wir bekämen einen dritten Packen chinesischer Schrift hereingereicht, gemeinsam mit Anweisungen in unserer Muttersprache, wie wir diese Schriftzeichen mit jenen des ersten und des zweiten Packens in Verbindung bringen können. Aufgrund der letzten Anweisungen werden wir dazu aufgefordert, bestimmte Symbole nach außen zurückzugeben.

Die Quintessenz dieses Gedankenexperiments ergibt sich aus dem folgenden Umstand. Wir wissen nicht, dass die Leute, die uns die verschiedenen Packen nacheinander hineinreichen, den ersten Packen als 'Skript', den zweiten als 'Geschichte' und den dritten als 'Fragen' bezeichnen. Ferner verstehen sie unter den in unserer Muttersprache gegebenen Anweisungen ein 'Programm' und die herausgereichten Symbole deuten sie als 'Antworten auf die Fragen'. Der springende Punkt bei diesem Gedankenexperiment ist also der folgende: Von außen betrachtet, vom Standpunkt des Beobachters des im Zimmer Eingeschlossenen, sind die Antworten auf die Fragen nicht von den Antworten eines chinesischen Muttersprachlers zu unterscheiden. Dennoch wäre es - was dieses Gedankenexperiment eben demonstrieren soll - völlig verfehlt, dem im Zimmer eingeschlossenen ein Verstehen der chinesischen Schriftzeichen zuzuschreiben. Solange wir selber nach dem Arbeitsprinzip eines Computerprogramms operieren, werden wir niemals eine Ahnung davon haben, worum es sich bei den von uns manipulierten Kritzel-Zeichen handelt. Die chinesischen Symbole bleiben von unserem Standpunkt aus betrachtet - aus der Perspektive des im Zimmer Eingeschlossenen - ein unergründbares Rätsel. Aufgrund des bloßen Hantierens mit bedeutunslosen Kritzelzeichen können wir niemals deren Bedeutung erkennen.

Vergleicht man die Versuchsanordnung des chinesischen Zimmers mit jener des Turing-Tests, so zeigt sich eine verblüffende Ähnlichkeit. Auch bei Turing finden wir ein Zimmer vor, in das jemand eingeschlossen ist, desgleichen eine Verbindung zur Außenwelt über eine Lade

oder einen Fernschreiber sowie auch außenstehende Beobachter, die Fragen an den im Zimmer Eingeschlossenen stellen. Searles Pointe besteht allerdings darin, dass er Turings Prüfverfahren gewissermaßen auf den Kopf stellt. Nicht die außenstehenden Beobachter sollen darüber entscheiden, ob ein Computerprogramm im buchstäblichen Sinne verstehen kann, sondern der im Zimmer eingeschlossene Mensch, der sich so verhält wie eine Maschine. Noch eine weitere Parallele zu Turing lässt sich herstellen: Das Hantieren mit den chinesischen Schriftzeichen erinnert stark an Turings Papiermaschine, bei der ja auch ein Mensch sich strikt an die Anweisungen einer Turingmaschine hält. Eine solche Papiermaschine, so würde jedenfalls Searle argumentieren, hat aber nur eine Syntax und eben keine Semantik. Dass dem so ist, finden wir aber nicht aus der Beobachterperspektive heraus, sondern nur so, dass wir selber Papiermaschine spielen und aus dieser Mitspielerperspektive heraus entscheiden, ob wir etwas verstehen oder nicht.
Lässt man die spezifischen Details des Chinesischen-Zimmers-Gedankenexperiment einmal beiseite und reduziert sie auf die dabei verwendete Logik der Beweisführung, so kann man Searles Argumentation folgendermaßen rekonstruieren: Sie besteht im wesentlichen aus drei Axiomen und einer zentralen Schlussfolgerung. Die Axiome sind: 1. Computerprogramme funktionieren allein nach syntaktischen Prinzipien. 2. Menschliches Verstehen erfolgt auf der Grundlage geistiger Inhalte (Semantik). 3. Semantik lässt sich nicht durch Syntax erklären. Daraus ergibt sich, dass Computerprogramme menschliches Verstehen nicht erklären können.
Damit soll aber nicht behauptet werden, dass *Maschinen* nicht denken können. Schließlich seien, so Searle, wir selber nichts anderes als eben denkende Maschinen (vgl. Einsicht ins Ich, 351). Searle bestreitet lediglich vehement, dass *Programme* denken bzw. verstehen können (vgl. a.a.O., S. 352). Denn Intentionalität – die unabdingbare Voraussetzung dafür, um sich mit Symbolen auf Dinge in der Welt beziehen zu können – sei schließlich ein *biologisches* Phänomen. Damit vertritt Searle eine naturalistische Identitätstheorie von kognitiven Phänomenen und deren physikalischer Realisierung. Das Mehrebenenmodell der Intelligenz, insbesondere die These von der multiplen Instantiierbarkeit kognitiver Phänomene, ist von dieser Warte aus betrachtet ein letztes Residuum des Dualismus, der geistige Vorgänge unabhängig von materiellen Prozessen betrachtet (vgl. a.a.O., S. 355).

Die Roboter-Replik

Searles Gedankenexperiment war für die KI-Forschung eine Herausforderung und hat denn auch eine ganze Bandbreite verschiedenster Reaktionen ausgelöst. Eine dieser Repliken ist die Roboter-Replik von der Yale University, die ich hier gesondert darstellen möchte. Ich treffe diese Auswahl nicht ohne Grund. Diese Replik dient hier als Aufhänger, um zu dem im Anschluß daran behandelten *Symbol Grounding Problem* überzuleiten.
Folgen wir dem Einwand von Yale, so hat Searles Kritik nur eine beschränkte Berechtigung. Tatsächlich könne ein Computerprogramm, das strikt formale Regeln befolgt, nicht die Bedeutung der Symbolketten verstehen, die von dem Programm verarbeitet werden. Ganz anders verhielte es sich aber mit einem Computerprogramm, das in einem Roboter eingebaut ist. Dank eines solchen Roboters wäre das Computerprogramm über eine Videokamera auf der Eingabeseite und über eine motorische Steuerung von Greifarmen auf der Ausgabeseite direkt

in eine Welt eingebettet. Was Searle in seinem Gedankenexperiment dahingegen beschreibt, sei die Situation eines Programms, dem diese Einbettung in eine Welt fehle. Und es sei eben diese Einbettung in eine Welt, durch die Symbolketten eine Bedeutung bekommen.
Um den geistigen Nährboden zu verstehen, der bei diesem Einwand der Yale University eine gewisse Rolle gespielt haben könnte, müssen wir uns jene weiter oben angegebene Definition von Bedeutung wiederum in Erinnerung rufen: Bedeutung sei jenes mentale Vehikel, wodurch wir uns mit Worten auf Gegenstände in der Welt beziehen. Bedeutung hat also, um es vorerst noch ganz salopp zu formulieren, etwas mit der Bezugnahme von sprachlichen Zeichen auf Dinge in einer Welt zu tun.
Dass aber diese noch sehr vage Vorstellung keineswegs ausreicht, um erklären zu können, wie die Symbole in einem Computerprogramm zu einer inhärenten Bedeutung kommen, zeigt Searles Antwort auf die Roboter-Replik.
Ein Programm, eingebettet in einen Roboter, das über eine Videokamera Signale empfängt, ändere nämlich nicht im geringsten die Situation, die Searle in dem Chinesichen-Zimmer-Argument beschrieben hat und sei daher auch kein schlagender Einwand gegen seine Kritik an der starken KI. Was nämlich ein solches Programm tatsächlich als Eingabe erhalte, seien weitere bedeutungslose Symbole, etwa in Form eines Bitmusters. Und die Hinzufügung weiterer bedeutungsloser Symbole ändere nichts an dem Grundproblem des im Zimmer Eingeschlossenen, dass er nämlich kein Wort Chinesisch verstehe.
Wir stehen hier vor einem grundsätzlichen Problem der KI-Forschung: Wie können die Symbole eines Computerprogramms eine *intrinsische* Bedeutung erhalten, die nicht nur im Auge des Betrachters liegt (also eine nur zugeschriebene respektive geborgte Bedeutung ist), sondern sondern von dem Programm selber verstanden werden kann? Dazu müsste das Programm die Bedeutung seiner Symbole irgendwie erlernen. Ein System, das aber überhaupt keine Sprache versteht, kann die Bedeutung von Symbolen nicht über andere Symbole erlernen, die ihrerseits zuallererst interpretiert werden müssen. Daher ist auch eine Einbettung eines Computerprogramms in eine Umwelt so lange wenig hilfreich, als die Ein- und Ausgabe des Programms nach wie vor über Symbole erfolgt. Damit steht die KI-Forschung aber vor dem grundsätzlichen Problem, wie die Symbole eines symbolverarbeitenden Systems *gegründet* werden müssen, damit das System dazu in der Lage ist, selbständig (also ohne Zuhilfenahme eines externen Interpreten) deren Bedeutung zu erkennen.

The Symbol Grounding Problem

Stevan Harnad hat dieses Problem in einem 1990 erschienenen Artikel *The Symbol Grounding Problem* klar auf den Punkt gebracht. In Anlehnung an Searles Kritik an der starken KI stellt er fest: Solange Computerprogramme Zeichenketten lediglich aufgrund ihrer äußeren Gestalt (also nach ausschließlich formalen Kriterien) verarbeiten, wird die Interpretation dieser Zeichenketten stets parasitär bleiben gegenüber jenen Bedeutungen, die der Programmierer in die Zeichenketten des Programms hineinliest. In einem solchen Falle gründet die Bedeutung der Zeichenketten im Kopf eines externen Interpreten. Sie ist nur eine abgeleitete Bedeutung, keine im Programm selber inhärent verankerte Bedeutung.
Von dieser Überlegung ausgehend, greift Harnad die Frage auf, worin denn die Symbole eines nach ausschließlich syntaktischen Prinzipien arbeitenden Systems denn nun gegrün-

det werden müssen, damit ihre Bedeutung auch vom System selber unmittelbar verstanden werden könne. Dabei gilt es aber zu beachten, dass bei der Beschreibung jenes Bereichs, in dem die Symbole gegründet werden sollen, keine semantischen Termini mehr verwendet weden dürfen, wenn wir einen Zirkel in der Erklärung vermeiden wollen.
Harnad beschreibt seine Version von Searles Gedankenexperiment in zwei verschiedenen Varianten.
Nach der *ersten Variante* beherrschen wir bereits eine erste Sprache und versuchen, Chinesisch als zweite Sprache ausschließlich unter Nutzung eines Chinesisch-Chinesisch Wörterbuchs zu erlernen. Dabei drehen wir uns jedoch ständig im Kreis. Wie sollen wir eine Fremdsprache als zweite Sprache erlernen, wenn wir als einzige Informationsquelle nur ein einsprachiges Wörterbuch zur Verfügung haben? Schlagen wir beispielsweise unter einem bestimmten uns unbekannten Symbol nach, so werden wir auf ein anderes Symbol verwiesen, dessen Bedeutung uns gleichermaßen unbekannt ist. Es ist offensichtlich, daß wir auf diesem Wege nur geringe Aussichten haben, jemals die Bedeutung der chinesischen Symbole herausfinden zu können.
Vergleichen wir nun diese erste Variante mit der Situation eines Computerprogramms, das bedeutungslose Zeichenketten verarbeitet, und dabei Chinesisch als *erste* Sprache erlernen soll. Stünde einem solchen Programm nun ebenfalls nur ein Chinesisch-Chinesisch Wörterbuch zu Verfügung, so wäre diese Aufgabe für das Programm praktisch unlösbar. Dies ist die *zweite Variante* von Searles Gedankenexperiment. Während ein deutscher Muttersprachler, ausgerüstet mit einem Chinesisch-Chinesisch Wörterbuch aufgrund seiner ersten Sprache zumindestens erkennen kann, dass er hier zu dechriffrierende Symbole vor sich hat, die sich auf irgendwelche Dinge in der Welt beziehen, hat ein Programm von Hause aus ja gar keine Ahnung, dass die von ihm verabeitenden Zeichenketten etwas *symbolisieren*, also eine Bedeutung haben. Ein solches Programm müsste ja nicht nur erkennen, welche spezielle Bedeutung bestimmte Zeichenketten haben, was einem solchen Programm dahingegen von vornherein abgeht, ist ein Grundverständnis von Semantizität. Die zweite Variante von Searles Gedankenexperiment lässt sich auf den Punkt bringen: Wie lernt man Chinesisch als erste Sprache, wenn man nur ein Chinesisch-Chinesisch Wörterbuch zur Verfügung hat?
Darauf gibt es nach Harnad nur eine einzig mögliche Antwort: Um die Bedeutung chinesischer Symbole als erste Sprache erlernen zu können, müssen die Symbole irgendwie mit der Außenwelt verknüpft werden (da ja eine Erklärung dieser Symbole durch andere chinesiche Symbole in einem solchen Falle nicht möglich ist). Fraglich ist allerdings, auf welche Weise dies zu erfolgen hat. Bereits Searle hat diesen Lösungsvorschlag (in der oben erwähnten Roboter-Replik) aufgegriffen und kritisiert. Auch Harnad hält die Idee, Symbole mit einer Bedeutung auszustatten, indem man sie mit Gegenständen in der Außenwelt verknüpft, für viel zu unpräzise. Man kann sich die dabei auftretenden Probleme anhand der folgenden beiden Beispiele überlegen.
Erstes Beispiel: Vor vielen Jahren versuchte ich meinen Hund das Apportieren beizubringen. In einem Moment der Unaufmerksamkeit entging ihm, in welche Richtung ich einen Stock gerade geworfen hatte. Um ihm das Auffinden des Stockes zu ermöglichen, zeigte ich mit meiner Hand in die Richtung, in der er zu suchen hatte. Zu meiner Überraschung schaute der Hund aber statt in die angegebene Richtung von meinem Zeigefinger in Richtung meiner Schulter. Der Hund konnte offensichtlich mit der Gestik des Zeigens überhaupt nichts anfangen.

Dieses Beispiel wirft nun aber ein zweifelhaftes Licht auf alle Versuche, die Bedeutung von Symbolen über die *hinweisende Definition* einzuführen. Wollen wir beispielweise jemand die Bedeutung des deutschen Wortes „Apfel“ erklären, indem wir auf einen Apfel zeigen, so setzt auch eine solche Erklärung ein Grundverständnis von *Referenz* voraus. Die hinweisende Definition ist ihrerseits ein Zeichen, das zuallererst verstanden werden muß. Ein solches Vorverständnis können wir aber nur bei jemandem voraussetzen, der bereits eine Sprache beherrscht.
Dieser Gedankengang führt mich zum zweiten Beispiel: Nehmen wir an, wir hätten ein Bildwörterbuch, das in einer links stehenden Spalte zu erklärende Wörter und in der rechten Spalte bildliche Darstellungen von Dingen enthielte, auf die sich die Wörter beziehen sollen. In der Mitte der beiden Spalten befände sich ein Pfeilzeichen. Die Anordnung würde also in etwa so ausschauen:

Wort	Pfeilzeichen	Abgebildeter Gegenstand
„Tisch“	->	Abbildung eines Tisches
„Apfel“	->	Abildung eines Apfels
Usw.		

Mit einem solchen Wörterbuch kann allerdings nur derjenige etwas anfangen, der zumindestens die Bedeutung des Pfeilzeichens versteht. Wie kann nun aber jemandem die Bedeutung dieses Pfeilzeichens erklärt werden? Man versuche es einmal damit, das Pfeilzeichen in unserem Bildwörterbuch zu erklären, also etwa wie folgt:

Wort	Pfeilzeichen	Abgebildeter Gegenstand
„->“	->	Abbildung eines Pfeiles

Dass ein derartiger Versuch wenig zielführend ist, ist offensichtlich. Hinter diesem Beispiel steht aber ein viel weiter reichendes Problem, nämlich: Wie kann grundsätzlich die Interpretation von Zeichenketten auf einem Wege erfolgen, der nicht bereits ein Verstehen von Zeichenketten voraussetzt? Wie kann also der Zirkel in der Erklärung vermieden werden? Es schaut fast so aus, als müsse man, um eine Sprache überhaupt erlernen zu können, bereits eine Sprache verstehen (vgl. Wittgenstein). Solange wir dieses Problem nicht lösen, wird die den Symbolketten eines Programms zugeschriebene Bedeutung immer nur parasitär bleiben gegenüber einem menschlichen Interpreten (Programmierer oder Anwender des Programms).
Worum es hier letzten Endes geht, ist das folgende: Es geht darum, eine Brücke schlagen zu wollen zwischen einer phänomenologischen Beschreibung unserer intuitiven Semantik (eine Beschreibung aus der Perspektive eines Verstehenden) und möglichen naturwissenschaftlichen Erklärungen, wie wir sie unter anderem auch in der Neurophysiologie vorfinden. Die zentrale Frage ist, ob die semantische Kluft (Lenk) zwischen einer naturalistischen Erklärung vom Standpunkt eines unbeteiligten Beobachters und einer phänomenologischen Beschreibung jemals geschlossen werden kann. Mit einer positiven Antwort auf diese Frage steht und fällt der Forschungsanspruch der Kognitionswissenschaft. Anders gefragt: ist eine naturalistische Reduktion unserer intuitiven Semantik möglich? Diese strittige Frage wurde in verschiedenen Fachbereichen und wissenschaftlichen Traditionen kontroversiell diskutiert, Diskussionsbeiträge stammen aus der KI-Forschung, der Neurophysiologie, der analytischen Sprach-

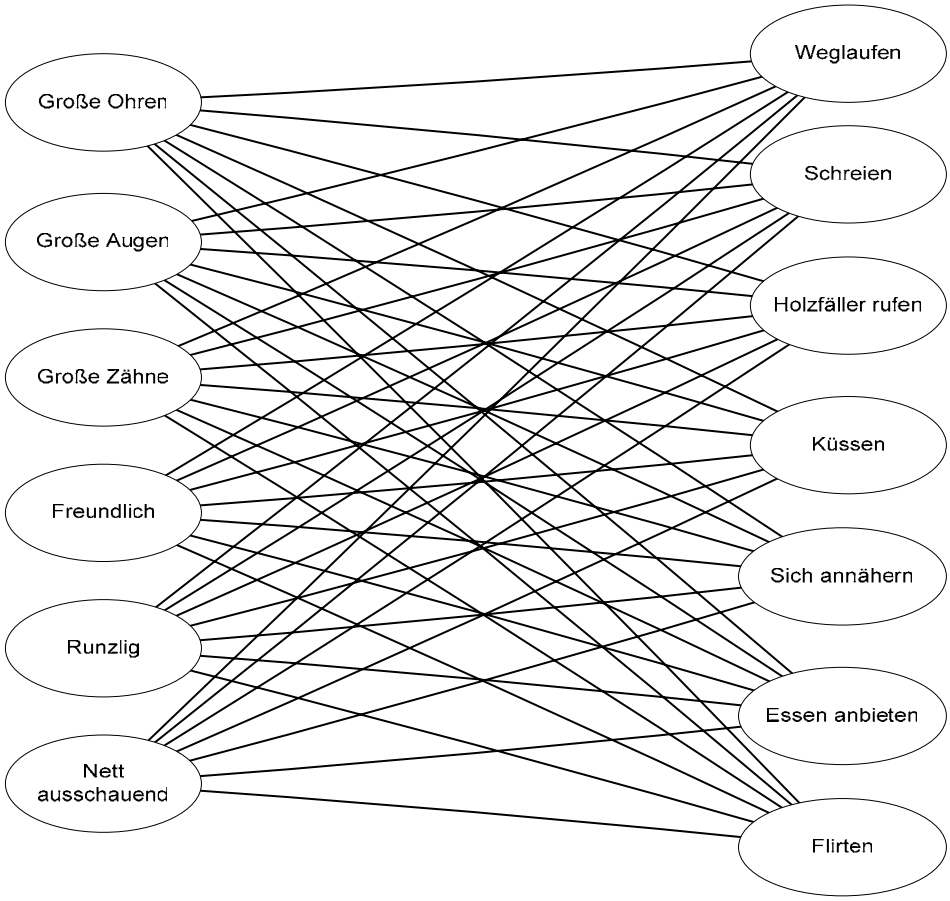
philosophie und auch der Phänomenologie. Harnads *Symbol Grounding Problem* ist davon nur ein beredtes Beispiel.
Nachdem die Symbole in einem Computersystem aus den dargelegten Argumenten nicht in ihrerseits interpretationsbedürftigen Symbolen gegründet werden können, schlägt Harnad vor, die elementaren Symbole eines Computersystems in nichtsymbolischen Repräsentationen zu gründen. Es handelt sich hier um Projektionen proximal sensorischer Reize, wobei Harnad hier zwei Arten von nichtsymbolischen Repräsentationen unterscheiden, nämlich *ikonische* Repräsentationen (beispielsweise die sich auf der Retina abzeichnende Kontur eines Pferdes) und *kategoriale* Repräsentationen. Erstere sollen dazu dienen, Reize voneinander *unterscheiden* zu können, letztere haben die Aufgabe Klassen von Stimuli *identifizieren* zu können. Zu betonen ist, dass beide Arten von nichtsymbolischen Repräsentationen nichts symbolisieren, also keine Symbole sind, die Dinge in der Außenwelt repräsentieren. Projektionen auf der Retina sind reine Beobachtungsdaten noch vor jeder möglichen Interpretation. „Iconic representations no more ‚mean' the objects of which they are the projections.." (343). Statt einer semantischen Verbindung zur Außenwelt haben wir es in diesem Falle nur mit einer rein kausalen Verbindung zu tun. Eine solche kausale Verbindung ist aber auch mit einem ausschließlich naturalistischen Vokabular beschreibbar.
Ein Computersystem, das neben den (gegründeten) Symbolen auch (gründende) nichtsymbolische Repräsentationen enthält, bezeichnet Harnad als ein hybrides System, eine Mischform von Elementen aus der neuronalen Netzwerkforschung und dem klassisch formalistischen Ansatz der KI-Forschung.
Harnads Lösungsvorschlag für das *Symbol Grounding Problem* geht also in die Richtung, die beiden Hauptströmungen der KI-Forschung, den Konnektionismus und die syntaktische Therie des Geistes, in einem System zu vereinen, wobei die Gründung von Symbolen eines nach syntaktischen Prinzipien arbeitenden Programms über eine Schnittstelle zur Außenwelt erfolgt. Anders als in der oben beschriebenen Roboter-Replik besteht die über eine solche Schnittstelle erfolgende Ein- und Ausgabe allerdings aus nichtsymbolischen Repräsentationen, die in einem neuronalen Netz realisiert sind.
Harnads Vorschlag, so bestechend er sich auf den ersten Blick liest, ist weniger einfach in der Durchführung. Man kann sich dies am Beispiel eines relativ simplen neuronalen Netzes überlegen.

Das Rotkäppchen und der böse Wolf

Wir alle kennen die Geschichte vom Rotkäppchen und dem bösen Wolf. Der Wolf lässt sich durch bestimmte Merkmale beschreiben (beispielsweise große Augen, große Ohren und große Zähne). Begegnet das Rotkäppchen einem Wesen mit diesen Merkmalen im Wald, so soll es am besten weglaufen, schreien und den Holzfäller um Hilfe rufen. Simuliert man nun ein derartiges erlerntes Verhalten in einem neuronalen Netz, so benötigen wir hierzu die folgenden Komponenten:

1. Den verschiedenen Merkmalen des Wolfes entsprechen Eingabeeinheiten, die im einfachsten Falle entweder aktiviert sind (also den Wert 1 erhalten) oder nicht (den Wert 0 erhalten).

2. Das Verhaltensmuster ‚Weglaufen, Schreien und den Holzfäller um Hilfe rufen' wird in Ausgabeeinheiten realisiert, denen ebenfalls der Wert 0 oder 1 zugeschrieben wird.
3. Benötigen wir eine Lernregel, durch die das neuronale Netz während der Lern(Trainings)phase schrittweise lernt, auf ein bestimmtes Eingabemuster (Wolfsmerkmale) mit einem bestimmten Ausgabemuster zu reagieren. Im nachstehenden Beispiel wird die so genannte Delta-Regel beschrieben.
4. Das ‚Wissen' des neuronalen Netzes steckt in den numerisch gewichteten Verbindungen zwischen den Eingabeeinheiten und den Ausgabeeinheiten. Da von Verarbeitungseinheit zu Verarbeitungseinheit nur numerisch gewichtete Aktivitäten weitergegeben werden, spricht man in diesem Falle von einer *subsymbolischen* Beschreibungsebene bzw. von einer *verteilten Wissensrepräsentation*.



Die Verbindungen zwischen den sechs möglichen Eingabeeinheiten und den sieben möglichen Ausgabeeinheiten lassen sich auch in einer Matrix darstellen, wobei den Eingabeeinheiten Zeilen entsprechen und den Ausgabeeinheiten Spalten in der Matrix. In den Verbindungsstellen von Zeilen und Spalten stehen Gewichte (w_{nn}), die die Stärke der Verbindung von Eingabeeinheit und Ausgabeeinheit numerisch quantifizieren.

	O_1	O_2	O_3	O_4	O_5	O_6	O_7
I_1	W_{11}	W_{12}	W_{13}	W_{14}	W_{15}	W_{16}	W_{17}
I_2	W_{21}	W_{22}	W_{23}	W_{24}	W_{25}	W_{26}	W_{27}
I_3	W_{31}	W_{32}	W_{33}	W_{34}	W_{35}	W_{36}	W_{37}
I_4	W_{41}	W_{42}	W_{43}	W_{44}	W_{45}	W_{46}	W_{47}
I_5	W_{51}	W_{52}	W_{53}	W_{54}	W_{55}	W_{56}	W_{57}
I_6	W_{61}	W_{62}	W_{63}	W_{64}	W_{65}	W_{66}	W_{67}

Nehmen wir fürs erste an, das System habe bereits das richtige Verhalten erlernt. Das bedeutet, dass in den Verbindungsstellen von Zeilen und Spalten bereits die richtigen Gewichte stehen. Die aktivierten Eingabeinheiten ‚Große Ohren', ‚große Augen' und ‚große Zähne' sollen also zur Aktivierung der Ausgabeeinheiten ‚Weglaufen', ‚Schreien' und ‚Holzfäller rufen' (dieses Verhalten ist angezeigt, wenn das Rotkäppchen dem Wolf begegnet) führen.
In Matrixschreibweise lässt sich dies auch so darstellen: Der Eingabevektor ‚1 1 1 0 0 0' (Erscheinung des Wolfs) soll den Ausgabevektor ‚1 1 1 0 0 0 0' auslösen.

Zur Aktivierung der Ausgabeeinheit O_1 ($O_1 = 1$) kommt es beispielsweise folgendermaßen: Wir multiplizieren die einzelnen Elemente des Eingabevektors I_i mit den jeweiligen Gewichten w_{ij} (gemeint sind die numerisch gewichteten Verbindungen zwischen den Elementen des Eingabevektors mit jenen des Ausgabevektors – in der oben angegebenen Matrix rot markiert) und addieren die Produkte:

$$\sum_{i=1}^{6} I_i \cdot w_{ij}$$

Ist also j = 1 (der Laufindex j bezieht sich auf die Spalten der Matrix) wie im Falle der Ausgabeeinheit O_1 so gilt:

$$I_1 \cdot w_{11} + I_2 \cdot w_{21} + I_3 \cdot w_{31} + I_4 \cdot w_{41} + I_5 \cdot w_{51} + I_6 \cdot w_{61}$$

Sind I_4 bis I_6 gleich 0, wie dies im Falle des Wolfes gegeben ist, so ist die Produktsumme dieser Gewichte gleich 0! Da in diesem Falle I_1 bis I_3 gleich 1 sind, ist die Produktsumme von I_1 bis I_3 mit den jeweiligen Gewichten die Summe der Gewichte!
Das Ausgabeverhalten „Weglaufen" ($O_1 = 1$) wird nun so aktiviert: Überschreitet die Produktsumme einen bestimmten Schwellwert, beispielsweise 0.7, dann wird O_1 aktiviert (O_1 ist dann gleich 1)
Diese Aktivierung setzt freilich voraus, dass das Rotkäppchen bereits das richtige Verhalten gelernt hat.

Damit unser neuronales Netz auf die Eingabeinheiten ‚Große Ohren', ‚große Augen' und ‚große Zähne' mit der Aktivierung der Ausgabeeinheit ‚Weglaufen' reagiert, müssen in unserer Matrix lediglich die richtigen Gewichte stehen (beispielsweise jeweils 0.3). Völlig offen blieb in diesem Beispiel allerdings, wie dem System diese Gewichte antrainiert werden können. Dies ist erst der eigentliche Lernvorgang.
Vor dem Lernvorgang ist es nämlich so, dass alle Gewichte in der Matrix in der Nähe des Wertes 0 randomisiert sind. Dadurch ist am Beginn des Lernvorgangs die Produktsumme von Gewichten und aktivierten Eingabeeinheiten unter dem in unserem Beispiel angenommenen Schwellwert von 0.7. Dies führt wiederum dazu, dass beim ersten Lernzyklus selbst bei Aktivierung der ersten 3 Eingabeeinheiten O_1 bis O_3 gleich 0 sind und eben nicht, wie erwartet, auf 1 gesetzt werden.

In diesem Falle müssen die Gewichte schrittweise so erhöht werden, dass das System das richtige Verhalten zeigt. Diese Änderung der Gewichte ist das eigentliche Lernen des Systems.
Eine einfache Rechenvorschrift, wie eine solche Änderung der Gewichte erfolgen kann, gibt uns die so genannte delta Regel. Wir benötigen hierzu die Differenz des erwarteten Ausgabevektors mit dem aufgrund der Gewichte errechneten Vektor. Bezeichnen wir j als den Index für eine Ausgabeaktivität, so ist Tj – Oj die Differenz zwischen dem j-ten Element des erwarteten Ausgabevektors Tj und dem j-ten Element des beobachteten Ausgabevektors für die Ausgabeeinheit Oj. Bevor das System das richtige Verhalten gelernt hat, wird Tj = 1 sein (wir erwarten ‚Weglaufen') und Oj = 0 sein (wir bekommen vor dem Lernvorgang kein ‚Weglaufen').
Die Gewichtsänderung (das Delta) nenne ich delta w_{ij}. Sie wird wie folgt berechnet:

$$\Delta_{ij} = n \cdot (T_j - O_j) \cdot I_i$$

N ist die so genannte Lernrate. Sie sei beispielsweise auf 0.1 gesetzt.
Betrachten wir nur jene Gewichte, die zur Aktivierung der für das Weglaufen zuständigen Ausgabeeinheit (O_1 =1) führen, so ergibt sich das folgende:

$$\Delta_{11} = 0.1 \cdot (1 - 0) \cdot 1 = 0.1$$
$$\Delta_{21} = 0.1 \cdot (1 - 0) \cdot 1 = 0.1$$
$$\Delta_{31} = 0.1 \cdot (1 - 0) \cdot 1 = 0.1$$
$$\Delta_{41} = 0.1 \cdot (1 - 0) \cdot 0 = 0$$
$$\Delta_{51} = 0.1 \cdot (1 - 0) \cdot 0 = 0$$
$$\Delta_{61} = 0.1 \cdot (1 - 0) \cdot 0 = 0$$

Das jeweilige Delta wird nach dem ersten Lernzyklus zu den bereits bestehenden Gewichten dazu addiert.
Nach dieser Gewichtsänderung sind die neuen Gewichte, die von I_1 bis I_3 zu O_1 führen, 0.1 (unter der grob vereinfachenden Annahme, dass die randomisierten Gewichte zuallererst auf nahezu 0 gesetzt worden sind). Berechnet man aufgrund der neuen Gewichte wiederum das Ausgabeverhalten, so bekommen wir bei einem Schwellwert von 0.7 immer noch ein O_1 von

0. Der Vorgang wird solange wiederholt, bis $0_1 = 1$ ist! Dies wird durch eine schrittweise Erhöhung der Gewichte erzielt.

Was kann uns jetzt aber dieses kleine Beispiel eines neuronalen Netzes über Harnads *Symbol Grounding Problem* verraten? So fällt auf den ersten Blick auf, dass an keiner Stelle in diesem Netz ein Symbol für ‚Wolf' erkennbar ist. Was also in dem Beispiel augenscheinlich fehlt, ist eine komplexere symbolische Ebene, die zuallerst in nichtsymbolischen Verarbeitungseinheiten gegündet werden könnte. Eine solche symbolische Ebene könnte nach dem Vorschlag von Konnektionisten durch die Einführung von verborgenen Einheiten (so genannter hidden units) zwischen den Ein- und Ausgabeschichten realisiert werden. In einem Multilayer-System werden die Gewichtsänderungen in einer Erweiterung der Delta-Regel Schicht für Schicht rückwärtsgerechnet. Eine solche Erweiterung der Delta-Regel ist das Back-Propagation-Verfahren. Ein derartiger Hidden Layer enthielte beispielsweise eine Verarbeitungseinheit für ‚Wolf', eine andere für ‚Großmutter' usw. Inwiefern sich aber mit solchen Symbolen in einem neuronalen Netz eine kompositionale Semantik realisieren ließe (elementare Symbole definieren komplexe Symbole, die ihrerseits noch komplexere Symbole definieren), ist ein in der Literatur kontroversiell diskutiertes Thema (vgl. Smolenski,..).
Unser Rotkäppchenbeispiel deckt aber ein noch viel größeres Problem auf, nämlich: Sowohl Ein- als auch Ausgabeeinheiten haben immer noch eine *symbolische* Bedeutung (so bedeutet die Aktivierung der ersten Eingabeeinheit „Große Ohren" und der ersten Ausgabeeinheit ‚Weglaufen' usw.). Die Lösung des *Symbol Grounding Problem* verlangt aber eine *subsymbolische* Ebene, also eine verteilte Wissensrepräsentation sowohl auf der Seite der Eingabeeinheiten als auch auf der Seite der Ausgabeeinheiten! Dies wäre nur dann der Fall, wenn – wie in der Mustererkennung – die Aktivierung der Eingabereize bereits auf nicht symbolische Weise erfolgt. Die meisten neuronalen Netze erhalten ihre Informationen stattdessen in symbolischer Form, verarbeiten diese Informationen zwar intern im Sinne einer verteilten Wissensrepräsentation, um dann aber in der Ausgabe wieder Informationen in symbolischer Form zu produzieren.
Es scheint nicht ganz einfach zu sein, die Ebene der nichtsymbolischen Repräsentationen in dafür geeigneten Worten zu beschreiben. Harnads eigene Beispiele sind in dieser Hinsicht auch nicht sehr hilfreich. Wenn das Symbol für Zebra über die elementaren Symbole ‚Pferd' und ‚Streifen' eingeführt wird, wobei letztere in ikonischen und kategorialen Repräsentationen gegründet werden sollen, so ist dies im besten Falle ein Vorschlag, der aber erst erklärt werden müßte.
Auf welche Weise also Symbole einer Sprache gegründet werden könnten, ohne dass dabei das Verstehen von Symbolen bereits vorausgesetzt werden muß, wie und ob überhaupt die semantische Kluft zwischen phänomenologischer Beschreibung und naturalistischer Erklärung geschlossen werden könnte, ist nach wie vor eine offene Frage.

Wir stehen somit nach wie vor vor dem bislang ungelösten Problem: Ist eine naturalistische Erklärung unserer intuitiven Semantik überhaupt möglich? Kann die Bedeutung von Symbolen reduziert werden auf Beobachtungsdaten noch vor jeder Interpretation? Gibt es dagegen etwas kritisch einzuwenden? Die Antwort darauf ist ein deutliches Ja. So hat bereits Jahrzehnte vor Harnads Vorschlag der amerikanische Sprachphilosoph und Logiker Quine in einem Gedankenexperiment zu zeigen versucht, dass unsere intuitive Semantik, was ihre Unterfütterung durch Sinnesdaten anbelangt, prinzipiell unbestimmt sei. Dies ist die These von der *Unbestimmtheit der Übersetzung*. Quines Gedankenexperiment lässt sich im Wesentlichen so beschreiben: Ein Sprachforscher wird zu einem Eingeborenenvolk geschickt, um ausschließlich auf der Grundlage von Beobachtungsdaten dessen Sprache zu erlernen. Wesentlich an diesem Gedankenexperiment ist das Fehlen eines Dolmetschers als Übersetzungshilfe. Denn dadurch wird sichergestellt, dass das Erlernen der Eingeborenensprache tatsächlich nur unter Nutzung beobachtbarer Sinnesreize erfolgt. Quine spricht in diesem Zusammenhang auch von einer „radikalen Übersetzung“ (63). Hätte der Sprachforscher einen Dolmetscher dabei, der des Deutschen oder des Englischen mächtig wäre, so wäre diese Situation vergleichbar mit dem Erlernen einer Fremdsprache durch ein zweisprachiges Lexikon. In Quines Gedankenexperiment sind dahingegen nur nonverbale Stimuli die einzigen Übersetzunghilfen. Interessanterweise spricht Quine in diesem Zusammenhang von "Mustern der Bestrahlung des Auges" (67). Dahinter steht ein ganz ähnlicher Gedankengang, wie ihn dann später Harnad über seine ikonischen Repräsentationen aufgegriffen hat, nämlich die Bedeutung von Symbolen herausfinden zu wollen, indem man sich ausschließlich Muster der Bestrahlung des Auges bedient.
Huscht also ein Kaninchen vorbei und ein Eingeborener sagt beispielsweise "Gavagai" (63), so notiert sich der Sprachforscher als erste Übersetzungshypothese den Satz "Sie da, ein Kaninchen" (63).
Dabei stellt sich nun aber das folgende Problem: Allein unter Zuhilfenahme derartiger Stimuli ist es dem Sprachforscher unmöglich, herauszufinden, ob mit "Gavagai" zeitliche Kaninchenstadien, Kaninchenteile oder ganze Kaninchen gemeint sind (101f.). Aber damit nicht genug: Selbst dass es sich überhaupt um Kaninchen handelt, lässt sich auf der Ebene nonverbaler Stimuli nicht eindeutig entscheiden. Nehmen wir an, es gäbe am Ort der Eingeborenen eine dem Sprachforscher unbekannte Kaninchenfliege, die bereits von weitem an ihren langen Flügeln und ihren ruckartigen Bewegungen zu erkennen sei und die den Eingeborenen das Erkennen eines nur schwach wahrgenommenen Kaninchens erleichtere (77). Sagt ein Eingeborener in einem derartigen Falle "Gavagai", so würde der Sprachforscher den wahrgenommenen Reiz (schwaches Bestrahlungsmuster eines Kaninchens verbunden mit starkem Bestrahlungsmuster von Kaninchenfliege) nicht seiner Übersetzungshypothese "Sie da, ein Kaninchen" zurechnen.
Daraus lässt sich zunächst das folgende schließen: Unter *unveränderten* Reizbedingungen sind *verschiedene* Übersetzungen eines sprachlichen Ausdrucks und in der Folge davon auch unterschiedliche Bedeutungszuweisungen möglich. Dies bedeutet ferner, dass die Bedeutungen sprachlicher Ausdrücke nur relativ zu bestimmten linguistischen Grundannahmen einer Sprechergemeinschaft - so genannten "analytischen Hypothesen" (129-136) - fixiert werden können. Es sind *verschiedene* Übersetzungsmanuale denkbar, die von der *gleichen*

sinnlichen Evidenz als 'objektiver Basis' ausgehen und dennoch wechselseitig unverträglich sind.
Damit ist aber Harnads Versuch, die Bedeutung elementarer Symbole über nichtsymbolische Repräsentationen erklären zu können, vor dem Hintergrund von Quines These von vornherein zum Scheitern verurteilt. Die Bedeutung von Symbolen lässt sich nicht zurückführen auf die Ebene des beobachtbaren Verhaltens. Semantik ist und bleibt hinsichtlich ihrer sensorischen Reizgrundlage unterbestimmt.
Diese These, so unbestreitbar richtig sie für sich allein betrachtet auch immer sein mag, stellt nun für Quine aber nicht das Endresultat seiner Überlegungen dar, sondern markiert den Ausgangspunkt für zwei gänzlich verschiedene und miteinander unverträgliche Schlußfolgerungen. Sie kann nämlich auf zwei verschiedene Arten interpretiert werden.
Nach der ersten Deutung enthalten semantische Termini eine eigenständige, nicht in die Sprache der Naturwissenschaften überführbare Beschreibungsebene. Dieser Auffassung zufolge ist also die Unbestimmtheit der Übersetzung ein Argument zugunsten einer naturalistisch nicht erklärbaren Semantik. Dies ist aber nur *eine* mögliche Schlussfolgerung aus Quines These. Eine andere, zur ersten diametral entgegengesetzte Schlussfolgerung erklärt sich die sensorische Unterbestimmtheit unserer intuitiven Semantik aus der Zweideutigkeit unserer Alltagssprache, die aus einer streng normierten Wissenschaftssprache eliminiert werden müsste.
Quine bezeichnet die erste Deutung der Unbestimmtheit der Übersetzung auch als Brentanos These und beschreibt die beiden alternativen Schlussfolgerungen folgendermaßen:
"Man kann Brentanos These akzeptieren und entweder so verstehen, dass sie die Unabdingbarkeit der intentionalen Ausdrucksformen und die Bedeutung einer eigenständigen Wissenschaft von den Intentionen zeigt, oder aber so, dass sie die Grundlosigkeit der intentionalen Ausdrucksformen und die Gehaltlosigkeit einer Wissenschaft von den Intentionen zeigt. Im Gegensatz zu Brentano bin ich der letzteren Auffassung." (381)
Quine selber verwendet die These von der Übersetzungsunbestimmtheit also als indirekten Beweis *gegen* den Mentalismus. Die Nichtreduzierbarkeit semantischer Ausdrucksformen in die Sprache des Naturalisten wird als Argument verwendet, um deren ontologische Grundlosigkeit zu demonstrieren. Diese Überlegung folgt gewissermaßen dem Motto, ist die Semantik unbestimmt, so sei dies umso schlimmer für die Semantik. "Semantik ist so lange von einem schädlichen Mentalismus verderbt, wie wir die Semantik eines Menschen als irgendwie in seinem Geiste über das hinaus festgelegt ansehen, was in seinen Dispositionen zu beobachtbarem Verhalten enthalten sein könnte". (1975: 42 f.)
Was die ultimative Struktur der Realität betrifft, sind semantiche Ausdrucksformen daher ein überflüssiges Beiwerk (vgl. 1980: 382) Quine bekennt sich zum *eliminativen Materialismus*: "Wollen wir die wahre und letzliche Struktur der Realität darstellen, ist das schmucklose Schema das für uns kanonische, also dasjenige, das keine Zitierung kennt außer der direkten Rede und keine propositionalen Einstellungen außer der physischen Konstitution und dem physischen Verhalten der Organismen." (382)
Quine räumt semantischen Ausdrucksformen lediglich den Status einer dramaturgischen Geste ein (378), die uns dabei behilflich ist, sich in die Überzeugungen und Wünsche unseres Gegenüber hineinzuversetzen und mit deren Hilfe wir die verschiedenen miteinander unverträglichen Übersetzungsmanuale normieren können.

Eine Zwischenbilanz

Mit den beiden von Quine beschriebenen alternativen Interpretationen seiner These von der Unbestimmtheit der Übersetzung schließt sich der Kreis unserer Überlegungen und wir sind wieder dort angekommen, wovon wir ganz am Anfang bei der Vorstellung des Mehrebenen-Modells der Intelligenz ausgegangen sind. Dieses Modell soll nämlich, so wurde jedenfalls behauptet, eine echte Alternative darstellen zu zwei Traditionslinien, die beide je für sich betrachtet nur unzureichend kognitive Phänomene erklären können. Ich denke hier an den Dualismus, wie er uns noch im alten Mentalismus der Würzburger Schule begegnet und dem Behaviorismus. Es sind diese beiden Traditionslinien, die Quine als einzig zulässige Schlussfolgerungen aus seiner These von der Übersetzungsunbestimmtheit im Blick hat.
So erwies sich der Dualismus als inkompatibel mit unserem naturwissenschaftlichen Weltbild, das physikalisch kausal geschlossen ist. Reklamiert man für intentionale Ausdrucksformen einen auch ontologisch eigenständigen Wissenschaftsbereich (wie dies Brentanos These entspricht), so kann nicht erklärt werden, wie denn diese intentionalen Ausdrucksformen mit unserem physikalisch beobachtbaren Verhalten interagieren können. Auf der anderen Seite widerspricht der Behaviorismus, der ja nur statistische Korrelationen zwischen Sinnesreizen und beobachtbarem Verhalten ermitteln kann, unserer Alltagspsychologie. Alltagspsychologisch betrachtet schreiben wir unserem Gegenüber intentionale Einstellungen zu, um damit eben dessen Verhalten erklären zu können.
Im Mehrebenen-Model wird stattdessen der Versuch unternommen, unseren alltagspsychologischen Vorstellungen entgegenzukommen, ohne dabei aber das naturwissenschaftliche Weltbild zu verletzen. Dies geschieht durch die Nutzung zweier Aspekte des Schichtenmodells, nämlich des Prizips der multiplen Instanziierung und jenes der Übersetzbarkeit. Aufgrund des Prinzips der multiplen Instanziierung sind kognitive Prozesse unabhängig von ihrer physikalischen Realisierung beschreibbar. Wenn Searle also dem Mehrebenen-Modell eine gewisse Nähe zum Dualismus attestiert, so hat er insofern recht, als diesem Modell zufolge kognitive Prozesse eben nicht auf ganz bestimmte physikalische Trägerprozesse reduziert werden können. Das Prinzip der multiplen Instanziierung rechtfertigt zuallererst eine eigenständige Wissenschaft von der Kognition unabhängig von ihrer empirischen oder technischen Realisierung (also unabhängig von Kognitionspsychologe und Artificial Intelligence).
Dem Prinzip der Übersetzbarkeit zufolge sind kognitive Prozesse aber nur eine andere Beschreibungsebene der zugrunde liegenden physikalischen Trägerprozesse und in letztere auch übersetzbar. Kognitive Prozesse sind somit auch naturalistisch beschreibbar, womit wiederum einer dualistischen Sichtweise eine klare Absage erteilt werden kann. Wir haben gemäß dem Mehrebenen-Modell nur in einem eingeschränkten Sinne einen Realismus der intentionalen Einstellungen.
Das Modell bietet sich somit als eine Art Reparaturhilfe an, um konzeptionelle Engpässe und Erklärungsnotstände des klassischen Dualismus aber auch des Behaviorismus ausräumen zu können. Wir dürfen dabei aber nicht übersehen, dass es sich bei den Forschungsparadigmen dieses Modells zunächst einmal um Annahmen handelt, die sich zuallererst in der Forschungspraxis dann auch bewähren müssen. So ist es ja gerade das Prinzip der Übersetzbarkeit semantischer Ausdrucksformen, das durch Quines These in Frage gestellt wird. Damit ist aber auch fraglich, inwiefern das Mehrebenen-Modell den Dualismus tatsächlich aushebeln

und unsere alltagspsychologischen Vorstellungen einer naturalistischen Betrachtungsweise zugänglich machen kann.
Probleme der Übersetzbarkeit von Semantik in Physik haben dann auch dazu geführt, die Grundannahmen der komputationalen Theorie des Geistes in einem kritischen Licht zu überdenken und sie durch andere Forschungsparadigmata, verbunden mit neueren ‚turns' in der Cognitive Science, abzulösen. Dies ist jedoch eine andere Geschichte.

Supervenienz

Fassen wir die im vorangehenden Kapitel behandelten Probleme in der folgenden Frage zusammen: Warum orientiert sich sich die Cognitive Science in ihrem terminologischen Apparat hauptsächlich an der formalen Beschreeibungsebene? Warum glaubt sie, die semantische und die physikalische Beschreibungsebene aus ihren Betrachtungen ausklammern zu können?
Entscheidend für die Beantwortung dieser Frage ist das Verhältnis zwischen den verschiedenen Beschreibungsebenen, die gemäß der komputationalen Theorie des Geistes als ineinander übersetzbare Schichten gedeutet werden.
Aufgrund des Prinzips der multiplen Instanziierung kann ein und derselbe Vorgang auf der Softwareebene in verschiedenen Vorgängen auf der Hardwareebene realisiert werden. Wir können daher bei der Beschreibung mentaler Prozesse deren physikalische Realisierung ausklammern. Die physikalische Realisierung ist eine Frage der Performance, nicht aber eine Frage des Prinzips. Dieser Grundgedanke ist maßgeblich für die technische Umsetzung einer künstlichen Intelligenz.
Warum aber wird die semantische Ebene für eine Beschreibung mentaler Prozesse nicht benötigt? Dies wiederum erklärt sich durch die Doktrin der *Übersetzbarkeit* von Semantik in Syntax. Es handelt sich hierbei um das bereits mehrfach angesprochene Formalistenmotto des Kognitivismus, nämlich *syntax mirrors semantics*. Semantische Eigenschaften mentaler Zustände und Prozesse spiegeln sich sozusagen in deren syntaktischen Strukturen. Diese Spiegelung wird mit dem technischen Ausdruck *Supervenienz* bezeichnet. Auch mit diesem Prinzip wird das Verhältnis zweier Schichten untereinander beschrieben, in diesem speziellen Falle handelt es sich um das Verhältnis zweier Schichten von *unten nach oben* betrachtet. Aufgrund der Übersetzbarkeit von Semantik in Syntax ergibt sich nämlich: Ein und derselbe Prozeß auf der syntaktischen Beschreibungsebene zieht auch ein und denselben Prozeß auf der semantischen Ebene nach sich. Das Verhältnis zweier Schichten von unten nach oben betrachtet darf also nicht zweideutig sein, wenn man jedenfalls von einer Supervenienz einer höheren Schicht über einer tieferen Schicht sprechen will.
Man betrachte hierzu die folgende knappe Definition von Supervenienz nach Fodor: „States of type X supervene on states of type Y iff there is no difference among X states without a corresponding difference among Y states." (Fodor, Psychosemantics, p. 30)
Damit wird zum Ausdruck gebracht, dass höhere Beschreibungsebenen durch tiefere Beschreibungsebenen determiniert sind, respektive auf letztere reduziert werden können. Sagt man also, die semantische Beschreibungsebene superveniere über der syntaktischen Ebene, so ist mit diesem Ausdruck nichts anderes als das Formalistenmotto gemeint, demzufolge das Semantische eine Spiegelung des Syntaktischen sei. Wenn ich also kognitive Prozesse voll-

ständig formal beschreibe, so habe ich demzufolge auch erschöpfend deren Interpretation (mit-)beschrieben.
Es war Hilary Putnam, der zunächst in den 50er Jahren des vergangenen Jahrhunderts als einer der Mitbegründer des Funktionalismus und damit der komputationalen Theorie des Geistes in Erscheinung trat, von dem dann jedoch in weiterer Folge ein folgenschweres und viel diskutiertes Gedankenexperiment entwickelt wurde, welches sich als schlagendes Argument gegen die Supervenienz der Semantik über die Syntax herausstellte. Es handelt sich hierbei um das Gedankenexperiment von der sogenannten Zwillingserde.
Man kann dieses Gedankenexperiment von zwei Seiten betrachten. Auf der einen Seite kann man die innere Logik der dort vorgebrachten Argumente analysieren. Dies läßt sich relativ kurz darstellen. Ungleich schwieriger dagegen ist eine Darstellung der epistemologischen und semantiktheoretischen Voraussetzungen bzw. Implikationen dieses Gedankenexperiments. Bevor ich daher auf die technischen Details eingehe, möchte ich quasi als Vorbereitung auf Putnams Gedankenexperiment eine kurze Anekdokte vorbringen, die mir als didaktischer Einstieg in die schwierigen Analysen Putnams geeignet erscheint. Es handelt sich hierbei um eine Geschichte zur perspektivischen Verkürzung, wie sie in der abendländischen Kultur seit der Renaissance üblich ist.

Die Perspektive im Spiegel der Kulturen

Wir alle kennen die in der Renaissance wiederentdeckte *Zentralperspektive*, die unter anderem gleich große Objekte bei unterschiedlicher Entfernung in verschiedenen Größen darstellt (weiter entfernte Objekte werden kleiner gezeichnet). Im Unterschied dazu verwendete das Mittelalter die sogenannte *Bedeutungsperspektive*. Personen und Gegenstände wurden damals in verschiedenen Größen je nach ihrer unterschiedlichen sozialen Bedeutung wiedergegeben.
Betrachten wir nun Gemälde der alten Meister aus der Zeit der Renaissance, so entsteht der Eindruck, als wäre Räumlichkeit ein selbstverständliches Merkmal der bildlichen Darstellung als solchen, so als *müsse* das Gemälde räumlich sein, als ob also die bildliche Darstellung zwingend Räumlichkeit vorschreibe. Räumlichkeit wäre demzufolge keine Eigenschaft, die zuallererst von einem Interpreten in das Bild hineingelesen wird, sondern intrinsich in dem Bild selber verankert. Man muß Ihnen beispielsweise nicht erklären, „der kleine Mann am oberen Rand des Bildes ist deshalb kleiner, weil er weiter entfernt ist“, sondern der kleine Mann wird unmittelbar und quasi automatisch als weiter entfernt wahrgenommen.
Trotzdem möchte ich die Frage stellen: Wo steckt nun wirklich dieses Verständnis von Räumlichkeit? Steckt es tatsächlich im Bild oder nicht vielmehr im Auge des Betrachters? Haben wir es hier mit einer Information zu tun, die völlig unabhängig vom Kontext verschiedener Kulturen direkt mit der bildlichen Darstellung verbunden ist?
Stellen Sie sich hierzu folgende Geschichte vor: Ein Kind sagt zu seinem Vater, „Schau, schau, was für ein kleiner Mann steht dort oben auf dem Hügel“. Diese kleine Demonstration zeigt deutlich, dass nicht einmal an der Projektion auf unserer Retina unmittelbar die Entfrenung von Gegenständen abgelesen werden kann, sondern zuallererst erlernt werden muß. Ausgehend von Quines Dokrin der Bedeutungsunbestimmtheit könnte man auch sagen, die Entfernung von Objekten sei keine Information, die unmittelbar im Bestrahlungsmuster der Retina enthalten ist, sondern eine Information, die in ein solches Muster erst hineingelesen

wird. Dieser Umstand liegt an der referentiellen Unbestimmtheit derartiger Bestrahlungsmuster. Und in eben diesem Sinne ist auch die perspektivische Verkürzung kein intrinsiches Merkmal des Bildes selber, sondern eben eine Information, die zuallererst im Kontext einer Kultur erlernt werden muß.
Vergleichen wir dazu die Funktion unterschiedlicher Größendarstellungen in der chinesischen Malerei. Ähnlich wie im Europa des Mittelalters, bedeuten dort Menschen, die in verschiedener Größe dargestellt werden, nicht Entfernungsunterschiede, sondern markieren einen jeweils verschiedenen sozialen Status. Größenunterschiede werden also im *Kontext* der chinesischen Kultur ganz anders interpretiert, wie dies einem Europäer geläufig ist. An dieser Gegenüberstellung zwei verschiedener Kulturen hinsichtlich der Deutung von Größenunterschieden in blidnerischen Darstellungen zeigt sich in aller Deutlichkeit, dass die Interpretation der perspektivischen Verkürzung als Räumlichkeit eben kein intrinisches Merkmal des Bildes ist. Auf den Punkt gebracht bedeutet dies: Nicht das Gemälde selbst enthält intrinisch die objektive Eigenschaft Räumlichkeit, sondern nur wir selber sind es, die die perspektivische Verkürzung als Räumlichkeit interpretieren.
Diese Interpretation erfolgt vor dem Hintergrund unserer Kultur. Das Gemälde selber ist ein Artefakt und als solches hinsichtlich der Bedeutung seiner unterschiedlichen Größendarstellungen *unbestimmt*. Es ist dieselbe Bedeutungsunbestimmheit, die auch in Quines berühmtem Gavagai-Beispiel zum Ausdruck kommt. Auch das durch ein vorbeihuschendes Kaninchen ausgelöste Bestrahlungsmuster auf der Retina ist hinsichtlich der Bedeutung von „Gavagai“ unbestimmt. Bei gleichem Bestrahlungsmuster sind schließlich verschiedene Übersetzungen des Wortes „Gavagai“ möglich. Und am Rande sei erwähnt, dass Quine unter der Bedeutung eines Satzes (beispielsweise des Einwortsatzes „Gavagai“) das versteht, was ein Satz mit seiner Übersetzung gemeinsam hat (vgl. Quine, Wort und Gegenstand, S. 69).
Welche Konsequenzen diese Überlegungen haben, kann man sich am folgenden konstruierten Fall überlegen. Ein europäisches Gemälde wurde nach China verschickt. Stellen wir uns vor, in China zeige ein Chinese auf einen kleinen Mann in dem Gemälde und kommentiere dessen Größe wie folgt: „Schau, was für eine unbedeutende Person!“. Die entscheidende Frage ist nunmehr, ob unser Chinese einen Fehler gemacht hat. Ist seine Aussage also falsch? Die Antwort im Anschluß an die obigen Überlegungen ist ein klares Nein! Warum aber ist seine Antwort nicht falsch? Nehmen wir nur für einen Augenblick mal an, die über die perspektivische Verkürzung erzielte Raumwirkung wäre ein objektives Merkmal des Gemäldes selber. Nur in diesem speziellen Falle könnte man mit Fug und Recht den Ausruf unseres Chinesen als falsch einstufen. Nun handelt es sich bei unserem nach China verschickten Gemälde unterdessen um ein Artefakt, dessen bildliche Darstellungen keine intrinsische Bedeutungen haben und daher auch nicht von sich aus räumliche Gegebenheiten widerspiegeln können. Die vom europäischen Betrachter in Objekte verschiedener Größe hineingelesene Entfernung ist eine Bedeutungszuschreibung, deren Gültigkeit vom Kontext einer bestimmten Kultur, in diesem Falle der abendländischen Kultur abhängt. Kleinere Objekte bedeuten daher auch nicht *wirklich* entferntere Objekte.
Stellen wir uns nun ein etwas anderes Szenario vor. Nehmen wir, unser Chinese kommt bei der Betrachtung des Gemäldes zur Auffassung, es wäre in der Intention des europäischen Malers gelegen, Personen mit geringerem sozialen Status kleiner darzustellen. Der europäische Maler hätte also mit der Größe des abgebildeten Mannes einen niedereren sozialen Status *gemeint*. Hätte er auch in einem solchen Falle keinen Fehler gemacht? Dieser zweite Fall unter-

scheidet sich indes deutlich vom ersteren. Während sich nämlich im ersteren Falle die Aussage auf ein Artefakt (die zeichnerische Darstellung eines Objekts, eines Zeichens, eines Symbols) bezieht, wird im letzteren Falle eine Aussage getroffen über die Intention des Malers selber. Im letzteren Falle hätte unser Chinese, so möchte ich jedenfalls behaupten, in der Tat einen Fehler gemacht, u.zw. aus dem folgenden Grund: Nicht die Zeichnung selber bedeutet *wirklich* dieses oder jenes (räumlich-geoemtrische Gegebenheiten, sozialer Status usw.), sondern nur wir selber meinen mit der Zeichnung *wirklich* dieses oder jenes. Dies liegt an dem Umstand, dass unsere Intentionalität eben keine abgeleitete, sondern eine ursprüngliche, intrinsich in unserem Denken und Handeln verankerte ist. Schließlich sind wir keine Artefakte, denen erst von außen eine Bedeutung zugeschrieben wird, sondern autonom denkende Wesen. Wir sind gewissermaßen diejenigen, die uns eine bedeutungsvolle Welt zuallererst genuin erzeugen. Was wir also meinen, ist nicht relativ zu einer jeweiligen Kultur, sondern behält seine Gültigkeit *in allen möglichen Kontexten*. Aus diesem Grunde macht der Chinese einen Fehler, wenn er dem europäischen Maler unterstellt, mit dem kleiner dargestellten Mann einen Mann mit geringerem sozialem Status *gemeint* zu haben.
Lassen Sie mich den Punkt hervorheben, um den es mir in diesem Zusammenhang geht: Es geht mir an dieser Stelle allein um eine schlichte Explikation unserer intuitiven Vorstellung von einer ursprünglichen Bedeutung im Unterschied zu einer abgeleiteten Bedeutung. Und ursprünglich etwas bedeuten heißt eben, dies ist die Quintessenz der obigen Überlegungen, eine Bedeutung zu haben, die in allen möglichen Kontexten gleich bleibt. Alle weitergehenden Überlegungen – einschließlich der Frage, welchen Sinn es überhaupt macht, dem menschlichen Denken eine ursprüngliche Bedeutung zuzusprechen – seien daher, jedenfalls für den Augenblick, ausdrücklich ausgeklammert.
Von hier ausgehend ist es ein nicht mehr allzu großer Schritt zu Putnams Gedankenexperiment von der Zwillingserde, auf das ich im Folgenden in mehreren Varianten eingehen möchte.

Putnams Zwillingserde I (Das Gedankenexperiment)

Stellen Sie sich vor, irgendwo im Universum gäbe es ein Duplikat zu unserer Erde, die oberflächlich betrachtet in allen Merkmalen unserem Planeten ähnlich ist. Auf dieser Zwillingserde, nennen wir sie kurz die Zwerde, gäbe es also die gleichen Kontinente, die gleichen Länder, die gleichen Bewohner usw. Auf der Zwerde gäbe es auch ein Pendant zu unserem Europa mit einer kompletten Kopie von Österreich. In diesem ‚Zwösterreich' gäbe es sogar die gleichen Einwohner, die auch die gleiche Sprache verwenden wie wir hier auf der Erde. Mein Zwillingsbruder auf der Zwerde würde mit mir zudem die gleiche Geschichte teilen und sich auch im gleichen psychischen Zustand befinden wie ich hier auf der Erde.
Auf der Zwerde gäbe es natürlich auch Wasser. Das zwirdische Wasser erfüllt die gleiche Funktion wie irdisches Wasser, es füllt die Meere, Seen und Flüsse und lässt sich wie irdisches Wasser zum Trinken und Waschen verwenden. Trotz dieser oberflächlichen Ähnlichkeit unterscheide sich jedoch das zwirdische Wasser in einem Punkt von unserem Wasser: Seine chemische Struktur wäre nicht H_2O, sondern hätte die chemische Formel *XYZ*. Stellen wir uns nun ferner zu diesem Gedankenexperiment zwei Varianten vor.
I. Variante: Im Jahr 2050 landet ein Raumschiff von der Erde auf der Zwillingserde. Am Anfang werden die Astronauten zunächst annehmen, das zwirdische Wasser habe die gleiche

Bedeutung wie auf der Erde. Erst nach einer genaueren Untersuchung im Labor stellt sich dann heraus, dass das zwirdische Wasser eine andere molekulare Struktur hat wie das irdische Wasser. Unsere Astronauten werden auf die Erde einen Funkspruch durchgeben des Inhalts: „Das Wort ‚Wasser' hat auf der Zwerde die Bedeutung *XYZ*."
II. Variante: Verlegen wir unser Gedankenexperiment in das Jahr 1750. Zu dieser Zeit war die Chemie noch nicht dazu in der Lage, um irdisches Wasser von Wasser auf der Zwerde unterscheiden zu können. Nicht einmal die Experten wussten damals, dass irdisches Wasser aus Wasserstoff und Sauerstoff besteht. Wir können hier – ganz analog zu unserem weiter oben beschriebenen Beispiel eines nach China verschickten Gemäldes aus Europa – die Frage stellen: Würde ein Bewohner der Erde einen Fehler machen, wenn er zwirdisches Wasser (man stelle sich dazu nur vor, XYZ käme durch irgendeinen Vorfall, etwa den Einschlag eines Meteoriten, auf die Erde) als Wasser (in seiner Wortbedeutung) bezeichnete? An diesem Punkt scheiden sich die Geister. Die richtige Antwort auf diese Frage hängt von ganz bestimmten erkenntnistheoretischen und semantiktheoretischen Vorbedingungen ab.
Putnam meint jedenfalls, dass bereits 1750 ‚Wasser' auf der Erde eine *andere* Bedeutung hatte als auf der Zwillingserde. Es wäre daher auch falsch, zwirdisches Wasser als ‚Wasser' (in der Wortbedeutung der Erde) zu bezeichnen. Die Bedeutung von Ausdrücken, die eine natürliche Art bezeichnen (Wasser ist eine natürliche Art) sei nämlich unabhängig von unseren momentan verfügbaren Identifizierungsmöglichkeiten natürlicher Arten. Dass nämlich irdisches Wasser H_2O bedeute, heiße nämlich nicht, dass wir dies auch wissen müssen. Der Grund dafür läge daran, dass die Bedeutung von ‚Wasser' eben nicht über Wasseridentifizierungstests festgelegt werde. Seine Bedeutung lässt sich daher auch nicht auf das Wissen einer bestimmten Zeit reduzieren. Unser Erdenbewohner hätte nur dann einen Fehler gemacht, wenn für die Festlegung der Bedeutung irdischen Wassers einzig und allein operationale Testkriterien ausschlaggebend wären. Wenn also das, was wir mit Wasser meinen, sich reduzieren ließe auf unser Wissen über Wasser.
Doch was heißt das eigentlich, bereits 1750 mit ‚Wasser' H_2O zu meinen, obwohl zu dieser Zeit niemand – nicht einmal die Experten – wissen konnte, welche molekulare Struktur irdisches Wasser tatsächlich hat? Man kann sich diesen sonderbaren Umstand nur so erklären, dass die Festlegung der Wortbedeutung von ‚Wasser' eben über Kriterien erfolgt, die nicht von operationalem Charakter (wie etwa wissenschaftliche Tests oder ähnliches) sind. Doch mit welchen Mitteln soll die Bedeutung von ‚Wasser' (oder einer anderen natürlichen Art) festgelegt werden, wenn wir einmal alle wissenschaftlichen Verfahren zur Überprüfung von Wasser ausklammern? Eine mögliche Antwort auf diese Frage gibt uns der semantische Externalismus.

Putnams Zwillingserde II (semantischer Externalimus)

Nach dem semantischen Externalismus wird die Bedeutung von Bezeichnungen natürlicher Arten über die bezeichnete Substanz selber festgelegt. Diese Bedeutungsfestlegung erfolgt über die sogenannte *starre Referenz*. Was also Wasser ist bzw. nicht Wasser ist, entscheiden keine wissenschaftlichen (oder auch außerwissenschaftlichen) Überprüfungsverfahren, nicht das, was wir aufgrund unseres Wissens für Wasser halten (mit welchen Überprüfungsmethoden auch immer), zähle als Wasser, sondern das, was mit der bezeichneten Substanz identisch

sei. Kriterium dieser Identität (die Identitätsrelation, in der alle Wassermoleküle zur bezeichneten Substanz stehen) sei eben die Identität mit *dieser* (!) Substanz selber und nicht unser subjektives Wissen über die Substanz. Oder, wie es Putnam selbst auch ausdrückt, spielen „*die natürlichen Arten selbst* eine Rolle bei der Bestimmung der Extension jener Ausdrücke", die auf sie referieren (vgl. Putnam, Von einem realistischen Standpunkt, S. 139) bzw. sei es „'die Substanz selbst', die die Aufgabe erfüllt, die Extension des Terminus zu bestimmen." (Vgl. Putnam, Vernunft, Wahrheit und Geschichte, S. 45) Schließlich werde die Extension von Ausdrücken natürlicher Arten zumindestens „teilweise *durch die Welt festgelegt*" , es läge also an den „objektiven Gesetzen" der Natur, was zu einer Klasse gehöre und was nicht (vgl. Putnam, Von einem realistischen Standpunkt, S. 135).

Eine ganz ähnliche Theorie in Bezug auf Eigennamen hat Saul Kripke unabhängig von Putnam in *Name und Notwendigkeit* (1972) entwickelt. Kripke kritisiert dort die Auffassung, dass die Bedeutung eines Eigennamen synoym sei mit bestimmten Beschreibungen. So sei etwa ‚Moses' nicht synonym mit ‚der Mann, der die Isrealiten über das Rote Meer geführt hat'. Wäre nämlich die Bedeutung von ‚Moses' synonym mit ‚der Mann, der die Israeliten über das Rote Meer geführt hat', dann wäre er gar nicht Moses gewesen, hätte er die Israeliten nicht über das Rote Meer geführt. Diese Schlußfolgerung wird von Kripke in Frage gestellt. Wir können uns schließlich eine anders verlaufende Geschichte vorstellen, in der zwar derselbe Moses vorkommt, ohne dass aber die meisten uns geläufigen Beschreibungen über Moses zutreffen. Diese Auffassung hat gravierende Konsequenzen für die Modallogik, auf die ich hier aber nicht näher eingehen möchte.

Ein Motiv von Kripkes Theorie der Eigennahmen ist jedenfalls seine Kritik an der sogenannten *Cluster Theory of Meaning*, zu Deutsch der *Bündeltheorie (oder Beschreibungstheorie) der Bedeutung*. Was darunter zu verstehen ist, lässt sich an einem Beispiel demonstrieren, das Kripke, in Anlehnung an John Stuart Mill, in *Name und Notwendigkeit* gebracht hat (vgl. Kripke, Name und Notwendigkeit, S. 35f.): Der Name ‚Dartmouth' bezeichnet eine bestimmte Hafenstadt in England, gelegen an der Mündung des Flusses Dart. Wir können jetzt die Bedeutung von ‚Dartmouth' auf zwei verschiedene Weisen analysieren. Nach der Beschreibungstheorie bedeutet ‚Dartmouth' eben jene Stadt die an der Mündung des Dart River gelegen ist. Würde der Dart seinen Verlauf ändern und an einer anderen Stelle ins Meer münden, so könnten wir diese Stadt nicht mehr länger ‚Dartmouth' nennen. Verwenden wir dahingegen ‚Dartmouth' als Eigennamen, der in einem hinweisenden Sinne auf *diese* Stadt referiert, so würde sich an unserer Benennung selbst dann nichts ändern, wenn der Fluß Dart seinen Verlauf geändert hätte. Denn im letzteren Sinne kommt die Hafenstadt zu ihrem Namen über einen Taufakt, der ‚starr' (unabhängig von allen möglichen Konnotationen, die wir mit Dartmouth verbinden) *diese* ganz bestimmte Stadt eben bezeichnet.

Gegen die Bestimmung der Extension eines sprachlichen Ausdrucks über die starre Referenz lässt sich nun aber einwenden, wie denn überhaupt der Referent eines Namens festgestellt werden kann, ohne dabei auf irgendeine Beschreibung zurückgreifen zu müssen. Schließlich lässt sich der Referent eines Namens nur von demjenigen über einen hinweisenden Sprechakt bestimmen, der in einem unmittelbaren Bezug zu dem Referenten steht und bei dieser Gelegenheit vielleicht auf eine Stadt mit den Worten zeigt: ‚Diese Stadt ist Dartmouth.' So hat bereits Hegel in seiner Kritik an der unmittelbaren Sinneswahrnehmung ironisch gefragt: Wie steht es um den Wahrheitswert der zu nächtlicher Stunde ausgesprochenen Aussage: ‚Dieses ist die Nacht', wenn sie aufgeschrieben und am nächsten Tag dann abgelesen wird? Ein ganz

ähnliches Problem stellt sich im Übrigen auch bei Putnams Referenztheorie. Wie sollen wir bereits im Jahre 1750 mit ‚Wasser' H_2O meinen können, wenn uns zu dieser Zeit sämtliche operationalen Methoden fehlen, um dies feststellen zu können? Putnams Antwort darauf lautet, sprachliche Ausdrücke für natürliche Arten enthielten eine verborgene *indexikalische Komponente* (vgl. Putnam, Die Bedeutung von ‚Bedeutung', S. 46). Die Extension von ‚Wasser' beinhaltet all jene Wassermoleküle, die *von derselben Natur* sind wie paradigmatische Fälle von Wasser (vgl. Putnam, Von einem realistischen Standpunkt, S. 136).
Wie sollen wir aber auf solche Dinge in Situationen referieren können, in denen wir nicht dazu in der Lage sind, auf sie unmittelbar zeigen zu können? Kripkes Antwort auf diese Frage ist die ‚Kette der Referenz'. Er stellt sich dies so vor: Anfänglich erfolgt die Namensgebung in einem Taufakt, also in einer Situation, in der Menschen direkt auf den gemeinten Referenten zeigen können. Dieser urspünglich durch eine deiktische Handlung gestiftete Bezug zum Referenten wird dann von einem Sprecher in einer Gemeinschaft an den nächsten weitergegeben und bleibt dergestalt erhalten. Es ist die faktische Geschichte dieser Kommunikationskette, wodurch die Bedeutung eines Eigennamens erhalten wird. Eine auf diese Weise weitergegebene Bedeutung hat ihren Ursprung in der Geschichte. Sie ist nicht im Kopf jenes Teils der Sprechergemeinschaft, die den Namen als die derzeit letzten lebenden Glieder der Kommunikationskette gerade aktuell verwenden. Dies ist im Kern die These des semantischen Externalismus. Nach dieser These wird auch die Bedeutung von Ausdrücken natürlicher Arten über einen Taufakt festgelegt. Wir zeigen beispielsweise auf ein Glas Wasser und erklären dabei: Wasser sei all das, was mit *dieser* Flüssigkeit – jenseits meiner Fingerspitzen (!) - identisch ist. Die Flüssigkeit in dem Glas vor mir ist gewissermaßen ein Muster über das wir festlegen, was mit ‚Wasser' gemeint ist. Was also Wasser ist, wird, salopp gesprochen, über etwas festgelegt, das jenseits unserer Fingerspitzen (jenseits unserer Verifikationsmethoden) angesiedelt ist. Es ist die Substanz selber, die bei der Festlegung der Bedeutung eine Rolle spielt. Die Tragweite dieser These lässt sich ermessen, wenn wir verschiedene Pros und Kontras, die für oder gegen den semantischen Externalismus sprechen, einander gegenüber stellen.

Putnams Zwillingserde III (Pros und Kontras)

Zu den Pros gehören:

- Der Antioperationalismus
- Der Antiverifikationismus
- Argumente gegen den Kulturrelativismus
- Argumente gegen den *phenomenal internalism*
- Argumente gegen die referentielle Unbestimmtheit

Zu den Kontras gehören:

- Magische Bezugnahme
- Previligierter Zugang zu Wirklichkeit
- Metaphyischer Realismus

Die Pros

Putnam weist darauf hin, dass man bei Einhaltung eines strikt *operationalistischen* Standpunkts niemals interessante Entdeckungen machen könnte. Wäre Wasser einfach das, was den momentan gültigen operationalen Kriterien entspricht, so könnte man nicht sagen, dass unsere neu entwickelte Chemie *entdeckt* hat, dass Wasser H_2O bedeutet, denn sie hätte es ja nur über diese Kriterien so festgelegt (vgl. Putnam, Vernunft, Wahrheit und Geschichte, S. 44). Erst durch die neu entwickelte Chemie bekäme das Wort Wasser die Bedeutung von H_2O. Vor dieser Zeit wäre die Bedeutung von ‚Wasser' im Hinblick auf seine molekulare Struktur unbestimmt. Wasseridentifizierungsmethoden wären demzufolge keine Methoden, um zu entdecken, was Wasser ist, sondern definitorisches Merkmal von ‚Wasser'.
Wahr wäre in diesem Falle nur das, *was wir für wahr halten*. Dies ist die Position des *Verifikationismus*. Der Verifikationist glaubt nämlich, die Wahrheit wissenschaftlicher Theorien lasse sich reduzieren auf ihre Überprüfbarkeit. Nach dem Verifikationismus ist eine Aussage in einer wissenschaftlichen Theorie nur dann wahr, wenn sie sich *innerhalb* dieser wissenschaftlichen Theorie als wahr herausstellt. Wahrheit ist also gemäß der Doktrin des Verikationismus ein unseren wissenschaftlichen Theorien *untergeordneter* Begriff (vgl. Putnam, Die Bedeutung von ‚Bedeutung', S. 49).
Damit ist aber jede wissenschaftliche Entdeckung und Weiterentwicklung einer Wissenschaft nicht erklärbar. Denken wir an jene Zeit, in der die Menschen die Erde noch für eine flache Scheibe hielten. Vertritt man nun einen verifikationistischen Standpunkt, dann war die Aussage ‚Die Erde ist eine Scheibe' zu der Zeit, als Menschen noch nicht dazu in der Lage waren, das Gegenteil zu überprüfen, nicht falsch gewesen.
Nun ist es aber intuitiv plausibel, dass eine Aussage auch falsch (oder wahr) sein kann, wenn wir sie zu einem bestimmten Zeitpunkt nicht verifizieren können. Betrachten wir hierzu noch das folgende Beispiel. Gold und Katzengold sind, sofern man sich nur an deren Oberflächeneigenschaften orientiert, auf den ersten Blick nicht unterscheidbar. Katzengold hat eine goldene Farbe und einen ähnlich metallischen Glanz wie echtes Gold. Können wir deshalb ernsthaft behaupten, dass zu einer Zeit, als unsere Goldidentifizierungsmethoden noch nicht so ausgereift waren, um hier einen Unterschied feststellen zu können, mit dem sprachlichen Ausdruck ‚Gold' auch Katzengold gemeint war? Hat ‚Gold' eine Bedeutung nur relativ zu den gerade gültigen Verifikationsmethoden?
Eine Spielart und Steigerung des Verifikationismus ist der *Kulturrelativismus*. Es handelt sich hierbei um einen Standpunkt, der die Gültigkeit auch jeder ethisch-moralischen Aussage immer nur relativ zu den gerade geltenden Werten einer historisch kontingenten Kultur ansieht. Zu dem Dunstkreis solcher Vorstellungen zählt auch das postmoderne Diktum des *anything goes*, eines radikalen Liberalismus, der jede objektive ethische und ätshetische Verbindlichkeit negiert.
Leugnet man die konstruktive Rolle, die die objektiven Gesetze der Natur bei der Wahrheitsfindung wissenschaftlicher Theorien spielen, so führt ein solcher Standpunkt zum *phenomenal internalism*. Es handelt sich hierbei um eine erkenntnistheoretische Position, derzufolge unsere Wahrnehmung einzig und allein durch Fakten bestimmt ist, die innerhalb des Wahrnehmenden angesiedelt sind. Ein anschauliches Beispiel für eine solche Haltung findet sich in dem Block Buster *Matrix*. Cypher, der Verräter von Neo und Morpheus, begründet seinen

Verrat folgendermaßen: Was interessiert es mich, ob dieses Stück Fleisch ein echtes Steak ist oder nur ein virtuelles Fleisch, das von Maschinen als Programm in meinem Gehirn induziert wurde? Für Cypher zählt nur, dass es sich wie echtes Fleisch anfühlt und auch so schmeckt! Es zählt also nur das, was subjektiv als Fleisch verifiziert werden kann, nicht aber das, was die im Gehirn von Cypher induzierte Vorstellung *wirklich* ist, nämlich eine nur künstlich erzeugte Illusion.
Der Film Matrix bedient sich hier einer gängigen Spekulation, die sich an der Frage entzündet, ob es so etwas wie ‚Gehirne im Faß' geben könnte. Tatsächlich sind die Menschen in der Matrix nichts anderes als Gehirne im Faß. Eingeschlossen in unterirdischen Verließen dienen ihre Körper nur der Energieversorgung der Maschinen und die ganze Wirklichkeit ist lediglich eine von den Maschinen suggerierte Wirklichkeit, eine programmierte Illusion, eben die Matrix. Für einen Vetreter des *phenomenal internalism* macht es aber keinen Unterschied, ob jemand ein Gehirn im Faß ist oder aber in einer ‚echten' Umgebung lebt. Gewisse Ähnlichkeiten mit derzeit bereits bestehenden Parallelwelten, etwa mit *second life*, sind nicht von der Hand zu weisen.
Für einen Vertreter des semantischen Externalismus ist es jedoch ein großer Unterschied, ob das Geschehen, das wir gerade wahrnehmen, nur halluziniert wird oder wirklich so stattfindet, wie wir es auch subjektiv empfinden. Was also ein Steak ist, um auf das Beispiel von Cypher zurückzukommen, darüber entscheiden nicht unsere subjektiven Empfindungen, sondern die (über die starre Referenz festgelegte) Identität mit *echtem* Fleisch.
Darüber hinaus wäre die Bedeutung sprachlicher Ausdrücke von natürlichen Arten *referentiell unbestimmt*, wenn wir die starre Referenz als Methode zur Festlegung von Bedeutung nicht zulassen wollen. Denken Sie an das Beispiel, das ich als Einstieg in Putnams Gedankenexperiment gewählt habe, die Geschichte jenes europäischen Gemäldes, das nach China verschickt wurde. Macht ein Chinese, so haben wir jedenfalls gefragt, einen Fehler, wenn er eine kleiner dargestellte Person als eine Person mit geringerem sozialem Status interpretiert? Die Antwort auf diese Frage war ein klares Nein, da nämlich ein Gemälde ein Artefakt ist und das in ihm dargestellte Motiv daher auch keine intrinische Bedeutung hat. Das Gemälde ist, so habe ich argumentiert, hinsichtlich seiner geometrisch-räumlichen Denotation referentiell unbestimmt. Nicht die perspektivische Verkürzung meint von sich aus Räumlichkeit, sondern nur wir selber meinen mit der perspektivischen Verkürzung Räumlichkeit. Dies ist eben der Unterschied zwischen einem Artefakt und einem autonom denkenden Wesen.
Wäre nun die Bedeutung aller von uns verwendeten sprachlichen Ausdrücken referentiell unbestimmt, so wäre die Semantik kognitiver Zustände nicht autonom, sondern immer nur relativ zum Kontext eines ganz bestimmten Interpretationsschemas. Folgen wir dem Operationalismus, so meinen wir im Jahre 1750 mit ‚Wasser' nicht *wirklich* H_2O, was wir unter Wasser verstehen, ist abhängig von den in einem bestimmten geschichtlichen Kontext gerade gültigen Wasseridentifizierungstests. Nach der Doktrin des semantischen Externalismus ist es dahingegen so, dass ‚Wasser' in allen möglichen Kontexten H_2O bedeutet.
Die These von der referentiellen Unbestimmtheit führt also dazu, dass wir uns selber nicht mehr als selbständig denkende Wesen verstehen, sondern als Artefakte, hervorgebracht durch den blinden Evolutionismus der Natur. Es war Daniel Dennett, der diesen Zusammenhang zwischen referentieller Unbestimmtheit und Darwins Evolutionstheorie klar erkannt hat.

Wenn bei der Bestimmung der Extension von sprachlichen Ausdrücken natürlicher Arten als Kriterium auf die Substanz selber verwiesen wird und nicht auf unser Wissen über die Substanz, wie können wir dann aber jemals dessen sicher sein, worauf wir mit unseren Worten referieren? Haben wir es hier nicht mit einem *metaphyischen Realismus* zu tun, der seine Legitimation aus historischen Fakten ableitet, die jenseits aller Überprüfungsmethoden angesiedelt sind?

Ich erinnere mich noch gut an eine Diskussion im österreichichen Fernsehen, die vor einigen Jahren im damaligen *Club 2* stattgefunden hat. Die Diskussion drehte sich um die Abtreibung. Bei der Diskussion war auch ein äußerst konservativer Weihbischof zugegen. Einer der Teilnehmer an dem Gespräch hat dabei dem Bischof vorgehalten, dass seine Behauptungen nicht wahr wären, worauf letzterer geantwortet hat: Ich bin die Wahrheit! Seine Antwort war, wenn man von der inneren Logik des katholischen Fundamentalismus ausgeht, völlig korrekt. Denn was geglaubt werden soll oder nicht ist keine Frage, die dem Ermessen der Vernunft des Einzelnen anheimgestellt ist. So verurteilte bereits um die vergangene Jahrhundertwende Papst Pius der X in einem Schreiben, dem sogenannten „Motuproprio", den *Modernismus,* eine damals herrschende Strömung innerhalb und auch außerhalb der katholischen Kirche, welche eine direkt von Gott der Kirche verliehene Autorität in Frage stellte und derzufolge alles Äußere und Geschichtliche nur Geltung habe als symbolische Einkleidung für religiöse Erlebnisse. Was wahr ist, entscheidet nach dem päpstlichen Schreiben dahingegen die Amtskirche und nicht das subjektiv empfundene Gefühl und auch nicht das vorherrschende Wissen einer bestimmten Zeit.

Die objektive Wahrheit der christlichen Glaubensbotschaft wird, so jedenfalls im katholischen Fundamentalismus, allein durch die Autorität der Kirche tradiert. Der Bischof in jener Fernsehsendung verstand sich daher auch nur als ein Glied in der Kette der Referenz, die aus der Gemeinschaft aller von der Kirche ernannten Amtsträger besteht und die ihre Autorität direkt auf Gott zurückführt. Läßt man die speziellen inhaltlichen Aspekte einmal beiseite, so ist dieser Grundgedanke nicht allzu weit entfernt von Kripkes Theorie der Kommunikationskette zur Bestimmung der Extension von Eigennamen. Einziger Garant für die Wahrheit der Glaubensinhalte ist die Vertrauenswürdigkeit der Überlieferung. Es sind die realen Glieder in der Kommunikationskette, die allein einen *priveligierten Zugang* zum Referenzobjekt (Christusereignis oder ähnliches) haben. Eine solche priveligierte Art der Bezugnahme hat einen *magischen* Charakter, denn sie ist jenseits aller Verifikationsmethoden. Es handelt sich hierbei um einen, wie es gelegentlich auch so genannt wird, Gottesstandpunkt ohne Rücksicht auf den Zuschnitt menschlichen Erkennens. Denn wir haben es hier mit einer Vorstellung von Wahrheit zu tun, die nurmehr dem engen Kreis der Initiierten zugänglich ist.

Dazu kommt aber noch, dass ein derartiger metapyhsicher Realismus mit wissenschaftlichen Methoden überhaupt nicht zu fassen ist und daher auch nicht naturalistisch beschrieben werden kann. Dies wird deutlich, wenn man Quines Reizbedingungen zur Übersetzung von ‚Gavagai' mit Putnams operationalen Methoden zur Bestimmung der Bedeutung vergleicht. Übertragen wir Quines Doktrin der Übersetzungsunbestimmtheit auf Putnams Analyse der Bedeutung sprachlicher Ausdrücke von natürlichen Arten, so kann man sagen: Ähnlich wie wir bei gleichen Bestrahlungsmustern auf der Retina verschiedene Übersetzungen von ‚Gavagai' (und damit verschiedene Bedeutungszuschreibungen) bekommen können, ist auch die

Bedeutung von ‚Wasser' hinsichtlich der operationalen Methoden zur Bestimmung von Wasser unbestimmt. Schließlich sind die operationalen Methoden zur Bestimmung von Wasser auf der Erde und der Zwillingserde im Jahre 1750 gleich und dennoch haben die sprachlichen Ausdrücke eine verschiedene Bedeutung. In diesem Punkt sind die Positionen von Quine und Putnam nicht allzu weit voneinander entfernt. Während jedoch Putnam eine Bestimmbarkeit der Bedeutung jenseits der operationalen Verfahren über die Methode der starren Referenz für möglich hält, findet sich Quine mit der Bedeutungsunbestimmtgheit ab. Er warnt sogar ausdrücklich vor einem schädlichen Mentalismus, der die Bedeutung sprachlicher Ausdrücke irgendwo jenseits des beobachtbaren Verhaltens ansiedelt. Schließlich gibt es keine magischen Strahlen, mit deren Hilfe wir die Bedeutung eines Wortes über alle wissenschaftlichen Verfahren hinaus festlegen können.
Damit stellt sich aber umso eindringlicher die Frage, wie wir einen schädlichen Mentalismus vermeiden können, ohne dabei aber in die Falle des Verifikationismus respektive des Kulturrelativismus zu tappen? Gesucht ist eine Mittellösung, die beide Extremstandpunkte vermeidet. Eine solche Mittellösung hat Putnam in verschiedenen Anläufen zur Interpretation des Realismus angestrebt, an prominenter Stelle sind hier der interne Realismus, der pragmatische Realismus sowie der direkte Realismus zu erwähnen. Ich beschränke mich im Folgenden auf eine Besprechung des internen Realismus.

Interner Realismus

Mit der Einführung des internen Realismus hat Putnam seine starke realistische Position, wie er sie noch in *Die Bedeutung von ‚Bedeutung'* vertreten hat, deutlich abgeschwächt. Nach dieser Auffassung gibt es, kurz gesagt, keine Dinge an sich irgendwo dort ‚daußen' jenseits unseres Verstandesvermögens, die von Hause aus zu bestimmten Arten gehören, sondern erst aufgrund unserer Begriffsschemata wird die Welt in Gegenstände aufgespalten. Es sind also wir selber, die mit Hilfe unserer Begriffsschemata die Dinge zu Arten ordnen. Putnam sagt dazu auch, es gäbe in der Welt keine Gegenstände, die sich aufgrund ihrer ontologischen Natur von selber identifizieren. Diese Position erinnert im Kern stark an die Transzendentalphilosophie von Kant. Eine der Grundbotschaften der *Kritik der einen Vernunft* ist ja, dass unser Denken nicht so etwas wie Gegenstände an sich ausmachen könne, sondern erst aufgrund der Kategorien unseres Verstandes konstituieren sich Gegenstände. So ist beispielsweise auch die Kausalität nach Kant keine in der Natur vorkommende Eigenschaft, sondern nur Teil jenes subjektiven Begriffs- und Ordnungssystems, das wir den Sinnesdaten – dem Gewühl der Empfindungen - überstülpen. Betrachtet man nur das, was uns die Sinnesdaten liefern, so handelt es sich hierbei um ein ungeordnetes Gewühl von Empfindungen, das erst im Verstand zu Gegenständen verarbeitet wird. Für Putnam bedeutet dies nunmehr, dass es keine sich selbst identifizierenden Gegenstände geben könnte.
Vergleicht man nun diese Position mit Putnams Überlegungen zur starren Referenz, so haben wir aber ein Problem. Denn dort hat es ja geheißen, dass die Extension von Ausdrücken natürlicher Arten nicht über irgendwelche Tests festgelegt werden könne, sondern über objektive Gesetze der Natur. So sei beispielsweise Wasser all das, was in einer bestimmten Identitätsrelation zu paradigmatischen Fällen von Wasser steht. Wasser ist das, was *zur selben Art* gehört wie das Wasser, auf das ich da zeige.

Damit stellt sich die Frage: Wie kann es sein, dass auf der *einen* Seite nur wir selber die Dinge in Arten aufspalten und auf der *anderen* Seite das, was zu einer natürlichen Art gehört, teilweise durch die Welt festgelegt wird? Brauchen wir im letzteren Falle nicht doch so etwas wie sich selbst identifizierende Gegenstände? Anders ausgedrückt: Wie kann die in unseren Begriffsschemata angesiedelte Identitätsrelation – nach dem internen Realismus ist das, was zu einer bestimmten natürlichen Art gehört, nur relativ zu unseren Begriffsschemata – eine Identität außerhalb unserer Begriffsschemata ausmachen?
Die spezielle Antwort, die Putnam auf diese Frage gibt, soll dabei helfen, sowohl die Fehler des Verifikationismus als auch jene des metaphyischen Realismus zu vermeiden. Eine erste Antwort findet sich bereits in *Die Bedeutung von ‚Bedeutung'*, also noch lange bevor Putnam den Gedanken des internen Realismus aufgegriffen hat. Er meint dort, die Identitätsrelation, in der Wassermoleküle zu paradigmatischen Fällen von Wasser stehen, sei eine *theoretische Relation* (vgl. Die Bedeutung von ‚Bedeutung', S. . 35). Um festzustellen, ob etwas die gleiche Flüssigkeit ist wie diejenige in dem Glas, auf die ich zeige, bedarf eines *unabsehbaren* Forschungsaufwands. ‚Flüssidentität' (sic!) ist also nur ein idealtypisches Konzept. Natürlich wird man diese Identität früher oder später auch überprüfen müssen. Es gibt nicht so etwas wie eine bereits feststehende Identität aller Wassermoleküle unabhängig von allen Möglichkeiten, diese Identität auch überprüfen zu können. In diesem Sinne befindet sich auch Putnam auf dem Boden des Verifikationismus. Auf der anderen Seite räumt er aber ein, dass ‚Flüssidentität' zwar nicht außerhalb *aller* Überprüfungsmöglichkeiten angesiedelt ist, ‚Wasser' aber dennoch mehr bedeutet als unsere *momentanen* Überprüfungsmöglichkeiten uns verraten. Dies ist der interne Realismus: Die Identität natürlicher Arten liegt nicht außerhalb *jedes* Begriffsschemas, sie liegt nur außerhalb unseres derzeit gültigen Begriffsschemas. Ich lege zwar momentan fest, was eine natürliche Art ist und verwende hierzu auch meine mir zur Verfügung stehenden Möglichkeiten, dies zu überprüfen. Aber ich behalte mir vor, dass diese Überprüfungsmöglichkeiten auch falsch sein können. Schließlich könnten ja in hundert Jahren neue Tests entwickelt werden, bei denen sich herausstellt, dass das, was ich momentan für Wasser halte, in Wirklichkeit überhaupt nicht Wasser ist. Was ich mit ‚Wasser' meine, ist daher auch unabgängig von den mir momentan zur Verfügung stehenden Testverfahren. Würde sich nämlich das, was als eine natürliche Art zählt, reduzieren lassen auf das, was wir zu einem jeweiligen Zeitpunkt als diese Art bestimmen können, so könnten wir unsere wissenschaftlichen Annahmen über die Natur niemals revidieren. Wahr wäre einfach das, was im Rahmen einer wissenschaftlichen Theorie gerade gültig wäre. Doch dies hatten wir ja schon.
Inwiefern gibt es nun aber so etwas wie selbstidentifizierende Gegenstände und inwiefern gibt es sie nicht?
Auf der einen Seite gilt: Was zu einer natürlichen Art gehört, ist nicht unabhängig von allen Testverfahren, sondern nur unabhängig von unseren derzeit gültigen Testverfahren. In diesem Sinne ist Wasser keine natürliche Art, die wir in einer von denkenden Wesen völlig unabhängigen Fertigwelt vorfinden. Schließlich sind es nur wir selber, die die Dinge zu Arten ordnen. In diesem Sinne gibt es also keine selbstidentifizierenden Gegenstände. Auch die Identitätsrelation, die wir aufgrund der starren Referenz festlegen, muß irgendwann einmal in einem Testverfahren nachgewiesen werden.
Inwiefern gibt es aber doch so etwas wie selbstidentifizierende Gegenstände? Wenn Wasser keine natürliche Art ist, die für sich existiert, wie kann es dann überhaupt die Substanz selber sein, die die Aufgabe erfüllt, die Extension des Wortes ‚Wasser' zu bestimmen? Dies ist ja im

Kern die Grundaussage der starren Referenz: Bedeutung ist nicht im Kopf. Inwiefern gibt es also doch selbstidentifizierende Gegenstände, wenn nicht auf die metaphysische Weise? Prinzipiell muß die Frage, ob etwas die gleiche Flüssigkeit ist wie paradigmatische Fälle von Wasser irgendwann einmal durch unsere Forschungsvorhaben geklärt werden können. Was kann uns dann aber die starre Referenz, der zufolge ja Wasser all das ist, was mit paradigmatischen Fällen von Wasser identisch ist, mehr über Wasser verraten als bestimmte Kriterien zur Überprüfung von Wasser?
Hier kommt der interne Realismus ins Spiel: Wenn wir mittels der starren Referenz die Extension von Ausdrücken natürlicher Arten festlegen, so transzendieren wir unsere derzeit gültigen Testverfahren. Wir behalten uns vor, dass diese auch falsch sein können. Wir halten, wie Putnam auch sagt, die Extension von ‚Wasser' ein wenig offen. Die Extension offen halten meint, dass wir nicht mit gerade geltenden operationalen Verfahren von vornherein exakt festlegen, was zu einer natürllichen Art gehört
Das, was aber jenseits unserer derzeit gültigen Testverfahren liegt, sind keine objektiven Fakten in einer ‚Fertigwelt', sondern es sind spätere, momentan noch unbekannte Testverfahren, die zum gegenwärtigen Zeitpunkt noch nicht durchgeführt wurden bzw. nicht durchführbar sind. Indem wir die Extension eines Ausdrucks für eine natürliche Art ein wenig offen halten, machen wir gewissermaßen eine Anleihe auf spätere Testverfahren – eine Anleihe, die früher oder später aber zurückbezahlt werden muß. Mit Hilfe der starren Referenz markieren wir sozusagen die Grenzen unserer Vernunft, ohne sie jedoch in unzulässiger Weise zu überschreiten. Die Substanz, deren Aufgabe es ist, die Extension zu bestimmen, ist nichts anderes als eine Marke für die Grenzen unserer Vernunft und die Dinge, auf die wir uns starr beziehen, sind eine Art von Platzhalter, von Indikatoren für die Grenzen unserer wissenschaftlichen Theorien.

Eine ontologische Intuition

Hinter der Vorstellung, man müsse die Extension von Ausdrücken für natürliche Arten ein wenig offen halten, steckt eine *schwache ontologische Intuition*. Was heißt es also, das mit unseren sprachlichen Ausdrücken gemeinte lasse sich nicht reduzieren auf das, was uns bzw. der Gemeinschaft der Wissenschaftler einer Zeit im Lichte bestimmter operationaler Verfahren als dieses gemeinte erscheint und was ist schwach an dieser ontologischen Intuition?
Ich habe bereits darauf hingewiesen, dass wir es hier mit einem ganz ähnlichen Gedankengang zu tun haben, wie er bereits bei Quine in seiner These von der Übersetzungsunbestimmtheit begegnet. Wie wir bereits wissen, lässt sich die Bedeutung von ‚Gavagai' nicht auf nonverbale Hinweisreize reduzieren. Bei gleichen Bestrahlungsmustern auf der Retina sind verschiedene Übersetzungen von ‚Gavagai' möglich. Was wir also mit ‚Gavagai' meinen, geht über die nonverbalen Hinweisreize hinaus. Ein ganz ähnliches Problem haben wir auch in dem Beispiel mit der Perspektive vorgefunden: Nicht die perspektivische Verkürzung im Bild meint *wirklich* räumliche Distanz, sondern nur wir selber meinen mit der perspektivischen Verkürzung wirklich Distanz. Kein Signal, kein beobachtbares Verhalten ist dazu in der Lage, eindeutig festzulegen, was wir mit unseren Worten meinen. Schließlich stellt sich dann die Frage, wie wir mit unseren Worten jemals etwas wirklich meinen können, wenn alle Signale, alle Informationen, die wir zu den Worten erhalten, zu kurz greifen. Wenn also die Bedeutung unserer Worte, die wir verwenden, über alle Oberflächenmerkmale hinaus festgelegt

ist, wie können wir dann jemals diese Bedeutung erlernen? Stellen Sie sich vor, Sie versuchen jemandem zu erklären, was die perspektivische Verkürzung bedeutet, indem sie ihm dazu ein Bild zeigen. Wenn aber das, was wir mit der perspektivischen Verkürzung meinen, mehr ist, als das, was unmittelbar im Bild enthalten ist, wie können wir dann jemals die Bedeutung der perspektivischen Verkürzung lernen? Wie können wir unter dieser Voraussetzung überhaupt jemals sinnvoll miteinander kommunizieren? Der schädliche Mentalismus, von dem Quine spricht, lässt schön grüßen.

Hier begegnet uns wiederum, allerdings in einem ganz anderen Zusammenhang, das bereits bekannte *symbol grounding problem*. Stellen Sie sich folgende Situation vor. Man würde in eine Kapsel sämtliche kulturelle Errungenschaften der Menschheit aufzeichnen und diese Kapsel dann ins All befördern. Wie könnte eine fremde Zivilisation jemals diese Schrift entziffern? Wenn wir in die Kapsel auch einen Dechriffierschlüssel zur Entzifferung der Schrift hineingeben, wie kann dieser Dechriffierschlüssel wiederum verstanden werden?

Vor einiger Zeit hatte ich eine Diskussion über die textbasierte Kommunikation im Internet. Im Laufe der Diskussion wurde behauptet, dass man über diese Art der Kommunkaition nichts Neues erlernen könnte, da die textbasierte Kommunikation bereits ein gemeinsames Vorverständnis voraussetze. Würde die Wissensvermittlung allein über die Schiene der textbasierten Kommunikation erfolgen, so käme dies einer *Einfrierung* unseres Wissens gleich auf eben dieses gemeinsam geteilte Vorverständnis. Man stelle sich nur den Versuch vor, einem Angehörigen einer schriftlosen Gesellschaft das Schreiben über eine textbasierte Kommunikation beizubringen.

Ein kleines Beispiel aus dem Bereich des E-learnings beleuchtet deutlich, welche Grenzen der Informationsvermittlung im Internet gesteckt sind und inwiefern es hier zu einer Einfrierung des Wissens auf ein Vorverständnis kommt. In meinen Statistikkursen verwende ich im Netz interaktive Übungsmodule, mit deren Hilfe unter anderem sich die Regressionskoeffizienten einer Regressionsgerade direkt berechnen lassen. Darüber hinaus ist es auch möglich, diese Regressionsgerade dynamisch am Bildschirm dazustellen. Bei den parallel noch stattfindenden Übungen mit verpflichtender Präsenz hat sich dann aber des Öfteren herausgestellt: Erst durch das Zeichnen einer solchen Gerade mit der Hand wurde erst wirklich die Bedeutung der Regressionsgerade verstanden. Das Wissen über die Gerade steckt also ein Stück weit in der Hand des Zeichners und kann daher auch nicht allein durch eine Betrachtung der Gerade am Bildschirm erworben werden. Es kann auch nicht über eine textbasierte Kommunikation erworben werden.

Als ich in jener Diskussion über die Grenzen der Internetkommunikation diesen Einwand brachte, wurde von einem meiner Kollegen die prinzipielle Frage gestellt: Können wir die Beschränkungen der textbasierten Kommunikation sprengen, indem man weitere Kommunikationskanäle nutzt? Wenn wir also nicht nur textbasiert, sondern über Webcam, Joystick oder über akustische Kanäle (Stimme) miteinander kommunizieren?

Wenn wir beispielsweise via Internet ein Hologramm übertragen, das es uns erlaubt unser Gegenüber von allen Seiten zu betrachten, können wir dann die Bedeutung unserer Worte direkt an den übermittelten Informationen ablesen oder brauchen wir auch in diesem Falle bereits ein Vorverständnis, das bei der Kommunikation im Netz vorausgesetzt werden muß?

Nachdem die Diskussion sich so immer mehr ins prinzipielle und futuristische steigerte, wurde dann als letztes und ultimatives Beispiel das folgende vorgebracht: Es gibt eine neue Variante von Cyber Sex, bei der sich die Partner durch ferngesteuerte Werkzeuge berühren kön-

nen. Kann man in diesem Falle sagen, dass wir hier ein Beispiel einer unmittelbaren sexuellen Erfahrung vor uns haben oder brauchen wir, um eine Berührung mit einem ferngesteuerten Werkzeug als Berührung unseres entfernten Partners zu verstehen, wiederum ein Vorverständnis? Gesetzt den Fall, die technische Kommunikation entwickelt sich so weit, dass sie alle Kommunkationskanäle erfasst, können wir dann immer noch behaupten, dass für das Verständnis der übermittelten Informationen ein Vorverständnis notwendig ist? Sollte dies der Fall sein, so ist zu fragen, wie denn in der direkten Kommunikation überhaupt ein unmittelbares Verstehen möglich sein kann. Ist unter dieser Voraussetzung nicht jede Art der Kommunikation unbestimmt?
Kehren wir zu jener taktilen Variante von Cyber Sex zurück. Hier kann man jedenfalls sagen: Die ferngesteuerten Berührungen sind keine unmittelbaren sexuellen Erfahrungen per se, sie meinen nicht von sich aus den Partner bzw. die Partnerin, sondern nur wir selber verstehen die ferngesteuerten Berührungen *als* Berührungen einer Frau oder eines Mannes. Wir brauchen also bereits ein Vorverständnis, um die übermittelten Signale als Berührungen einer Frau, eines Mannes zu interpretieren.
Worin besteht dann aber hier, so ist schließlich zu fragen, noch ein Unterschied zur direkten Kommunikation? Das eigentliche Problem bei der computervermittelten Kommunikation besteht darin, dass in ihr grundsätzlich nur solche Informationen übertragen werden, deren Kommunikationskanäle von vornherein feststehen. Nur das, was wir über die Kommikation also bereits wissen, kann auch technisch realisiert werden. Damit wird aber grundsätzlich die Kommunikation auf das eingefroren, was wir über die Kommunikation an Wissen mitbringen. Nun besteht aber unser Umgang mit anderen Personen in einer direkten Begegnung ja gerade darin, dass wir unser Gegenüber niemals reduzieren wollen (und auch können) auf irgendwelche Informationen, die wir zu einem bestimmten Zeitpunkt über eine Person zur Verfügung haben. Dies ist es auch, was ich unter einer schwachen ontologischen Intuition verstehe. Hätten wir nicht diese Intuition, so könnten wir auch niemals zwischen Schein und Wirklichkeit, zwischen Illusion und Wahrheit unterscheiden. Schwach ist die Intuition insofern, als wir uns mit unseren Worten auf ein unbekanntes X beziehen, ohne uns dabei anzumaßen, dieses unbekannte ‚Dieses-da' jemals endgültig unserem Wissen einverleiben zu können. Statt dessen finden wir uns damit ab, dass wir alles, was wir mit unserem Wissen überblicken können, stets transzendieren können und all unser Wissen stets beschränkt, vorläufig, kontingent bleiben wird. Man kann hier in gewisser Weise, allerdings in starker Pointierung, von einem „blinden" Realismus sprechen. Blind ist dieser Realismus insofern, als er sich auf eine Dimnension der menschlichen Erfahrung einläßt, die sich jeder Beschreibung und jedem positiven Wissen entzieht.
Als letztes und abschließendes Beispiel zu dieser Problematik verweise ich noch auf eine Geschichte, die mir bei meiner Untersuchung von Kennenlernprozessen in Chatforen untergekommen ist. Ein Pärchen in einem dieser Foren hat mir folgendes berichtet. Die Partner haben sich über das Netz kennengelernt, wobei beide um möglichst große Authenizität bemüht waren. Es wurden also bewußt keinerlei falschen Angaben während der Zeit des Kennenlernens im virtuellen Raum des Chat gemacht. Dennoch war die erste reale Begegnung dann eine Enttäuschung. Beide Partner hatten den Eindruck, einer ganz anderen und fremden Persönlichkeit gegenüber zu stehen als dies vorher im Chat den Anschein hatte. In einem längeren Gespräch mit den beiden versuchte ich dieser Diskrepanz auf die Spur zu kommen. Um es kurz zu machen, stellte sich dabei im wesentlichen das folgende heraus: Die einzigen Informationen, die

über das Medium Computer ausgetauscht wurden (und auch austauschbar waren), waren *Selbstdarstellungen* der beiden Gesprächspartner (ich beziehe mich in diesem Beispiel auf einen textbasierten Chat). Damit werden aber in der computervermittelten Kommunikation ganze Aspekte unserer Persönlichkeit ausgeblendet und auf Selbstsbeschreibungen eingefroren. Es kommen stattdessen nur jene Aspekte unserer Persönlichkeit zum Vorschein, die uns auch in der Reflexion zugänglich sind. Eine reale Begegnung spielt sich dahingegen auf einer ganz anderen Ebene ab. In einer realen Beziehung reduzieren wir unser Gegenüber nicht auf ein explizites Wissen, das wir von unserem Partner haben. Wir sind gleichermaßen offen für die unbekannte Seite in dem Anderen. Es ist zuallererst diese ontologische Intuition, die jeden echten Dialog und jede echte Begegnung zuallererst vorantreibt.

Putnams Antiindividualismus

Nach dieser längeren Reflexion über die erkenntnistheoretischen Rahmenbedingungen des semantischen Externalismus komme ich nun zu folgendem Fazit. Die Bedeutung sprachlicher Ausdrücke lässt sich nicht ausschließlich und allein über operationale Kriterien bestimmen. Was wir beispielsweise mit Wasser meinen, ist nicht reduzierbar auf unser zu einer bestimmten Zeit verfügbares Wissen über Wasser. Bedeutung ist nicht im Kopf. Sprachliche Ausdrükke gleicher Individuen haben in verschiedenen Kontexten daher auch eine verschiedene Bedeutung. Diese Position, die Putnam gemeinsam mit Quine teilt, nennt man den Antiindividualismus. Vor diesem Hintergrund kommt Putnam zu folgender Kritik an der klassischen Bedeutungstheorie. Diese gehe von zwei Prämissen aus:

I. Wir interpretieren einen Ausdruck, indem wir uns in einem ganz bestimmten individuellen Geisteszustand befinden. Anders formuliert heißt das: Die Bedeutung eines Ausdrucks wird über den individuellen Geisteszustand festgelegt. Sind die Geisteszustände gleich, so impliziert dies auch Bedeutungsgleichheit.
II. Die Extension eines Ausdrucks ist über seine Bedeutung festgelegt. Unter Extension ist der Begriffsumfang eines Ausdrucks zu verstehen (worauf sich ein Ausdruck bezieht). Kennen wir die Bedeutung eines Wortes, so wissen wir auch, wovon das Wort handelt bzw. worauf es sich bezieht. Aus Bedeutungsglechheit folgt daher auch Extensionsgleichheit.

Putnam behauptet im Anschluß an diese beiden Prämissen, dass nicht beide Prämissen *zugleich* zutreffen können und daher auch eine der beiden Prämissen falsch sein müsse. Dies liegt daran, dass – folgen wir *beiden* Prämissen – aus Gleichheit der Geisteszustände auch Extensionsgleichheit folgen müsste. Dies kann aber nach dem Gedankenexperiment von der Zwillingserde nicht der Fall sein. Schließlich beziehen sich ein Sprecher auf der Erde auf H_2O und ein Sprecher auf der Zwerde auf *XYZ*, obwohl sie sich im gleichen Geisteszustand befinden.
Nun besagt die erste Prämisse, dass verschiedene Bedeutungen auch auf verschiedene Geisteszustände zurückzuführen sind. Die Bedeutung sprachlicher Ausdrücke *superveniert* über dem individuellen Geisteszustand ihres Sprechers. Die zweite Prämisse besagt, dass Extensionsverschiedenheit auf Unterschiede in der Bedeutung zurückzuführen ist. Die Extension

sprachlicher Ausdrücke *superveniert* dieser Prämisse zufolge also über deren Bedeutung. Wenn nun aber eine der beiden Prämissen Putnams Gedankenexperiment zufolge falsch ist, so kann die Extension sprachlicher Ausdrücke nicht über dem individuellen Geisteszustand ihrer Sprecher supervenieren. Eine der beiden Supervenienzen muß also aufgegeben werden.
Behalten wir die erste Prämisse bei und verwerfen die zweite Prämisse, so hat Wasser auf der Erde bzw. auf der Zwerde die gleiche Bedeutung aber eine verschiedene Extension. Eine Alternative dazu wäre, wofür sich Putnam dann auch letztlich entscheidet, die erste Prämisse zu verwerfen und dafür aber die zweite Prämisse beizubehalten. Diese Alternative ist ganz im Sinne des Antiindividualismus: Sprecher auf der Erde und der Zwerde sind zwar im gleichen Geisteszustand und dennoch hat Wasser in den verschiedenen Kontexten auch eine verschiedene Bedeutung. Putnam entscheidet sich für diese Alternative, da wir, wie er jedenfalls meint, bei der Festlegung von Bedeutung auch den „Beitrag der Umwelt" mitberücksichtigen müssen. Schließlich hänge es von der Einbettung in eine äußerere Umgebung ab, was ein sprachlicher Ausdruck bedeutet. Die Bedeutung sprachlicher Ausdrücke superveniert nach dieser Auffassung nicht über dem individuellen Geistestzustand ihrer Sprecher.
Im Anschluß an dieser Überlegungen ist nun aber zu fragen, inwiefern durch Putnams Kritik auch das Supervenienzprinzip, wie es in der komputationalen Theorie des Geistes verwendet wird, mitbetroffen ist. Nach dieser Theorie geht es darum, menschliches Verstehen zu reduzieren auf eine Abfolge von (formal beschreibbaren) Computerzuständen. Es geht darum, den semantischen Gehalt unseres Denkens auf das formale Operieren mit abstrakten Symbolen zurückzuführen. „Syntax mirrors semantics", so lautet – wir wissen es bereits - das Formalistenmotto. Nun steht dieses Motto, wie sich leicht zeigen läßt, aber in einem krassen Gegensatz zur Doktrin des Antiindividualismus. Denn nach dem Formalistenmotto supervenieren semantische Eigenschaften eines kognitiven Systems über dessen formale Eigenschaften. Und diese formalen Eigenschaften supervenieren ihrerseits über ihre physikalischen Trägerprozesse. Die Festlegung semantischer Eigenschaften erfolgt daher rein ‚individualistisch', also ohne Berücksichtigung des Beitrags der Umwelt. Konfrontiert man nun diese Position mit Putnams Gedankenexperiement, so ergibt sich aber das folgende Bild: Ich und mein Zwillingsbruder auf der Zwerde befinden sich laut Annahme dieses Gedankenexperiments im gleichen physikalischen Zustand und gleiche physikalische Zustände ziehen gleiche psychische Zustände nach sich (Dies erfolgt unter der Prämisse, dass Geisteszustände über ihren physikalischen Zuständen supervenieren). Aus Gleichheit von Geisteszuständen folgt wiederum (siehe die erste Prämisse der klassischen Bedeutungstheorie) Bedeutungsgleichheit. Nun haben Ausdrücke für natürliche Arten jedoch eine *andere* Bedeutung auf der Erde als auf der Zwillingserde. Semantische Eigenschaften sind daher nicht über die (individualistisch beschreibbaren) physikalischen Eigenschaften festgelegt. Damit widerspricht der semantische Externalismus ganz klar einem zentralen Grundgedanken der komputationalen Theorie des Geistes.

Fodors zweigeteilte Semantik

Es war nun Jerry Fodor, einem der herausragendsten Vertreter der Cognitive Science und Schüler von Hilary Putnam, der den Versuch unternommen hat, diese verfahrene Situation zu bereinigen. Fodor strebte dabei eine salomonische Lösung an, die sowohl dem semantischen

Externalismus aber auch dem Formalistenmotto der komputationalen Theorie des Geistes Gerechtigkeit widerfahren läßt.
Er orientiert sich in seiner Analyse kognitiver Prozesse am Vorbild der Sprache. Kognizieren heißt für ihn so viel wie das Manipulieren von Symbolen in unserem Kopf. Eines seiner frühen Hauptwerke trägt denn auch den Titel „The Language of Thought“ (1975) – die Sprache des Geistes. Überträgt man nun Eigenschaften der expliziten Sprache auf mentale Prozesse, so zeichnen sich letztere durch folgende Eigenschaften aus: Symbole in unserem Kopf sind – in Analogie zu den Wörtern unserer Sprache – Repräsentanten einer ‚außergeistigen' Realität und sie haben eine logische Konstituentenstruktur, was so viel heißt, dass sich komplexe Strukturen durch Rekombinationen von atomaren Elementen konstruieren lassen. Auch in der Sprache können wir modular aufbauend aus einfacheren Sätzen komplexere Strukturen erzeugen. Weitere Thesen Fodors sind der Komputationalismus und der Formalismus. Während erstere von der Vorstellung ausgeht, die Symbole in unserem Kopf lassen sich auf ähnliche Weise verarbeiten wie dies bei den Zeichenketten eines Computers der Fall ist, sind nach der Vorstellung von letzterer semantische Zuschreibungen ausschließlich über die formalen Eigenschaften kognitiver Prozesse festgelegt. Sind zwei Beschreibungen von Geisteszuständen formal identisch, so sind auch ihre semantischen Eigenschaften identisch. Es ist vor allem diese letzte These, für die Putnams Gedankenexperiment eine besondere Herausforderung darstellt.
Fodors Position läßt sich recht gut nachvollziehen, indem man auf die beiden Prämissen der klassischen Bedeutungstheorie zurückgreift, wie sie von Putnam beschrieben wurden, und sich dabei nachstehende Strategie Fodors vergegenwärtigt: Grundsätzlich versucht er beide Prämissen beizubehalten und dabei aber gleichzeitig auch Putnams Kritik gerecht zu werden. Dies geschieht durch eine semantische Arbeitsteilung. Fodor unterscheidet – je nach Prämisse – zwei verschiedene Bedeutungen von ‚Inhalt' (Content), nämlich einen engen Inhalt mentaler Repräsentationen (der sich an die erste Prämisse der klassischen Bedeutungstheorie anschließt) und einen weiten Inhalt (der sich an die zweite Prämisse anschließt).
Da gleiche Geisteszustände den gleichen engen Inhalt haben, hat das Wort ‚Wasser' auf der Erde und der Zwillingserde auch den gleichen engen Inhalt. Der enge Inhalt wird gewonnen ohne Berücksichtigung der äußeren Umgebung, in die ein Geisteszustand eingebettet ist. Dies folgt zumindestens aus der ersten Prämisse der klassischen Bedeutungstehorie. Nun unterscheidet sich ‚Wasser' auf der Erde und der Zwillingserde hinsichtlich des weiten Inhalts. Sind die Extensionen sprachlicher Ausdrücke verschieden, so muß auch ihr weiter Inhalt verschieden sein. Dies entspricht wiederum der zweiten Prämisse der klassischen Bedeutungstheorie: Aus Inhaltsgleichheit (im Sinne von weitem Inhalt) folgt auch Extensionsgleichheit. Selbst wenn beide Prämissen zutreffen, folgt wegen der Verschiedenheit des engen Inhalts (erste Prämisse) vom weiten Inhalt (zweite Prämisse) nicht, dass gleiche Geisteszustände auch gleiche Extensionen nach sich ziehen, was wiederum dem semantischen Externalismus und Putnams Gedankenexperiment entgegenkommt. Schauen wir uns das im Detail näher an: Was ist der enge Inhalt mentaler Repräsentationen?

Enger Inhalt

Nachdem, wie Fodor meint, Geisteszustände gewissermaßen intrinisch mit einem Inhalt verbunden sind – „mental states have their contents essentially" (vgl. Psychosemantics, S.45) – müssen gleiche Geisteszustände auch gleiche (enge) Inhalte haben. Eine zentrale Rolle spielen hierbei die sogenannten „causal powers" (Kausalkräfte) von Geisteszuständen. Erstere sind nämlich das ausschlaggebende Kriterium für die Gleichheit bzw. Verschiedenheit von letzteren. Bei der Erläuterung dieser Kausalkräfte fragt sich Fodor zunächst, was denn überhaupt Vorbedingung sei für eine wissenschaftliche Psychologie. Nun sei es ein generelles Merkmal jeder Wissenschaft, Dinge aufgrund ihrer Kausalkräfte zu typologisieren: „What you need in order to do science is a taxonomic apparatus that distinguishes between things insofar as they have *different* causal properties, and that groups things together insofar as they have the *same* causal properties." (Psychosemantics, S.34) "Every (..) branch of science is in the business of causal explanation." (Psychosemantics, S. 33) Solange wir also an einer *wissenschaftlichen* Psychologie interessiert sind, muß sich eine Taxonomie von Geisteszuständen daher an deren Kausalkräften orientieren.

Betrachten wir zunächst einmal die Gehirnzustände (brain states) von mir und meinem Doppelgänger auf der Zwerde, so haben diese nach Annahme des Gedankenexperiments auch die gleichen Kausalkräfte. Wegen der Supervenienz des Mentalen über dem Physikalischen befinden sich daher auch ich und mein Doppelgänger im gleichen Geisteszustand, was wiederum Gleichheit des engen Inhalts zur Folge hat. Diese Supervenienz von Geist über Gehirn ist nun nach Fodor die einzige mögliche Erklärung dafür, wie auch Geisteszustände ihrerseits kausal wirksam werden können: „Mind/brain supervenience (and/or mind/brain identity) is, after all, the best idea that anyone has had so far about how mental causation is possible." (Psychosemantics, 30). Den Grund dafür kennen wir bereits. Schon weiter oben haben wir die Frage gestellt, wie denn etwas Mentales etwas Physikalische verursachen kann. Nach dem Schichtenmodell der Cognitive Science ist das Mentale nichts anderes als eine abstrakte Beschreibungsebene der zugrunde liegenden physikalischen Prozesse und in diese ohne Informationsverlust übersetzbar. Diese Übersetzbarkeit des Mentalen in das Phyikalische stellt einen Parallelismus zwischen den beiden Beschreibungsebenen her, wie er auch in dem Supervenienzprinzip zum Ausdruck kommt. Fodor faßt diesen Standpunkt auch wie folgt zusammen: „We have assumed a typology according to which the phyiological identity of organism guarantees the identity of their mental states (and, a fortiori, the identity of the contents of their mental states).

Nach dieser Auffassung haben – ich fasse zusammen - gleiche Geisteszustände in verschiedenen Kontexten den gleichen Inhalt. Der Inhalt der Geisteszustände wird also unter Ausklammerung der äußeren Umgebung festgelegt und aus eben diesem Grunde als enger Inhalt bezeichnet. Dies ist die Position des semantischen Internalismus: der Inhalt von Geisteszuständen wird ermittelt über eine Beschreibung dessen, was sich im Kopf eines Individuums abspielt.

Der Knackpunkt dieser Überlegungen ist nun der folgende: Fodor ist der Auffassung, dass eine Taxonomie von Geisteszuständen über ihre ‚causal powers' nur unter der Voraussetzung glücken kann, dass die Geisteszustände unter Absehung ihrer kontextuellen Eigenschaften individuiert werden. Mit anderen Worten: *Der semantische Internalismus erhält dadurch seine Legitimation, dass Geisteszustände über ihre Kausalkräfte individuiert werden.* Bevor ich

auf diesen wichtigen Punkt näher eingehe, möchte ich noch in einem kurzem Einschub abkläären, inwiefern die Kausalkräfte von mir und meinem Zwillingsbruder auch tatsächlich gleich sind (schließlich ist es diese Voraussetzung, auf die Fodor seinen semantischen Internalismus aufbaut).
Fodor konstruiert in diesem Zusammenhang den nachstehenden Einwand. Nehmen wir an, ich sage auf der Erde: Bring mir ein Glas Wasser! Diese Äußerung hat zur Folge, dass mir jemand H_2O bringt. Die gleiche Äußerung meines Zwillingsbruders auf der Zwerde hat jedoch zur Folge, dass ihm *XYZ* (also eine andere Substanz) gebracht wird. Somit würden sich auch die Kausalkräfte von mir und meinem Zwillingsbruder unterscheiden. Dagegen argumentiert Fodor folgendermaßen: Die Kausalkräfte von mir und meinem Zwillingsbruder wären nur dann verschieden, wenn sie im *gleichen* Kontext verschiedene Effekte auslösen könnten. Würde jemand beispielsweise auf dem Mond ein hundert Kilogramm schweres Gewicht stemmen, sein Zwillingsbruder auf der Erde dies aber nicht schaffen, so läge dies nicht an den unterschiedlichen Kausalkräften, sondern nur an den in den beiden Kontexten unterschiedlichen Schwerkraftsbedingungen.
(Welchen Sinn es allerdings macht, Kausalkräfte völlig losgelöst von ihrer Einbettung in eine reale Umgebung zu beschreiben, sei für den Augenblick dahingestellt. Ein möglicher Einwand gegen eine solche ‚individualistische' Erfassung von Kausalkräften wäre etwa der folgende: Eine Fußballmannchaft A gewinnt ein Spiel souverän gegen eine Mannschaft B und letztere gewinnt wiederum souverän gegen eine Mannschaft C. Wäre die Spielstärke der drei Mannschaften eine transitive Beziehung, so müsste in einem Spiel von A gegen C eindeutig der Gewinner feststehen. Nun gewinnt aber womöglich C gegen A. Der Grund hierfür liegt daran, dass sich spielerisches Können immer nur situativ entwickelt und daher auch nicht isoliert von seiner Einbettung in einen Kontext erfasst werden kann.)
Kommen wir nach diesem Einschub nun aber zur zentralen Frage: In welchem Sinne führt eine Taxonomie von Geisteszuständen über ihre Kausalkräfte zum semantischen Internalismus?
Eine klare Antwort auf diese Frage gibt Fodor in dem einige Jahre vor *Psychosemantics* erschienenen Artikel *Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology* (1980). Er argumentiert dort im Wesentlichen wie folgt: Alltagspsychologisch betrachtet schreiben wir jemandem einen Geisteszustand zu, um sein Verhalten voraussagen zu können. Weil Herr X sich beispielsweise vor etwas fürchtet respektive etwas erhofft, herbeiwünscht usw., wird er sich *so* oder *so* verhalten. Die Prognose seines Verhaltens ist jedoch völlig unabhängig vom Wahrheitsgehalt des Geisteszustands, in dem sich Herr X gerade befindet. Wir können und müssen sogar bei der Beschreibung seines Geisteszustands die Umgebung ausklammern, in die der Geisteszustand eingebettet ist. Nehmen wir mal an, Herr X sei paranoid und vermutet, dass ein ihm entgegenkommender Spaziergänger ein gefährlicher Mörder sei. Für die Prognose seines Verhaltens ist es nun aber völlig unerheblich, ob die Vermutung des Herrn X wahr oder falsch ist. Was wirklich zählt, ist nicht der Wahrheitswert seiner Vermutung, sondern der Wahrheitswert der Beschreibung seiner Vermutung! Um das Verhalten des Herrn X voraussagen zu können, müssen wir schließlich nur wissen, ob unsere Zuschreibung ‚Herr X vermutet, dass..' wahr ist, nicht aber ob die Vermutung des Herrn X zurecht besteht oder nicht. Denn das Handeln von Herrn X hängt ja nur davon ab, was er subjektiv für wahr hält, nicht aber davon, was objektiv auch wahr ist.

Daraus ergibt sich: Wir können bei der Beschreibung von Geisteszuständen deren semantische Bewertung ausklammern. Satzkonstruktionen, in denen Geisteszustände beschrieben werden, bezeichnet man in der analytischen Sprachphilosophie auch als opake Satzkonstruktionen. Der Wahrheitswert opaker Satzkonstruktionen hängt nur davon ab, was das Subjekt unter jenen Ausdrücken versteht, die in diese Satzkonstruktion eingebettet sind. Betrachten wir hierzu das folgende (gängige) Beispiel. Es ist das Drama von Ödipus, an dem sich klassisch die Differenz zwischen objektiver Wahrheit und subjektiver Meinung demonstrieren läßt. Sagen wir beispielsweise, Ödipus glaubt, dass Jokaste die für ihn geeignete Ehefrau sei, so schreiben wir Ödipus einen bestimmten Geisteszustand zu. Der Wahrheitswert dieser Zuschreibung ist unabhängig davon, ob Jokaste tatsächlich die für ihn geeignete Ehefrau ist, ja selbst davon, ob Jokaste überhaupt existiert. Der Grund dafür liegt an der semantischen Rolle, die der Ausdruck ‚Jokaste' in einer derartigen opaken Satzkonstruktion hat. Was unter dieser semantischen Rolle zu verstehen ist, zeigt eine Gegenüberstellung von opaken Satzkonstruktionen und Sätzen, die Sachverhalte (und eben keine Geisteszustände) beschreiben. So hat auch der einfache Satz ‚Der Tisch ist viereckig' einen Wahrheitswert. Der Satz kann wahr oder falsch sein. Trifft der beschriebene Sachverhalt zu, so ist der Satz wahr, andernfalls ist er falsch. Nun gilt für derartige Sätze das folgenden: Ersetzen wir in einem solchen Satz einen Ausdruck durch einen andern, der das gleiche Referenzobjekt bezeichnet, so wird sich am Wahrheitswert des betreffenden Satzes nichts ändern. Ausdrücke mit gleicher Extension können ‚salva veritate' ausgetauscht werden. Gilt also ‚X = Tisch', so ist auch der Satz ‚X ist viereckig' wahr, wenn ‚Der Tisch ist viereckig' wahr ist. Die mit einem solchen Ersetzungsverfahren verbundene Logik wird auch als eine *extensionale Logik* bezeichnet. Diese Ersatzbarkeit von Ausdrücken gleicher Extension unter Beibehaltung des Wahrheitswertes gilt aber nicht für opake Satzkonstruktionen, was folgendes Beispiel belegt. Ist der Satz ‚Ödipus glaubt, dass Jokaste die für ihn geeignete Ehefrau ist' wahr, so folgt daraus keineswegs, dass auch der Satz ‚Ödipus glaubt, dass seine Mutter die für ihn geeignete Ehefrau ist' wahr sein muß, obwohl die Ausdrücke ‚Jokaste' und ‚Mutter von Ödipus' extensionsgleich sind. Der Grund dafür liegt an der semantischen Rolle der in den dass-Sätzen eingebettten Ausdrücke. So dient ‚Jokaste' beispielsweise in dem oben erwähnten Dass-Satz nicht als Bezeichnung eines Referenzobjekts (man beachte: die Ausdrücke ‚Jokaste' und ‚Mutter von Ödipus' bezeichnen das gleiche Referenzobjekt), sondern als Beschreibung der subjektiven Sicht, die Ödipus von Jokaste hat. Was in opaken Satzkonstruktionen mit ‚Jokaste' (oder anderen in die Dass-Kontruktion eingebetteten Ausdrücken) gemeint ist, ist die *Intension* des Ausdrucks – und eben nicht die *Extension*. So haben die Ausdrücke ‚Jokaste' bzw. ‚Mutter von Ödipus' zwar die gleiche Extension, sie haben aber eine verschiedene *Intension*. Man spricht in diesem Zusammenhang auch von einer intensionalen Logik – extensionsgleiche Ausdrücke sind nicht *salva veritate* austauschbar.

Eine solche intensionale Logik ist nun die unabdingbare Voraussetzung dafür, dass Geisteszuständen überhaupt Kausalkräfte zugeschrieben werden können. Ohne das begriffliche Instrumentarium opaker Satzkonstruktionen wäre das Drama von Ödipus sprachlich gar nicht ausdrückbar. Würden nämlich Ausdrücke, die in Nebensatzkonstruktionen eingebettet sind, nur zur Bezeichnung von Referenzobjekten dienen, dann hätten wir nicht die Möglichkeit, die Intensionsverschiedenheit der in den Dass-Konstruktionen eingebetteten Ausdrücke zum Ausdruck zu bringen. Dieser - nur in opaken Kontexten ausdrückbare - Bedeutungsunterschied ist indes relevant zur Erklärung des Verhaltens jener Person, der wir einen Geistes-

zustand zuschreiben. (Wenn wir wissen, daß Ödipus die Absicht hat, Jokaste zu heiraten, so lassen sich daraus bestimmte Handlungen ableiten. Zugleich wissen wir aber, daß er nicht die Absicht hat, seine Mutter zu heiraten und daher im letzteren Falle auch keine Heiratsvorbereitungen treffen würde.)

Opake Satzkonstruktionen haben aber noch eine weitere Eigenschaft und es ist diese Eigenschaft, die Fodor als Argument dafür verwendet, dass eine Taxonomie von Geisteszuständen über ihre Kausalkräfte zum semantischen Internalismus führt (also zur Individuation von Geisteszuständen unter Ausklammerung der äußeren Umgebung). Diese Eigenschaft ist die folgende: Ausdrücke, die in opaken Kontexten eingebettet sind, lassen keine Rückschlüsse zu auf die Existenz ihrer Referenzobjekte. Der Existenzquantor der Prädikatenlogik ist nicht auf opake Satzkonstruktionen anwendbar. Betrachten wir dazu das folgende Beispiel: Ist der Satz ‚Der Tisch ist viereckig' wahr, so ist auch der folgende Satz wahr: ∃X (X ist viereckig). Ist aber der Satz ‚Ödipus glaubt, dass Jokaste die für ihn geeignete Ehefrau ist' wahr, so folgt daraus keineswegs, dass der Satz ∃X (Ödipus glaubt, dass X die für ihn geeignete Ehefrau ist) auch wahr ist. Schließlich ist die Intension des Ausdrucks ‚Jokaste' unbetroffen davon, ob Jokaste überhaupt noch am Leben ist. ‚Jokaste' drückt in diesem opaken Kontext nur das aus, was Ödipus über Jokaste weiß, ist Ausdruck seiner subjektiven Meinung. Und sollte Jokaste eventuell mittlerweile verstorben sein, so muß Ödipus dieses Faktum ja nicht zwangsläufig wissen.

Nachdem nun Beschreibungen von Geisteszuständen opake Satzkonstruktionen sind, schließt Fodor daraus, dass in solchen Beschreibungen nur die *formalen* Eigenschaften von Geisteszuständen Berücksichtigung finden. Unter formalen Eigenschaften eines kognitiven Systems versteht er solche Eigenschaften, die sich unter Ausklammerung deren semantischen Eigenschaften formulieren lassen. Er denkt hier insbesondere an die semantischen Eigenschaften von Wahrheit und Bezugnahme. Ob eine Person tatsächlich etwas wahrnimmt (vermutet, glaubt usw.) oder lediglich halluziniert, spielt, sofern wir allein an einer Beschreibung ihres Geisteszustands interessiert sind, keine Rolle. Dies liegt eben daran, dass Ausdrücke, die in opaken Kontexten verpackt sind, nicht in einem referierenden Sinne verwendet werden.

Diese formale Betrachtung mentaler Prozesse führt dann auch unmittelbar zu einer komputationalen Theorie des Geistes: Mentale Prozesse werden gleichgesetzt mit formalen Operationen von Symbolen eines Computers. Fodor führt als Beispiel ein Programm von Terry Winograd an, das einen Roboter simuliert. Der simulierte Roboter bewegt sich in einer Welt virtueller Blauklötzchen, die sich in bezug auf ihre Anordnung, Form, Größe und Farbe voneinander unterscheiden. In einem quasi natürlichsprachigen Dialog via Tastatur und Bildschirm vermag das Programm Anweisungen des Benutzers befolgen, Fragen beantworten, Strategien entwickeln, um die Blöcke neu anzuordnen, und neue Wörter lernen. Dies geschieht auf eine Art und Weise, die verblüffend natürlich wirkt. Sagen Sie beispielsweise dem Programm, stelle den zweiten kleineren Block auf den ersten größeren Block, so führt das Programm diese Anweisung aus. Nach einer weiteren Anweisung, aufgrund derer ein noch kleinerer dritter Block auf den zweiten Block gestellt wird, teilen Sie dem Programm mit, dass die drei aufeinander gestellten Blöcke ein Turm sind. Das Programm ist nun dazu in der Lage, sich diesen neuen Begriff zu merken und in der Folge dann auch die Anweisung ‚Baue einen Turm' auszuführen.

Die eigentliche Pointe dieses Beispiels, jedenfalls wie sich dies aus der Perspektive Fodors darstellt, ist nun aber die, dass weder der Roboter noch die Bauklötzchen real existieren. Fo-

dor meint dazu, Winograds Programm verhalte sich wie ein Computer, der nur träumt, ein Roboter zu sein. Alle Vorstellungen, die der Computer über Bauklötzchen hat, sind daher nur fiktiv. Schließlich fehlt die Einbettung in eine reale Umgebung. Aus diesem Grunde habe der Computer auch nur Zugang zu den formalen Eigenschaften der symbolischen Repräsentationen, die er verarbeitet (vgl. Fodor, Methodological Solipsism, S. 232).
Eine ganz ähnliche formale Betrachtungsweise läge nun dann vor, wenn wir Geisteszustände beschreiben. Denn der Wahrheitswert der Beschreibung von Geisteszuständen sei schließlich unabhängig von der semantischen Bewertung dieser Geisteszustände. Wenn sich jemand vor seinem Nachbarn fürchtet, weil er diesen für einen Mörder hält, so ist es für eine Beschreibung seines Geisteszustands unerheblich, ob er sich dies nur einbildet oder ob seine Vermutung zu Recht besteht.
(Gegen diesen Vergleich eines Computersprogramms mit einem Menschen, der etwas halluzinert, läßt sich allerdings das folgende einwenden: Ein Mensch, der etwas halluziniert, hat nur deshalb eine Halluzination, weil er fälschlicherweise das Halluzinierte für real hält. Ein Programm, das dagegegen eine Bauklötzchenwelt simuliert, hat nicht – wie Fodor glaubt – eine falsche Vorstellung von Bauklötzchen. Es interpretiert die von ihm verarbeiteten Symbole gar nicht als Bauklötzchen. Das Programm hat nicht eine falsche semantische Bewertung, sondern gar keine.)
Fassen wir Fodors Überlegungen zum engen Inhalt mentaler Repräsentationen zusammen: Solange wir an einer Beschreibung von Geisteszuständen interessiert sind, können wir die Einbettung dieser Geisteszustände in eine reale Umgebung ausklammern. Und dies ist genau das, was nach der Ansicht von Fodor eine wissenschaftliche Psychologie auszeichnet. Die Psychologie, sofern sie jedenfalls den Anspruch erhebt, eine Wissenschaft zu sein, hat demzufolge von einem engen Begriff des Inhalts auszugehen.
Gegen diesen strikt engen Begriff des Inhalts wendet nun aber Fodor selbst das folgende ein: Wenn uns jemand seine privaten Überlegungen mitteilt, so sind wir in der Regel nicht nur an seinen subjektiven Vermutungen interessiert. Wir wollen dahingegen sehr wohl wissen, ob seine Vermutungen auch zutreffend sind. Nicht dem subjektiven Geisteszustand des Vermutenden gilt also unser Interesse, sondern den Argumenten, die uns der Betreffende durch seine Vermutungen mitteilen will. Fodor bezeichnet diese Auffassung – im Unterschied zu einer wissenschaftlichen Psychologie - als jene Auffassung, die dem *Common Sense* Rechnung trägt. Der *Common Sense* fragt nach der semantischen Bewertung von Geisteszuständen. Und eine semantische Bewertung kann nur unter Berücksichtigung der äußeren Umgebung erfolgen, in dem die Geisteszustände eingebettet sind.
Als Beispiel bringt Fodor eine Kriminalgeschichte von Sir Arthur Conan Doyle. Es handelt sich um die Geschichte vom gefleckten Band. Sherlock Holmes kommt in dieser Geschichte einem Mörder auf die Spur, der jemanden vergiftet hat. Die Schweirigkeit bei der Aufklärung des Mordes bestand darin, dass das Opfer sich in einem Raum aufgehalten hat, in dem die Türen und Fenster von innen verschlossen waren. Bei der Begutachtung des Raumes entdeckte Holmes zunächst einen Ventilator. Unmittelbar über dem Bett des Ermordeteen befand sich zudem ein Seil mit einem Klingelzug, durch den man die Dienerschaft herbeiläuten konnte. Auffällig war, dass der Klingelzug bei näherer Betrachtung nur eine Attrappe war. Darüber hinaus konnte Holmes noch feststellen, dass das Bett am Boden festgeschraubt war. Aufgrund verschiedener Überlegungen kam Sherlock Holmes zu dem Schluß, dass der Tod durch eine giftige Schlange herbeigeführt wurde, die sich über den Ventilatorschaft und das

Seil nach unten auf sein Opfer herabgelassen hat. Holmes Schlußfolgerungen waren in etwa die folgenden: *Es wurde mir klar*, dass die drohende Gefahr nur von jemandem ausgehen konnte, der weder durch die Fenster noch durch die Türen kam. *Meine Aufmerksamkeit* wurde sofort von dem Ventilator und dem Seil, das als Klingelzug dienen sollte, angezogen. *Ich kam dann zum Schluß*, dass es sich hier nur um eine giftige Schlange handeln konnte.
Fodor kommentiert diese Gedankenkette nun folgendermaßen: Hören wir solche oder ähnliche Überlegungen, so werden wir sie nicht nur als Beschreibung des subjektiven Geisteszustands von Sherlock Holmes, sondern als Argumente interpretieren, mit denen er uns überzeugen will. Die Schlüssigkeit seiner Argumente hängt ab vom Wahrheitswert der von ihm geäußerten Überzeugungen.
Seine Geisteszustände lassen sich also von zwei Seiten betrachten: Solange wir ausschließlich an ihren ‚causal powers' interessiert sind, können und müssen wir sogar ihre semantische Bewertung ausklammern. Im Interesse einer wissenschaftlichen Psychologie geht es schließlich allein um eine formale Beschreibung von Geisteszuständen, um eine Beschreibung von ‚trains of thoughts', unabhängig davon, ob die beschriebenen Überzeugungen und Schlußfolgerungen wahr oder falsch sind. Parallel zu dieser Betrachtungsweise fragt aber der Common Sense nach dem Wahrheitsgehalt der beschriebenen Geisteszustände. Diese Doppelbetrachtung von Geisteszuständen ist nun so lange kein Problem, als zwischen der psychologischen Betrachtungsweise und dem Commone Sense eine Art Parallelismus besteht. Fodor geht davon aus, dass ein derartiger Parallelismus in den meisten Fällen gegeben sei. Er drückt dies so aus: „Causal relations very often respect semantic ones" (Fodor, 1987, S. 13) Diese Forderung führt uns allerdings wiederum zurück zu einer zentralen Voraussetzung der Cognitive Science, dass nämlich semantische Beziehungen ‚nachgeäfft' (mimicked) werden können durch ihre syntaktischen Korrelate. Vorausgesetzt wird die Supervenienz der Semantik über der Syntax.
Es ist nun aber eben dieses Supervenienzprinzip, das durch Putnams Zwillingserdegeschichte in Frage gestellt wird. Vor der Zwillingserdegeschichte hatten wir folgende Situation: Sind zwei Organismen physiologisch identisch, so sind auch ihre Geisteszustände identisch. Identische Geisteszustände wiederum garantieren identische Inhalte. Aus dieser Inhaltgleichheit von Geisteszuständen folgt wiederum Extensiongleichheit. Kennen wir den Inhalt unserer Geisteszustände, dann wissen wir auch, wovon sie handeln. „If you know what the content of a belief is then you know what it is about in the world that determines the semantic evaluation of the belief." Diese Position entspricht in etwa der Darstellung der klassischen Bedeutungstheorie bei Putnam. Nach der Zwillingserdegeschichte haben wir nun aber ein Problem. Aus Inhaltsgleichheit folgt nämlich nicht mehr Extensionsgleichheit! Fodor drückt dies so aus: „The Twin-Earth Problem *is* a problem *because it breaks the connection between extensional identity and content identity*." (Fodor, 1987, S. 47).
Fodor versucht dieses Problem abzufangen, indem er den Begriff des engen Inhalts so umformuliert, dass er zwar Putnams Intuition bezüglich der Zwillingserde Rechnung trägt, gleichzeitig aber immer noch eine eindeutige Zuordnung von Intensionen und Extensionen ermöglicht. Bei der Umformulierung seines Begriffs des engen Inhalts geht Fodor in zwei Schritten vor. *Vor der Zwillingserdegeschichte* definiert er Inhalt als eine Funktion, die Geisteszuständen eindeutig eine Extension zuordnet. Diese Eindeutigkeit ist *nach der Zwillingserdegeschichte* aber nicht mehr gegeben (ich und mein Zwillingsbruder sind ex hypothesi im gleichen Geisteszustand, beziehen uns aber auf verschiedene natürliche Arten).

Um nun diese Eindeutigkeit auch nach der Zwillingserdegeschichte aufrecht zu halten, bedient sich Fodor eines etwas kniffligen Schachzugs. Dieser Schachzug besteht im wesentlichen darin, dass er seine Definition des engen Inhalts ein wenig modifiziert: Inhalt sei – nach der Zwillingserde (!) – eine Funktion, die Geisteszuständen *in Abhängigkeit vom Kontext* eine Extension zuordnet. Fodor hat bei dieser modifizierten Definition des engen Inhalts nichts anderes getan als den Funktionsbegriff um ein zweites Argument, nämlich den Kontext, erweitert.
(Auf diese Weise wird allerdings trivialerweise sichergestellt, dass aus Inhaltsgleichheit auch Extensionsgleichheit folgt, denn der in dieser zweiten Definition vorkommende Kontext als zusätzliches Argument in der Funktion ist ja ein extensionales Kriterium zur Bestimmung von Extension.)
Ich und mein Zwillingsbruder haben nach dieser Definition den gleichen engen Inhalt (der Repräsentation von Wasser z.B.), da wir im gleichen Kontext uns (aufgrund der Gleichheit unserer Geisteszustände) auch auf die gleiche natürliche Art beziehen (aus eben diesem Grunde verfügen wir auch über die gleichen Kausalkräfte, denn, wie wir bereits weiter oben gesehen haben, wären die Kausalkräfte von mir und meinem Zwillingsbruder nur dann verschieden, wenn sie *im gleichen Kontext* verschiedene Wirkungen hätten).
Der Unterschied der mentalen Repräsentationen von mir und meinem Zwillingsbruder sei dahingegen ein Unterschied im *weiten Inhalt*. Diesen erhalten wir, indem wir bei der funktionalen Zuordnung von Geisteszuständen zu Extensionen den Kontext fixieren. Die Verschiedenheit der Bedeutung von ‚Wasser' auf der Erde und der Zwerde sei daher auch eine Verschiedenheit des weiten Inhalts.
Mit dieser genaueren Definition des weiten und engen Inhalts lässt sich nun, wie bereits gesagt, das von Putnam beschriebene Dilemma der klassischen Bedeutungstheorie umgehen. Den beiden Prämissen dieser Theorie (wie sie jedenfalls von Putnam skizziert wurden) liegen nämlich zwei verschiedene Bedeutungen von Bedeutung zugrunde. Besagt die erste Prämisse, dass gleiche Geisteszustände auch gleiche Bedeutungen zur Folge haben, so trifft dies nach Fodor insofern zu, als wir unter diesen gleichen Bedeutungen gleiche enge Inhalte verstehen. Dies ist deshalb der Fall, da ich und mein Zwillingsbruder uns – in Abhängigkeit vom Kontext – auf die gleiche natürliche Art beziehen. Besagt dahingegen die zweite Prämisse, dass bei Extensionsverschiedenheit auch die Bedeutungen verschieden sein müssen, so wäre in diesem Falle der weite Inhalt mentaler Repräsentationen gemeint. Die beiden von Putnam vorgestellten Prämissen beziehen sich auf verschiedene Bedeutungsbegriffe und können daher auch nicht zu jenen Widersprüchen führen, wie sie von Putnam aufgezeigt wurden.
Bei der Bestimmung des weiten Inhalts mentaler Repräsentationen hat Fodor nun aber ein Problem. Da zu seiner Bestimmung der Kontext fixiert werden muß und der Kontext ein extensionales Kriterium ist, läßt sich der weite Inhalt auch nicht über den individuellen physiologischen Zustand des Kognizierenden festlegen. Damit geht aber auch die Rechnung einer an den Naturwissenschaften orientierten Psychologie, über die ‚mind/brain supervenience' (Fodor, 1987: 42) auch den Inhalt von Geisteszuständen erfassen zu können, nicht auf. *Eine naturalistische Erklärung des weiten Inhalts kann daher nur unter Mitberücksichtigung der äußeren Umgebung erfolgen, in die ein Organismus eingebettet ist.* Die Erklärung darf allerdings in ihrem erklärenden Teil, wenn sie nicht zirkulär sein will, keine semantischen Termini enthalten: „If aboutness is real it must be really something else." (97)

Eine solche naturalistische Bedeutungstheorie haben wir bereits kennengelernt. Bereits Harnad hat versucht, die Bedeutung von Symbolen in subsymbolischen Repräsentationen zu gründen, die nur in einer kausalen Beziehung zur Außenwelt und nicht in einer semantischen Beziehung stehen. Harnads Unternehmen wurde dann allerdings unter Verweis auf Quines Unbestimmtheit der Bedeutung ausgehebelt. Bedeutung lässt sich nicht reduktiv über sensorischen Reize erklären, da bei gleichen Reizbedingungen (Strahlungsmuster auf der Retina) verschiedene Bedeutungen möglich sind. Auch Fodor schlägt für seinen weiten Inhalt mentaler Repräsentationen eine Kausaltheorie der Bedeutung vor. Ob und inwiefern er es besser macht als Harnad, sehen wir im nachstehenden Text.

Fodors Kausaltheorie der Bedeutung

Fodors Bedeutungstheorie handelt nicht von sprachlichen Zeichen. Vielmehr geht es um mentale Repräsentationen in unserem Kopf. Denn die Bedeutung sprachlicher Zeichen, sei dies in geschriebener oder ausgesprochener Form, ist immer nur eine abgeleitete Bedeutung. Ferner schränkt Fodor seine Bedeutungstheorie auf atomare Repräsentationen ein. Die Bedeutung komplexer Repräsentationen konstituiert sich über die Bedeutung atomarer Repräsentationen. So läßt sich etwa die mentale Repräsentation eines Zebra zurückführen auf die Repräsentation von Vierfüßler und gestreift etc. Die konkrete Frage ist also, wie die atomaren Symbole in unserem Kopf zu ihrer Bedeutung kommen.
Offensichtlich darf bei einer solchen Erklärung im erklärenden Teil nicht mehr ein Verstehen von Bedeutung vorausgesetzt werden. Andernfalls wäre die Erklärung ja zirkulär. Die Termini, die wir verwenden, um Bedeutung zu erklären, dürfen also keine semantischen Termini sein. Es geht somit darum, semantische Termini vollständig durch nicht semantische Termini zu ersetzen. Und genau dies soll die Kausaltheorie der Bedeutung leisten. Dabei soll der Ausdruck ‚x bedeutet y' durch den Ausdruck ‚x wird durch y verursacht' *ersetzt* werden. In dem Satz ‚x wird durch y verursacht' kommen keine semantischen Termini mehr vor, es wird ein rein physikalisches Geschehen beschrieben. Wenn diese Ersetzung ohne Informationsverlust durchführbar ist, dann hat die Kausaltheorie der Bedeutung ihren Zweck erfüllt.
Fangen wir mit einer ganz groben Definition an. Diese Definition könnte so lauten: eine atomare mentare Repräsentation bedeutet all das, wodurch sie ausgelöst wurde. Die Definition ist nicht zirkulär, denn wir haben den Ausdruck ‚bedeutet' ersetzt durch ‚wurde ausgelöst durch'.
In Zusammenhang mit dieser kruden Kausaltheorie stellen sich jedoch zwei Probleme: Wie steht es um mögliche Fehlauslöser? Gelegentlich schnappt ein Frosch statt nach einer Fliege nach einer Fliegenattrappe. Bedeutet die mentale Repräsentation, die der Frosch von Fliege hat, jetzt auch Fliegenattrappe? Nach der kruden Kausaltheorie müßte dies der Fall sein. Denn eine mentale Repräsentation bedeutet schließlich all das, wodurch sie ausgelöst wurde. Wird die mentale Repräsentation daher durch eine Fliegenattrappe ausgelöst, so bedeutet sie auch Fliegenattrappe. Die Fliegenattrappe wäre laut der kruden Kausaltheorie kein Fehlauslöser. In dieser Theorie sind Fehlauslöser gar nicht vorgesehen, denn der Ausdruck ‚bedeutet' wurde schlicht und einfach durch den Ausdruck ‚wurde ausgelöst durch' ersetzt. Eine Unterscheidung von Fehlauslösern und echten Auslösern ist auf diese Weise überhaupt nicht möglich. Um nämlich zwischen echten Auslösern und Fehlauslösern zu unterscheiden, müßten wir wiederum semantische Termini verwenden, die uns aber im Rahmen einer rein physikalischen

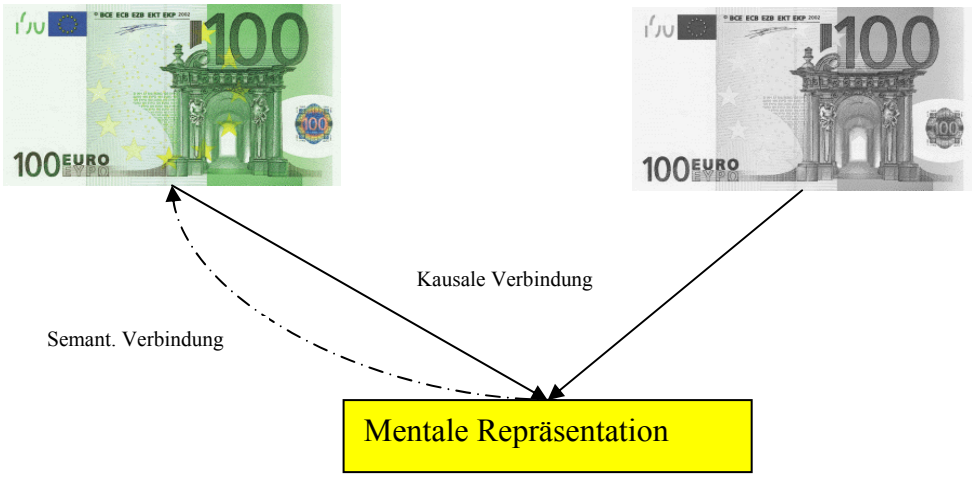
Kausaltheorie nicht zur Verfügung stehen. So wäre es etwa verfehlt, eine Fliegenattrappe deshalb als Fehlauslöser zu bezeichnen, da die mentale Repräsentation für Fliegen schließlich nur Fliegen *bedeutet* – und eben nicht Attrappen von Fliegen. Wie soll jetzt aber der Unterschied zwischen Fehlauslösern und echten Auslösern auf der Ebene der physikalischen Kausalität erklärt werden können?

Zu diesem ersten Problem unserer kruden Kausaltheorie gesellt sich noch das folgende: Definieren wir die Bedeutung einer atomaren mentalen Repräsentation über all das, wodurch sie ausgelöst wurde, so würden wir deren Bedeutung einschränken auf all jene Auslöser, die faktisch diese Repräsentation hervorgerufen haben. Denken wir beispielsweise an die mentale Repräsentation für Hasen. Verwenden wir keine Kausaltheorie und bemühen zunächst einmal unsere intuitive Semantik, so wissen wir, dass unsere mentale Repräsentation nicht nur jene Hasen bedeutet, die uns zufällig einmal in unserem Leben begegnet sind, sondern *alle* Hasen, also auch solche, denen wir noch niemals begegnet sind und denen wir auch niemals begegnen werden. Schränken wir daher die Bedeutung der mentalen Repräsentation auf jene Hasen ein, denen wir faktisch begegnet sind, so beschränken wir deren Extension auf eine Teilmenge jener Hasen, die die mentale Repräsentation eigentlich bedeutet.

Fassen wir zusammen: Die krude Kausaltheorie geht von zwei unplausiblen Voraussetzungen aus. Sie kann nämlich nur dann stimmen, wenn die mentale Repräsentation *nur* von Hasen (dann haben wir keine Fehlauslöser) und von *allen* Hasen ausgelöst würde. Beide Einschränkungen sind aber allerhöchst unrealistisch. Ich nenne die beiden Einschränkungen der Kürze halber das ‚nur'- und das ‚alle'-Problem. Es stellt sich die Frage, wie man die Kausaltheorie so modizifieren kann, dass sie einerseits auf rein naturalistische Weise das Problem der Fehlauslöser lösen kann und andererseits auch zu keiner Einschränkung in der Extension unserer mentalen Repräsentationen führt. Eine solche modifizierte Kausaltheorie bezeichnet Fodor als eine 'slightly less crude causal theory of content'.

Beginnen wir mit der ersten Korrektur der kruden Kausaltheorie: Wie können wir echte Auslöser von Fehlauslösern so unterscheiden, dass wir zur Beschreibung dieses Unterschieds keine semantischen Termini benötigen? Fodor präsentiert die Lösung zu diesem Problem in zwei Schritten. Im ersten Schritt beschreibt er den Unterschied zwischen Fehlauslösern und echten Auslösern noch unter Zuhilfenahme semantischer Termini. Dieser Schritt dient daher auch nur dazu, um das anstehende Problem deutlich zu machen. Erst in einem zweiten Schritt versucht Fodor dann, diese semantische Beschreibung in einem rein naturalistischen Vokabular zu reformulieren.

Den ersten Schritt kann man sich in etwa so vorstellen. Nehmen wir als Beispiel die mentale Repräsentation für einen 100 Euroschein. Stellen wir uns dazu vor, vor uns lägen zwei Scheine. Der eine Schein sei ein echter 100 Euroschein während es sich bei dem zweiten Schein um eine Fälschung handle. Nun lösen beide Scheine die mentale Repräsentation für 100 Euro aus. Worin besteht der Unterschied zwischen beiden Auslösern? Vergleicht man die beiden kausalen Verbindungen, die jeweils die mentale Repräsentation auslösen, so unterscheiden sie sich dadurch, dass zwischen einem echten 100 Euro Schein und der betreffenden mentalen Repräsentation *zusätzlich* zur kausalen Verbindung noch eine *semantische* Verbindung besteht (denn die mentale Repräsentation bedeutet eben echte Scheine aber nicht falsche Scheine).



Fodor behauptet nun, dass die kausale Verbindung, die zwischen der mentalen Repräsentation und falschen Scheinen besteht, *parasitär* sei gegenüber der Verbindung von echten Scheinen und ihrer Repräsentation: „It's an old observation – as old as Plato, I suppose – that falsehoods are ontologically dependend on truths in a way that truths are not ontologically dependent on falsehoods." (Psychosemantics, 107). Man versteht dies gleich, wenn man sich die folgende Situation vergegenwärtigt. Denken wir zurück an die Zeit vor Einführung des Euro als gängiger Währung. In dieser Zeit hätten wir niemals fälschlicherweise einen unechten 100 Euro Schein akzeptiert. Um überhaupt falsche Scheine für echte Scheine zu halten, müssen also erst echte Scheine im Umlauf sein. Das umgekehrte ist allerdings nicht der Fall. In einer Welt, in der überhaupt kein Falschgeld vorkommt, können dennoch echte 100 Euro Scheine im Umlauf sein!

Als nächstes stellt sich Fodor dann die Aufgabe, diesen Unterschied in der kausalen Verbindung von echten und falschen Scheinen zu ihrer mentalen Repräsentation in einem nicht semantischen Vokabular zu reformulieren. Demnach sei die zweite Verbindung (Falschgeld) asymmetrisch abhängig von der ersten Verbindung (echtes Geld). Das bedeutet: Die kausale Verbindung eines falschen 100 Euro Scheins zur mentalen Repräsentation ‚100 Euro Schein' setzt das Bestehen einer kausalen Verbindung von echten Scheinen zu ‚100 Euro Schein' voraus – das umgekehrte ist aber nicht der Fall. Diese asymmetrische Abhängigkeit läßt sich nach der Auffassung Fodors in einem rein naturalistischen Vokabular formulieren, da hier nurmehr von kausalen Verbindungen und eben nicht von semantischen Verbindungen die Rede ist.

Ich werde weiter unten allerdings zeigen, dass auch der Versuch, über die asymmetrische Abhängigkeit das Problem der Fehlauslöser allein auf der Basis von nicht semantischen (kausalen) Ausdrücken erklären zu können, zum Scheitern verurteilt ist, dass also auch bei diesem naturalistischen Erklärungsansatz wiederum gleichsam durch die Hintertür semantische Termini hereinkommen. Zuvor aber möchte ich noch auf das zweite Problem der kruden Kausaltheorie eingehen, nämlich auf das vorher erwähnte ‚alle'-Problem.

Wie kann eine mentale Repräsentation all das bedeuten, wodurch sie ausgelöst wurde, ohne dass es dadurch zu einer Einschränkung ihrer Extension auf ihre faktischen Auslöser kommt? Fodor selbst hält dieses Problem für wesentlich schwieriger als die Klärung von Fehlauslösern über die asymmetrische Dependenz. Er erwägt als erstes den folgenden Lösungsvorschlag:

Zur Extension der mentalen Repräsentation ‚Hase' gehören nicht nur jene Hasen, die faktisch diese Repräsentation ausgelöst haben, sondern alle Hasen, die ‚Hase' in unserem Kopf *auslösen würden*. Fodor versucht also das ‚alle'-Problem unter Zuhilfenahme von kontrafaktischen Konditionalsätzen in den Griff zu bekommen. Woher können wir aber wissen, so ist jedenfalls zu fragen, dass auch alle Vorkommnisse von Hasen dieser Aufgabe nachkommen *würden*? Für eine naturalistische Bedeutungstheorie müßten alle Hasen auf naturgesetzmäßige Weise ‚Hase' hervorrufen (was ja durch den kontrafaktischen Konditionalsatz zum Ausdruck kommt). Von einer derartigen naturgesetzlichen Verursachung könnte nur dann die Rede sein, wenn alle Vorkommnisse von Hasen auch *verläßlich* die Repräsentation ‚Hase' auslösen würden. Eine solche Verläßlichkeit ist jedoch für das angeführte Beispiel kaum nachweisbar.
Um dieses Problem zu lösen, unterscheidet Fodor zunächst drei verschiedene Typen von mentalen Repräsentationen: Theoretische, beobachtungsferne Termini (z.B. Protonen, Quarks usw.), die nicht unmittelbar wahrgenommen werden können, sondern nur im Kontext einer wissenschaftlichen Theorie indirekt faßbar sind, mittelgroße Objekte (dazu gehören alle in diesem Text bislang aufgeführten Beispiele, also Hasen, 100 Euroscheine usw.) und schließlich drittens Beobachtungstermini. Letztere sind Ausdrücke der Psychophysik. Und nur auf sie hält Fodor seine Kausaltheorie für unmittelbar anwendbar. Denn nur Daten auf der Ebene der Psychophysik würden nämlich *verläßlich* ihre entsprechenden Repräsentationen auslösen. Er denkt hier an Ausdrücke der unmittelbaren Sinneswahrnehmung, beispielsweise die Wahrnehmung von rot. Unter optimalen Beleuchtungsbedingungern kommen wir quasi automatisch und ohne lange darüber zu reflektieren zur Beobachtung ‚rot'. Fodor meint dazu, dass unter optimalen Beleuchtungsbedingungen „the Mentalese equivalent of ‚red there' will get stuffed into your belief box *willy-nilly*". (Psychosemantics, 112)
Wir werden die Farbe auch nicht *als* rot interpretieren. Fodor unterscheidet hier zwischen ‚seeing' und ‚seeing as' bzw. zwischen ‚imaging' und ‚judging'. Die jeweils erstgenannten Ausdrücke beziehen sich auf Beobachtungstermini, letztere auf Termini für mittelgroße Objekte und theoretische Termini. Dazu ein etwas drastisches Beispiel: Jemand legt seine Hand auf eine heiße Herdplatte, zieht sie dann gleich zurück und schreit: Heiß! Die anschließende Frage, bist Du Dir dessen sicher, dass Du Deiner Sinneswahrnehmung auch den richtigen Begriff zugeordnet hast (dass Du also den Wärmegrad der Kochplatte richtigerweise *als* heiß beschrieben hast), würde in diesem Falle mit Recht nur Unverständnis hervorrufen. Auch das mentalesische Äquivalent von ‚heiß' geht ‚willy-nilly' in unsere belief box! Die Sinneswahrnehmung einer heißen Kochplatte löst verläßlich die Repräsentation ‚heiß' aus. Im Falle von Termini der Psychophysik können wir daher getrost die weiter oben angegebenen konditionalen Satzkonstruktionen benützen, um das ‚alle' Problem zu lösen: Zur Extension von ‚rot' (‚heiß') gehört all das, was *verläßlich* die entsprechende Repräsentation auslösen würde. Und eine solche Verläßlichkeit ist eben nur bei psychophysikalischen Termini gegeben. Es sind daher die Ausdrücke der Psychophysik, auf die Fodor letzten Endes seine ganze Kausaltheorie der Bedeutung aufbaut. Andereseits stammen so gut wie alle Beispiele, die er zur Beschreibung der asymmetrischen Dependenz verwendet, aus dem Bereich von Terminie für mittelgroße Objekte. Damit auch diese innerhalb seiner naturalistischen Bedeutungstheorie plaziert werden können, müssen diese – so jedenfalls Fodors Schachzug – in den psychophysikalischen Ausdrücken gegründet werden. Fodor sieht hier eine kausale Verankerung aller Termini in den Ausdrücken der Psychophysik: „The causal chain runs from horses in the world to hor-

sy looks in the world to psychphysical concepts in the belief box to ‚horse' in the belief box. 'Horse' means *horse* because that chain is reliable." (Psychophysics, 122)
Aber gerade diese Verankerung aller Termini in der Psychophysik bringt mich nun zu folgendem kritischen Schluß. Dazu müssen wir uns nochmals der Problematik der Fehlrepräsentationen zuwenden. Daniel Dennett hat einmal in seinem Buch *Intentional Stance* behauptet, jedes eingehende Signal könne so weit *deinterpretiert* werden, dass wir niemals einen Fehler machen werden. Er meinte damit das folgende: Nehmen wir an, ein Frosch verschluckt statt einer Fliege eine Fliegenattrappe. Dies läßt uns zunächst vermuten, der Frosch habe fälschlicherweise die Attrappe für eine Fliege gehalten. Diese Vermutung entspricht jedoch nicht korrekt dem, was der Frosch unmittelbar wahrnimmt. Um seiner Wahrnehmung annähernd gerecht zu werden, müssen wir nämlich das im Froschauge eingehende Signal *deinterpretieren*. Was das Froschauge tatsächlich wahrnimmt, ist ein dunkler, sich bewegender Fleck von einer bestimmten Gestalt. Die Behauptung, der Frosch nehme diesen schwarzen Fleck *als* Fliege wahr, stützt sich ja nur auf unsere eigene Interpretation der Wahrnehmung des Frosches. Lassen wir diese Interpretation beiseite und reduzieren wir das in das Froschauge eingehende Signal auf das, was der Frosch tatsächlich wahrnimmt, nämlich einen schwarzen Fleck, so wird der Frosch auch niemals einen Fehler machen! Der Fehler kommt ja nur dadurch zustande, dass wir den schwarzen Fleck *als* Fliege interpretieren. Die Unterscheidung von echten Auslösern und Fehlauslösern setzt also bereits eine ganz bestimmte Interpretationsebene voraus.
Kehren wir vor diesem Hintergrund zu unserem Beispiel eines falschen 100 Euroscheins zurück. Wie wir mittlerweile wissen, gründet Fodor seine ganze Kausaltheorie in den psychophysikalischen Termini. Auf der Beschreibungsebene der Psychophysik ist es aber kein 100 Euroschein, der wahrgenommen wird, sondern lediglich ein Stück Papier, das sich so und so anfühlt, so und so ausschaut usw. Zu einem Fehlauslöser kann es daher erst dann kommen, wenn wir diese psychophysikalischen Prädikate *als* 100 Euroschein zu interpretieren beginnen! Damit kann es aber auf der Beschreibungsebene der Psychphysik auch nicht zu einer asymmetrischen Dependenz der Fehlauslöser von echten Auslösern kommen. Eine solche Asymmetrie entsteht erst dann, wenn wir unserer visuellen und taktilen Wahrnehmung eine ganz bestimmte Bedeutung zuschreiben, wenn wir also das zunächst wahrgenommene Stück knisternden Papiers hinaufstufen zu einem 100 Euroschein.
Damit ist aber Fodors Kausaltheorie der Bedeutung in ihrem Anspruch, eine rein naturalistische Erklärung von Bedeutung geben zu können, gescheitert. Auch bei dieser Erklärung bleibt ein unerklärter semantischer Rest. Kann die semantische Lücke zwischen menschlichem Verstehen und naturwissenschaftlicher Erklärung jemals geschlossen werden?

Ein neues Forschungsparadigma kündigt sich an

Im folgenden versuche ich die kritischen Einwände gegen den Versuch, Semantik und Bezugnahme (reference) in einer naturalistischen Terminologie beschreiben zu können, zusammenzufassen. Ich stütze mich hierbei in erster Linie auf Argumente, wie sie von Putnam vorgetragen wurden. Dabei geht es mir hauptsächlich darum, den geistigen Nährboden aufzuzeigen, der hinter diesen kritischen Einwänden steht. So möchte ich zeigen, wie Putnams Kritik an der Naturalisierung von Bezugnahme von einer ganz bestimmten kognitionswissenschaftli-

chen Sicht ausgeht, nämlich dem Repräsentationalismus. Diese Einschränkung kann dann zugleich als Motiv herangezogen werden, um das Projekt der Künstlichen Intelligenz unter Nutzung eines gänzlich anderen und neuen Forschungsparadigmas fortzuführen und vielleicht sogar eines Tages zu vollenden. Mit welchen Problemen sich dann aber auch dieses neue Forschungsparadigma konfrontiert sieht, gehört nicht hierher und wird an späterer Stelle erst erläutert.

Was spricht also gegen eine Naturalisierung von Bezugnahme? Betrachten wir hierzu das folgende kleine Beispiel, das Putnam in seinem Buch *Renewing Philosophy* angeführt hat. Stellen Sie sich vor, Sie würden einem Hund ein Stück Fleisch geben, das zur Gänze aus pflanzlichen Proteinen besteht. Nehmen wir an, dieses synthetische Fleisch würde exakt die gleichen Funktionen erfüllen, die auch echtes Fleisch erfüllt. Eine Unterscheidung von echtem und synthetischem Fleisch wäre für den Hund also überhaupt nicht möglich. Putnam meint dazu, dass es sich aus dem Blickwinkel eines Hundes damit auch schlicht und einfach um Fleisch handle. Denn für einen Hund sei nur entscheidend, dass er einen ‚successful belief' hat.

Vergleicht man die Situation jetzt aber mit einem Menschen, der einen solchen Tofu zu essen bekommt. Im ersten Moment wird er aufgrund des ähnlichen Geschmackes annehmen, dass er es hier mit Fleisch zu tun hat. Teilt man ihm dann aber mit, dass seine Speise in Wirklichkeit Tofu war, wird er seine erste Annahme sofort korrigieren. Dies liegt daran, dass der Mensch die Fähigkeit hat, unterscheiden zu können zwischen dem, was sich nur so anfühlt und so ausschaut wie Fleisch, und echtem Fleisch. Putnam bezeichnet dieses kognitive Vermögen als die Fähigkeit „succesful belief" und „true belief" auseinander halten zu können. Der Mensch ist eben dazu in der Lage, Schein und Sein, Illusion und Wirklichkeit gegeneinander abzugrenzen. Für den Hund hingegen ist der Begriff ‚Fleisch' *referentiell unbestimmt*. Aus seinem Blickwinkel ist überhaupt nicht festgelegt, ob er ein Stück echtes Fleisch oder synthetisches Fleisch vor sich hat.

Was hier referentiell unbestimmt bedeutet, läßt sich an einem kleinen Gedankenexperiment von Daniel Dennett anschaulich darstellen. Es handelt sich hierbei um das Gedankenexperiment von dem „Two Bitser". Dies sei ein Münzapparat hergestellt in den USA mit dem Zweck sogenannte „Two Bits" – das sind 25 Cent Stücke in US Währung – ‚erkennen' zu können. ‚Erkennen' heißt hier nur so viel, nach der korrekten Eingabe eines 25 Cent Stückes etwas (Zigaretten, Kaugummi oder was auch immer) auszugeben. Was ist aber unter einer korrekten Eingabe zu verstehen? Unser Apparat funktioniert nur dann korrekt, wenn er *alle* 25 Cent Münzen und *nur* 25 Cent Münzen akzeptiert (man denke an Fodors Kausaltheorie!). Verwirft unser Münzapparat eine echte 25 Cent Münze (akzeptiert also nicht *jede* 25 Cent Münze) oder akzeptiert er umgekehrt eine Fälschung (akzeptiert also nicht *nur* echte 25 Cent Münzen), so wird er einen Fehler machen.

Nehmen wir nun an, unser Münzapparat wird nach Panama exportiert. In der panamesischen Währung gäbe es sogenannte Viertel-Balboas, die in ihren Oberflächeneigenschaften (gleiche Dicke, gleiches Gewicht usw.) von einem Two-Bit in den USA nicht unterschieden werden können. Aufgabe unseres Two-Bitsers sei in Panama, nicht mehr 25 Cent Stücke, sondern eben Viertel-Balboas zu akzeptieren. Schließlich sind letztere in Panama die gültige Währung. Und nachdem die beiden Münzen überhaupt nicht voneinander unterschieden werden können, ist es auch nicht erforderlich, unseren Münzapparat nach dem Transport nach Panama nachzujustieren. Er ist im Gegenteil dort sofort betriebsbereit. Immer dann, wenn in Panama ein

Viertel-Balboa eingeworfen wird, wird der Apparat panamesische Zigaretten ausgeben. Dennett stellt sich jetzt die folgende Frage: Wirft jemand in Panama statt eines Viertel-Balboas einen Viertel Dollar ein, wird die Maschine dann einen Fehler machen, wenn sie diese Münze akzeptiert? Schließlich soll die Maschine in Panama ja den Zweck erfüllen, Viertel Balboas und eben nicht Viertel Dollars zu akzeptieren. Man kann sich aber auch den umgekehrten Fall vorstellen: Wirft jemand in den USA einen Viertel Balboa ein, wird die Maschine dann einen Fehler machen, wenn sie diese panamesische Münze akzeptiert? Ist das, was in den USA als Fehler zählt, in Panama die korrekte Eingabe und umgekehrt dajenige, das in Panama als Fehler zählt, in den USA die korrekte Eingabe? Dennett hat mit Absicht ein derartiges Beispiel gewählt. Er will damit sagen, dass im Falle eines solchen Münzapparates die Bedeutung der eingegebenen Münze (Viertel Dollar oder Viertel Balboa) *referentiell unbestimmt* ist. Bestimmt wird die Bedeutung erst durch unsere Zuschreibung. Was also als Fehler zählt und was als echte Münze, hängt von den Interessen des sozialen Umfeldes ab, das die Maschine verwendet und kann vom Standpunkt der Maschine aus betrachtet gar nicht beantwortet werden. Denn die Maschine ist schließlich nur ein Artefakt und die Bedeutung, die wir ihren phyikalischen Zuständen zuschreiben, ist immer nur eine *zugeschriebene*, also abgeleitete Bedeutung!
Dennetts Beispiel erinnert an mein weiter oben verwendetes Beispiel eines nach China transportierten Gemäldes. Ich habe dort behauptet: Nicht die im Gemälde verwendete Technik der perspektivischen Verkürzung bedeutet *intrinisch* räumliche Distanzen, sondern nur wir selber meinen mit der perspektivischen Verkürzung räumliche Distanzen. Denn das Gemälde sei schließlich nur ein Artefakt, dessen physikalische Eigenschaften referentiell unbestimmt sind.
Diese Beispiele lassen vermuten, dass referentielle Bestimmtheit grundsätzlich nicht den physikalischen Zuständen eines Artefakts zugesprochen werden kann. Vor dem Hintergrund dieser Überlegungen stellt uns Dennett vor die folgende Alternative: Entweder deuten wir unser eigenes kognitives System als ein (naturalistisch rekonstruierbares) physikalisches Artefakt, also als eine Art „Two-Bitser“ mit einer höheren Komplexität, und nehmen in Folge dieser Deutung die prinzipielle referentielle Unbestimmtheit unserer kognitiven Zustände in Kauf oder aber wir glauben an referentielle Bestimmtheit und bezahlen dafür aber den Preis, diese Art von Semantik naturalistisch nicht erklären zu können. Dennett stellt uns vor diese Alternative, da (wie sein Two-Beispiel zeigen soll) unter den Auspizien einer naturwissenschaftlichen Betrachtungsweise Bedeutung immer unbestimmt bleibt.
Dennetts Argumente gehen letztlich auf Quines These der Unbestimmtheit der Übersetzung zurück. Ich habe dies schon weiter oben dargestellt: Bei gleichen Bestrahlungsmustern auf der Retina sind verschiedene Bedeutungszuschreibungen (in Quines Beispiel zu „Gavagai“) möglich. Wir benötigen als kompetente Sprecher Zusatzhypothesen, um überhaupt fixieren zu können, ob „Gavagai“ ganze Kaninchen, Kaninchenteile oder zeitliche Kaninchenstadien bezeichnet. Jeder Versuch, Bedeutungen über das beobachtbare Verhalten hinaus festzulegen, wird von Quine als ein schädlicher Mentalismus geahndet.
Fodor lehnt in seiner kausalen Bedeutungstheorie indes die beiden von Dennett und Quine vorgeschlagenen Alternativen ab. Er hält einerseits an der Bestimmbarkeit der Bedeutung (der Symbole in unserem Kopf) fest und unternimmt andererseits den Versuch, für diese Bedeutungsbestimmtheit eine naturalistische Erklärung zu finden. Dass dieser Vermittlungsansatz dann letzten Endes aber gescheitert ist, habe ich weiter oben dazulegen versucht.
Behalten wir diese kritischen Einwände gegen Bedeutungsbestimmtheit im Auge und fragen vor diesem Hintergrund nun nach Putnams Bedeutungsbegriff. Zunächst einmal müssen wir

festhalten, dass Putnams semantischer Externalismus von einer Bedeutungsbestimmtheit ausgeht. Ich meine damit das folgende: Das, was wir mit unseren Worten meinen (Wasser oder was auch immer), hat nicht nur eine *relative* Bedeutung – relativ zum jeweiligen Kontext, in dem eine Aussage getätigt wird – sondern bedeutet *wirklich* Wasser (oder was auch immer). Das heißt aber wiederum nicht, dass unser Meinen über ein Wissen über eine absolute Wirklichkeit verfügt, es bedeutet lediglich, dass wir bereit sind, unser momentanes Wissen über das Gemeinte zu korrigieren! Wir sind sozusagen ständig dazu bereit, unsere Zugangswege und Interpretationsraster der Wirklichkeit *korrigierend zu transzendieren*. Diesen Grundgedanken hatten wir schon. Es ist der Grundgedanke des internen Realismus. Warum aber, so ist nun zu fragen, kann eine Bedeutungsbestimmtheit in dem soeben erläuterten Sinne naturalistisch nicht erklärt werden?

Dazu müssen wir uns überlegen, wie denn diese korrigierende Transzendenz näherhin ausschaut. Wir transzendieren unsere Interpretationsraster der Wirklichkeit, unsere operationalen Verfahren zur Bestimmung natürlicher Arten, indem wir die Dinge von einer neuen Warte aus anschauen. Dies erfolgt nicht von einem Stanpunkt von nirgendwo, sondern aufgrund geänderter operationaler Verfahren und neuer Interpretationsraster. Denn unser Bezug zu den Dingen ist ja nur innerhalb eines bestimmten Interpretationsschemas festgelegt. Um nun diesen Bezug kritisch hinterfragen zu können, kann dies nicht innerhalb jenes Interpretationsschemas erfolgen, das den Bezug zuallererst festlegt, sondern nur aus dem Blickwinkel eines neuen Interpretationsschemas, das das ursprüngliche Schema transzendiert. Nachdem nun aber nur innerhalb der Grenzen unserer Interpretationsschemata festgelegt ist, worauf wir uns beziehen, können wir objektiv jenseits der Grenzen unserer Vernunft überhaupt nicht wissen, worauf wir uns beziehen.

Putnam meint hierzu das folgende. Eine Theorie, die Bezugnahme naturwissenschaftlich erklären will, kann in ihrem erklärenden Teil nicht nur innerhalb der Grenzen unserer ‚Heimatsprache' verbleiben (damit sind Begriffsraster gemeint, die wir verwenden, um die Welt in Arten einzuteilen), eine solche Theorie müsste dahingegen *jede* Art von Intentionalität, also auch eine *speziesunabhängige* Intentionalität erklären. Schließlich können wir nicht nur aus dem subjektiven Blickwinkel unserer eigenen Intentionalität letztere erklären. Um Intentionalität zu erklären, müßte selbst die Intentionalität intelligenterer Lebensformen als wir selber erklärt werden. Ganz ähnliche Argumente finden sich bereits bei Kant in seinen *Paralogismen der reinen Vernunft*: Übertragen wir die Bedingungen, unter denen ich denke, auf alles, was denkt, so wird die logische Erörterung des Denkens fälschlicherweise für eine metaphysische Bestimmung des Subjekts gehalten (vgl. Kant, Kritik der reinen Vernunft, B 403f.). Wir müssten, um Intentionalität erklären zu können, eine Supertheorie entwickeln – eine Theorie, die gewissermaßen alle Theorien überblickt. Wir müssten ein ultimatives Interpretationsraster im Kopf haben, das alle möglichen Interpretationsraster in ihrer Gesamtheit miteinbezieht. Ein solcher Standpunkt wäre aber ein Standpunkt von nirgendwo. Denn die menschliche Vernunft kann aufgrund ihrer speziellen Verfasstheit überhaupt nicht über die Gesamtheit aller Interpretationsschemata verfügen. Als Begründung führt Putnam ein Argument an, das uns bereits in diesem Text in Zusammenhang mit Putnams Deutung von Gödels Unvollständigkeitstheorem begegnet ist. Die Vernunft könne nämlich über alles, was sie zu formalisieren vermag, hinausgehen. Es sei ein wesentliches Merkmal der Vernunft, dass sie sich nicht selbst überblicken kann, sondern alles, was sie überblicken kann, zu transzendieren vermag. Daher kann es auch kein Interpretationsraster geben, das alle Interpretationsraster in ihrer Gesamt-

heit überblickt. Es ist, so könnte man jedenfalls aufgrund der vorangehenden Überlegungen meinen, vielleicht der größte Triumph des menschlichen Geistes, dass er sich selbst nicht versteht.

Eine ähnliche Kritik an dem Versuch, Bezugnahme in einem physikalischen Artefakt zu realisieren, kommt auch von einer gänzlich anderen philosophischen Richtung, nämlich von Hubert Dreyfus, einem Heidegger Interpreten und langjährigen Kritiker der Künstlichen Intelligenz. Trotz gewisser Parallelen sind es aber die Unterschiede in den beiden Kritiken, die im Anschluß genauer besprochen werden müssen. Beginnen wir aber zunächst mit der Herausarbeitung der Parallelen. In seiner Kritik an der orthodoxen Künstlichen Intelligenz erwähnt Dreyfus immer wieder das so genannte „Frame"-Problem. Frames sind Interpretationsraster, die wir benötigen, um Signale aus der Umwelt überhaupt interpretieren zu können. Es sind stereotype Erwartungshaltungen, mit deren Hilfe wir uns in der Umwelt orientieren. So hilft uns das „Geburtstagsparty Frame" dabei, eine mitgebrachte Flasche Wein als ein Geschenk zu interpretieren. Dreyfus konnte nun aber zeigen, dass die Implementierung solcher Frames in einem Computerprogramm keine hinreichende Bedingung darstellt, um in diesem ein Verständnis für eine Situation erzeugen zu können. Denn woher soll denn ein derartiges Programm überhaupt wissen, dass es sich bei der zu interpretierenden Situation ausgerechnet um eine Geburtstagsparty handelt und dass daher das Geburtstagspartyframe zu aktivieren ist? Damit überhaupt dieses Frame aktiviert werden kann, muß bereits vorausgesetzt werden, dass wir ein bestimmtes Geschehen *als* Geburtstagsparty interpretieren. Wir bräuchten also eine Art Superframe, das uns die Entscheidung abnimmt, welches Frame für welche Situation gerade relevant ist. Die Suche nach einem derartigen Superframe führt indes zu einem unendlichen Regreß.

Betrachtet man Frames als Interpretationsschemata, so erinnert Dreyfus' Kritik an der Forschungsstrategie der Künstlichen Intelligenz deutlich an Putnams Kritik an einer reduktionistischen Naturalisierung von Bezugnahme. Dennoch gibt es einen gewaltigen Unterschied: Im Unterschied zu Putnam richtet sich nämlich Dreyfus' Kritik nicht grundsätzlich gegen einen naturalistischen Ansatz in den Kognitionswissenschaften, sondern nur gegen ein ganz bestimmtes Forschungsparadigma, das es seiner Meinung nach zu überwinden gilt. Seine Kritik richtet sich nämlich gegen die Idee des Rationalismus, Denken sei ein Vorgang, bei dem mentale Repräsentationen von Dingen oder Sachverhalten in der ‚Außenwelt' im Kopf des Repräsentierenden verarbeitet werden. Es ist nicht die Künstliche Intelligenz als solche, die von Dreyfus Kritik betroffen ist, sondern nur ihre Orientierung an der rationalistischen Tradition. Dreyfus schlägt in seinen neusten Arbeiten eine radikale Kehrtwende in der Künstlichen-Intelligenz-Forschung vor, die er als ‚Heideggerian AI' bezeichnet. Er schöpft seine Ideen aus dem Umfeld der Phänomenologie, hauptsächlich aus der Phänomenologie Heideggers und Merleau-Pontys. Dieser neue Ansatz von Dreyfus ist insofern bemerkenswert, als er jahrzehntelang im Namen von Heideggers Phänomenologie die Künstliche Intelligenz kritisiert hat. Es sei, wie er in seinen jüngsten Schriften behauptet, nunmehr an der Zeit, für einen positiven Beitrag zur KI-Forschung aus der Perspektive Heideggers. Er stellt sich in diesem Zusammenhang sogar die Frage, was denn zu tun sei, um die Künstliche Intelligenz „more Heideggerian" zu machen. Es ist, wenn man so will, eine Ironie der Geschichte, dass Putnam als einer der Mitbegründer des Funktionalismus in seinen letzten Arbeiten wesentlich kritischer einem physikalischen Reduktionismus gegenübersteht als dies bei Dreyfus, dem langjährigen Kritiker der Künstlichen Intelligenz, der Fall ist.

Man könnte von Dreyfus Kritik am Repräsentationalismus ausgehend Putnams Vorbehalte gegen eine Naturalisierung von Bezugnahme folgendermaßen aushebeln: Putnams Vorbehalte richten sich in erster Linie gegen eine ganz bestimmte Auffassung von Bezugnahme. Dieser Auffassung zufolge seien Gegenstände „außergeistige" Entitäten (vgl. Repräsentation und Realität, S. 211 und Vernunft, Wahrheit und Geschichte, S. 47, 73 und 75), die von mentalen Abbildern in unserem Kopf repräsentiert und verarbeitet werden. Aufgabe einer solchen Abbildtheorie des Geistes wäre es, eine eindeutige Korrespondenzbeziehung zwischen den mentalen Repräsentationen und den außergeistigen Entitäten aufzuspüren. Fodor hat versucht, dieser Korrespondenzbeziehung einen naturwissenschaftlichen Anstrich zu geben. Das von Putnam skizzierte und gleichzeitig kritisierte Bild entspricht im wesentlichen der klassischen Korrespondenztheorie der Wahrheit. Sein Haupteinwand gegenüber dem Versuch, eine eindeutige Korrespondenzbeziehung zwischen dem Denken und seinen Gegenständen ‚da draußen' ausfindig machen zu können lautet nun folgendermaßen. Schließlich sei ja nur innerhalb unserer Interpretationsschemata festgelegt, worauf wir uns beziehen. Um nun diesen Bezug erklären zu können, müßten wir von einem speziesunabhängigen Interpretationsschema ausgehen, das uns innerhalb der Grenzen unserer Vernunft nicht zur Verfügung steht. Wir können nicht objektiv die Korrespondenzbeziehung zwischen dem Denken und den Gegenständen erklären, da eben diese Beziehung erst innerhalb der Grenzen unserer ‚Heimatsprache' gestiftet wird. Mit dieser Kritik bleibt Putnam allerdings selbst noch der inneren Logik des Repräsentationalismus verpflichtet. Ausgehend von Interpretationsschemata, die wir dem Input unserer Erfahrung gewissermaßen überstülpen, wird die prinzipielle Unterscheidung zwischen einem denkenden Wesen und einer ‚außergeistigen' Realität im wesentlichen beibehalten.

Im Vergleich dazu ist der Ansatz von Dreyfus wesentlich radikaler. Aus seiner Sicht ist es die Redeweise von Interpretationsschemata ‚hier drinnen' und Gegenständen ‚dort draußen', also die Subjekt-Objekt-Spaltung der abendländischen Philosophie, die generell zu hinterfragen ist. Es ist die Idee des Repräsentationalismus als solches, die Fiktion einer externen und vom Denken abgetrennten Außenwelt, die durch subjektive Vorstellungen re-präsentiert wird, die zu einer unmöglichen Fragestellung führt. Die grundsätzliche Frage des Repräsentationalismus, wie sich Repräsentationen auf etwas beziehen können, das außerhalb des Geistes angesiedelt ist, kann mit dem begrifflichen Werkzeug des Repräsentationalismus nach der Ansicht von Dreyfus gar nicht beantwortet werden!

Dreyfus argumentiert dagegen mit Heidegger, Merleau-Ponty und dem Neurobiologen Walter Freeman, dass es gar keine Interpretationsschemata gäbe, die dem Input unserer Erfahrung erst im nachhinein eine Bedeutung zuschreiben. Statt dessen werden unsere Sinnesdaten direkt als signifikant erfahren und sind daher auch keine Abbilder einer externen Welt. Was wir in der Erfahrung an Wissen erwerben, werde von uns nicht *re-präsentiert*, sondern uns in bedeutungsvollen Situationen *präsentiert*. „The best representation of the world is (..) the world itself", behauptet Dreyfus (1998), wobei er hier ein Diktum Rodney Brooks vom MIT aufgreift. Vergleicht man diese Position mit Quines Theorem der Unbestimmtheit der Bezugnahme, so wird hier unter Berufung auf die Phänomenologie und die Neurobiologie eine unmittelbare referentielle Bestimmtheit der Sinnesdaten postuliert. Wie dies möglich sein soll, wird im nachstehenden Text behandelt.

Wie sich die Künstliche Intelligenz die Einsichten der Phänomenologie und der Neurobiologie nutzbar machen könnte, zeigt Dreyfus an drei Beispielen: 1) Am Beispiel neuronaler Netze, insbesondere so genannter FeedForward-Netze 2) An der Phänomenologie Merleau-Pontys und schließlich 3) an der Attraktortheorie Walter Freemans.

Beginnen wir mit den FeedForward-Netzen. In solchen Netzen sind die simulierten Neuronen in verschiedene Schichten unterteilt: Eine Eingabeschicht, eine oder mehrere innere, von außen unsichtbare Verarbeitungsschichten und eine Ausgabeschicht. Die Verbindungen in einem FeedForward-Netz verlaufen ausschließlich von den Neuronen der Eingabeschicht in Richtung der Neuronen der Ausgabeschicht. Verbindungen von Neuronen innerhalb einer Schicht sind diesem Modell zufolge nicht zulässig.

Lernen erfolgt in solchen Netzen durch Änderungen der Verbindungsstärke zwischen den Einheiten des neuronalen Netzes. Diese Verbindungsstärke wird in mehreren Lernzyklen so lange modifiziert, bis das Netz das gewünschte Verhalten produziert.

Am Beispiel derartiger neuronaler Netze will nun Dreyfus die schwierige Frage klären, wie es möglich sein soll, den Daten unserer Erfahrungen direkt eine Bedeutung zuzuschreiben.

Nehmen wir ein Beispiel aus der Mustererkennung. Ein neuronales Netz wird darauf trainiert, den Buchstaben A zu erkennen. Das Netz soll nach einer Trainingsphase dazu in der Lage sein, verschiedenen Inputs, die dem Buchstaben A ähnlich sind, einen ganz bestimmten Output zuzuordnen. Doch was heißt hier aber ähnlich?

Ein erstes Problem, das hier auftritt, ist, dass FeedForward-Netzen das gewünschte Lernziel in der Regel von einem Trainer vorgegeben ist. Man nennt dieses Verfahren *supervised learning* (überwachtes Lernen). Nun sind aber Menschen dazu in der Lage, direkt aus einer Situation, in der sie einbettet sind, zu lernen. Menschen sind, so Dreyfus (1998), „let alone networks". Dies ist freilich etwas überzogen – man denke an den Kaspar Hauser Effekt.

Nach dem Verfahren des supervised learning kann jedenfalls die Frage, was hier überhaupt ähnlich *für das System* heißt, nicht beantwortet werden. Denn so lange ein Trainer darüber entscheidet, was als ähnlich gilt bzw. nicht, kann die Frage, nach welchen Kriterien der Ähnlichkeit das System selber (sei es ein biologisches oder ein künstliches) entscheiden soll, gar nicht geklärt werden. Diesem Problem kann durch die Verwendung von Techniken des so genannten *Unsupervised Lerning* (Lernen ohne Lehrer) Abhilfe geschaffen werden. Es handelt sich hierbei um Lerntechniken, bei denen Kategorisierungen direkt aus der Lernumgebung des Netzes ohne Einschaltung eines externen Trainers vorgenommen werden.

Doch auch beim unüberwachten Lernen sieht Dreyfus ein grundsätzliches und scheinbar unüberwindbares Problem. Man stelle sich nur eine Situation vor, in der ein neuronales Netz eine falsche Zuordnung trifft, beispielsweise einen ganz anderen Buchstaben als A klassifiziert. Was heißt hier aber *falsche* Zuordnung? Dies hängt doch davon ab, nach welchen Kriterien der Ähnlichkeit unser Netz entscheidet. Nun bemerkt Dreyfus (1998) hierzu: „Everything is similar to everything else in an indefinitely large number of ways". Je nach Kriterium der Ähnlichkeit lassen sich daher auch ganz verschiedene Klassifizierungen vornehmen. Damit stellt sich die grundsätzliche Frage: Wie kann der theoretisch unendlich große Raum möglicher Klassifikationen auf ein menschliches Maß zugestutzt werden? Dreyfus' Antwort darauf lautet, wobei er sich in seiner Argumentation auf Merleau-Ponty stützt: Dies geschieht dadurch, dass wir einen Körper haben! Körperlosen neuronalen Netzen fehlt jedoch dieser we-

sentliche Aspekt menschlicher Fertigkeiten. Was für ein Netzwerk als ähnlich gilt oder nicht, kann daher auch innerhalb der Grenzen einer Netzwerkarchitektur nicht entschieden werden.
Nach Merleau-Ponty hat unser Körper die Tendenz, ein Gefühl des Ungleichgewichts zwischen ihm und der Umgebung, in die er eingebettet ist, zu reduzieren. Eine zentrale Rolle bei dieser Erklärung spielt Mereau-Ponty's Vorstellung eines „maximum grip". Gemeint ist damit das folgende: Wenn wir etwas mit der Hand greifen wollen, so tun wir es so, dass wir den besten Zugriff haben. Wenn wir etwas betrachten, so sind wir um eine optimale Distanz zu dem Ding bemüht - eine Distanz, die uns ein Maximum an Sichtbarkeit gewährleistet. Diese darf nicht zu nahe und nicht zu ferne sein.
Inwiefern nun durch diese Vorstellung eines „maximum grip" das Problem der Ähnlichkeit gelöst werden kann, zeigt sich durch eine Gegenüberstellung mit der Art und Weise, wie dieses Problem in der repräsentationalen Theorie des Geistes gehandhabt wird. Bleiben wir beim Buchstaben A. Nach letzterer Theorie wird eine Ähnlichkeit dadurch festgestellt, dass der Input mit einem prototypischen Muster des Buchstabens A, eine in unserem Gedächtnis gespeicherte mentale Repräsentation dieses Buchstabens, verglichen wird.
Die damit verbundenen Probleme kennen wir bereits: Schließlich ist alles mit allem in einer bestimmten Hinsicht ähnlich. Der Repräsentationalismus hat aber das Problem, kein verbindliches Kriterium für diese Ähnlichkeit finden zu können.
Nach Merleau-Ponty dahingegen wird der Input direkt, u.zw. in einem körperlichen Sinne, als Abweichung von einer Norm begriffen, also als eine Abweichung davon, wie der Buchstabe unter optimalen Wahrnehmungsbedingungen erscheinen sollte (vgl. Kelly, 2003, S. 14).
Dieses Optimum ist nicht ein gegenüber unserem Handeln extrinsisch vorgestelltes Ziel, sondern, wie Dreyfus betont, ein intrinsisches Merkmal von Fertigkeiten, die in unserer Hand, nicht aber in unserem Kopf, verankert sind. In diesem Sinne bezieht sich intentionales Verhalten direkt auf die Dinge, ohne sie dabei nur intellektuell vorzustellen.
Dinge sind in diesem Kontext keine theoretischen Gegenstände des kognitiven Verstandes, sondern etwas greifbares, also etwas, das wir zuallererst erfahren, wenn wir es in der Hand haben. Wir sehen die Welt nicht durch die Brille eines unbeteiligten Beobachters, sondern als Betroffene. Hier kommt eine alte Vorstellung von Dinglichkeit zum Tragen, wie sie uns bereits im Altgriechischen begegnet. Dinge sind Pragmata, Zeug. Es war Martin Heidegger, der im abgelaufenen 20. JH diese Bedeutung von Dinglichkeit besonders betont hat. Dazu ein einfaches Beispiel: Die Dinglichkeit eines Hammers lässt sich von zwei verschiedenen Seiten her betrachten: Auf der einen Seite ist ein Hammer ein Stück Stahl ausgestattet mit einem Stiel aus Holz. Holz und Stahl haben ganz bestimmte physikalische Eigenschaften, die wir auch objektiv im Labor feststellen können.
Dies ist jedoch nicht der Hammer, wie er uns zunächst im alltäglichen Umgang begegnet, der Hammer ist bei einer solchen Betrachtung lediglich Gegenstand einer wissenschaftlich objektivierenden Analyse. Eine solche Betrachtungsweise eines Hammers setzt voraus, dass wir von seiner Eigenschaft als alltäglichem Gebrauchsgegenstand absehen. Denn die Analyse im Labor erfolgt ja unter Ausklammerung seines ökologischen Settings.
Heidegger hat dahingegen darauf hingewiesen, dass eine solche Betrachtungsweise eines Dinges nur eine Reduktion auf ein materiell vorhandenes Objekt ist und dergestalt immer nur abgeleitet, also parasitär sein kann gegenüber jener Dinglichkeit, wie sie uns zunächst im alltäglichen Gebrauch begegnet. Die Verdrehung des ursprünglich phänomenalen Sachverhalts ist es auch, die Heidegger der rationalen Philosophie, hier vor allem in der Gestalt von Des-

cartes, vorwirft. Heidegger macht dann letztlich gerade diese Verdrehung dafür verantwortlich, dass der Rationalismus die strikte Trennung von Geist und Körper, von Subjekt und Objekt behaupten konnte.
Dieser Dualismus von Körper und Geist findet nun nach Dreyfus unterschwellig seine moderne Entsprechung im Repräsentationalismus der Cognitive Science. Denn auch der Repräsentationalismus unterstelle noch eine Trennung von mentaler Vorstellung und Inhalt, von Sinnesdaten und deren Interpretation, von unmittelbarer Wahrnehmung und deren Reflexion bzw. Repräsentation im kognitiven Apparat und übersieht dabei aber, dass unser Alltagshandeln immer schon in einer materiellen Welt verankert ist. Dass diese Koppelung an eine materielle Welt das primäre ist und dass jede Reflexion über diese Koppelung nur sekundär, d.h. aufgesetzt sein kann, zeigt sich nach Dreyfus schon allein daran, dass wir erst dann zu reflektieren beginnen, wenn es zu Unterbrechungen, zu Störungen in unserem Alltagshandeln kommt.
Aus diesem Grunde kann auch der Repräsentationalismus gar keine genuine Beschreibung unserer motorischen Fertigkeiten liefern. Denn er orientiert sich an Sondersituationen, in denen eben diese Fertigkeiten auf die eine oder andere Weise gestört sind. Mit anderen Worten: Der Repräsentationalismus versucht aus einer Ausnahmesituation eine Norm, eine generelle Regel abzuleiten.
Man versuche nur, den Wind als bloßes Geräusch wahrzunehmen, das dann erst in unserem Kopf zum Begriff ‚Wind' verarbeitet wird. Dies bedarf einer gehörigen Abstraktionsleistung. Eine solche Abstraktionsleistung gilt es zu sehen als das, was sie wirklich ist: als ein Absehen bzw. Abstrahieren von dem tatsächlich wahrgenommenen Phänomen. Was wir unmittelbar wahrnehmen, ist eben der Wind und kein bloßes Geräusch. Diese Wahrnehmung ensteht nicht erst dadurch, dass wir bedeutungslosen Sinnesdaten nachträglich ein Interpretationsschema darüber stülpen. Interpretationsschemata sind, so betrachtet, nur ein Versuch einer nachträglichen (Re-)Konstruktion unserer unmittelbaren Erfahrung. Ein solcher Erklärungsversuch kommt aber immer schon zu spät. Denn es ist nicht möglich, aus dem Konstrukt das Phänomen selber abzuleiten. Umgekehrt wird ein Schuh daraus: Zuerst ist das Phänomen und dann erst das Konstrukt. Das heisst: Es bedarf einer sehr künstlichen Einstellung, wollen wir gesprochene Worte zuerst als Geräusche hören, die wir nachträglich mit Bedeutungen ausstatten: Den Bedeutungen wachsen Worte zu, so Heidegger, nicht aber werden Wörterdinge mit Bedeutungen versehen (Heidegger, 1927, S. 161).
Die Vorstellung, Sinnesdaten würden passiv aufgenommen und erst danach intern in bedeutungsvolle Sinneseinheiten umgewandelt, entstünde nur dann, wenn wir uns von unserem eigenen beteiligten Handeln dissoziieren. Bei der Bewältigung unserer Alltagsaufgaben stünden wir statt dessen in einer wesentlich engeren Verbindung zu den Gegenständen unseres Handelns als dies aus der Perspektive eines distanzierten Beobachters der Fall ist. Streng genommen sind es die Dinge selber, die uns zu ganz bestimmten Handlungen nötigen.
Wie menschliches Verhalten immer schon in eine Welt eingebunden ist (être au monde), erklärt sich Merleau-Ponty über einen ursprünglich körperlichen und vorprädikativen Bezug zu den Dingen. Wie es aber unser Körper schaffen soll, ein Ziel anzustreben, ohne es vorher mental zu repräsentieren, klingt fürs erste betrachtet wie ein Stück Magie. Tatsächlich meint Merleau-Ponty, unsere motorischen Fertigkeiten seien magisch auf eine ständige Verbesserung bzw. Verfeinerung ausgerichtet, ohne dass wir aber diesen Umstand bewusst wahrnehmen.

Wie kann dies aber in einem existierenden Gehirn möglich sein? „After all the brain is not a wonder tissue“, meint Dreyfus (1996), der hier eine Formulierung Daniel Dennetts aufgreift.
Eine mögliche Erklärung auf der Grundlage der Neurobiologie für dieses Phänomen sieht Dreyfus nun in der Attraktortheorie von Walter Freeman. Freemans Theorie liefere somit die neurobiologische Unterfütterung für die antirepräsentationalistischen Ressentiments, wie wir sie bei Heidegger und Merleau-Ponty vorfinden. Freeman stützt sich auf neuere Forschungsergebnisse im Bereich der bildgebenden Verfahren in der Neurobiologie und versucht diese mit Termini aus der Chaostheorie zu beschreiben.
Ich möchte an Freeman’s Neurodynamik vor allem unter dem Blickwinkel der beiden folgenden Fragen herangehen:
1) Inwiefern wird der Input unserer Erfahrung von unserem Gehirn direkt als signifikant erfahren?
2) Was entspricht bei Freeman dem ‚maximum grip’?
Ein Ausgangspunkt für die Frage, wie unser Gehirn Verhalten steuern kann, ohne dabei von mentalen Repräsentationen Gebrauch zu machen, ist die Hebbsche Lernregel. Dieser Regel zufolge ändert sich die Verbindungsstärke zwischen den Neuronen auf der Grundlage von Erfahrungen. Lernen erfolgt dabei durch eine schrittweise Justierung dieser Verbindungsstärke. Das besondere bei Freeman ist, dass er diesen Justierungsprozeß nicht der Mikroebene, also der Ebene der Aktivitätslevel einzelner Neuronen zuschreibt, wie man sich dies etwa im Sinne eines bottom-up-Verfahrens erwarten könnte, die Aktivität eines einzelnen Neurons erfolgt dahingegen immer im Chor und in Abstimmung mit einem ganzen Neuronenensemble, in das es eingebettet ist.
Freeman nennt dieses Phänomen „the first building block of neurodynamics” (Freeman, 2000, S. 72) und meint damit, dass das Aktivitätslevel der Neuronen von einer Population bestimmt wird, nicht aber von einzelnen Neuronen. Der Lernvorgang besteht dann darin, dass für die verschiedenen Klassen antrainierter Stimuli so genannte chaotische Attraktoren im Gehirn erzeugt werden. Die Identifikation eines bestimmten Stimulus erfolgt dieser Theorie zufolge dann dadurch, dass der entsprechende Attraktor aktiviert wird.
Ein Attraktor ist nichts anderes als ein stabiler Gleichgewichtszustand, auf den ein dynamisches System sich im Laufe der Zeit zu bewegt. Pendelt das System zwischen mehreren solchen Zuständen und ist das Eintreten eines bestimmten Zustands nicht vorhersehbar, so spricht man von einem chaotischen Attraktor.
Freeman konzentriert sich in seinen Forschungen vor allem auf das olfaktorische System, denn dieses ist das einfachste und zugleich das älteste im Vergleich etwa zu Sehen und Hören. Da die Mechanismen, die im olfaktorischen System gefunden werden, sich auch auf andere Systeme der Sinneswahrnehmung übertragen lassen, dient es prototypisch dazu, um das Zustandekommen von Sinneswahrnehmung erklären zu können.
Ein wesentliches Problem, das es gleich vorneweg zu klären gilt, ist das Problem der so genannten Stimulus Konstanz (stimulus constancy, a.a.O., S. 88). So aktiviert ein bestimmter Stimulus bei jedem Feldversuch nur einen Bruchteil jener Rezeptoren, die auf die chemische Struktur dieses Stimulus ansprechen. Dieser Bruchteil aktivierter Rezeptoren bildet dann vor dem Hintergrund eines Neuronenaggregats ein räumliches Muster, vergleichbar mit bestimmten Sternbildern am Nachthimmel, in denen das Flackern der einzelnen Sterne als winzige mikroskopische Punkte wahrgenommen wird.

Dieses räumliche Muster im Aggregat der Rezeptoren wird dann an den olfaktorischen Bulbus weitergeleitet, um dort ein entsprechendes räumliches Muster zu erzeugen. Freeman spricht hier von einem „topographic mapping“ von Rezeptoren zum olfaktorischen Bulbus. Dieses Muster ändert sich nun bei jedem Feldversuch, den wir für einen bestimmten Geruch durchführen. Freeman deutet diesen Befund nun folgendermaßen: Auf der Mikroebene, also unter Betrachtung der mikroskopischen Aktivität der Neuronen im Bulbus, lässt sich keine Stimulus Invarianz feststellen.
Zur Überraschung für Freeman und sein Forschungsteam ließ sich dann aber auf der Makroebene, also unter Betrachtung der Neuronenaktivität des gesamten Bulbus ein räumliches Muster beobachten, das in allen Feldversuchen für den gleichen Geruch auch gleich geblieben ist. An der Erzeugung dieses Musters sind aber alle Neuronen des Bulbus beteiligt. Freeman hat daraus zunächst geschlossen, dass jedes Neuron im Bulbus an der Wahrnehmung jedes Geruchs seinen Anteil hat.
Dieses Muster isoliert betrachtet stellt nur einen relativ kleinen Baustein dar in Freemans Modell der Neurodynamik, das ja von einer hochgradigen Vernetzung aller Wahrnehmungskomponenten auf der Makroebene ausgeht. Dennoch eignet es sich gut dazu, um prototypisch erklären zu können, wie sich Freeman aus der Perspektive der Neurowissenschaften die Entstehung von Bedeutung erklärt.
Im EEG konnte Freeman diese für jeden Stimulus charakteristischen Muster im Hochfrequenzbereich (zwischen 20 und 100 Hz) beobachten. Es handelt sich hierbei um aperiodische und unvorhersehbare Wellen, ausgelöst durch Schwingungen von Millionen von Nervenzellen. Jedes Muster weist aber bei gleicher chaotischer Wellenform an verschiedenen Stellen des Bulbus eine unterschiedliche Amplitude auf. Die chaotische Wellenform dient daher als Träger für ein so genanntes Amplitudenmodulationsmuster (kurz: AM-Muster; vgl. a.a.O., S. 97).
Wichtig ist nun, dass dieses für jeden Stimulus charakteristische AM-Muster jede Übereinstimmung mit den auslösenden Stimuli vermissen lässt und statt dessen durch globale chaotische Vorgänge, an denen das gesamte olfaktorische System beteiligt ist, erzeugt wird. Das Muster ist keine Repräsentation eines Reizes, es entspricht den geschichtlich gewachsenen Bedeutungen, die ein Stimulus für das Individuum hat. Dies ist zumindest die Deutung, die Freeman dem Phänomen gibt.
Aus diesem Grunde hat auch jedes Individuum für den gleichen Geruchsreiz ein unverwechselbares, nur aus dem Kontext seiner individuellen Lebenserfahrung erklärbares AM-Muster.
Weiters hat sich herausgestellt, dass das Erlernen eines neuen Musters zu einer globalen Änderung aller Muster führt. Das Erkennen eines bestimmten Stimulus ist daher abhängig vom gesamten Kontext der Erfahrungen, die ein Individuum im Laufe seines Lebens gemacht hat.
Es ist diese Kontextabhängigkeit, die auch erklären kann, wie aus einer theoretisch unendlichen Menge von Substanzen in der Luft gerade ein bestimmter Geruchsstoff eine Reaktion auslösen kann. Dies zeigt zugleich, wie das von Dreyfus beschriebene Frame-Problem in der Terminologie von Freemans Neurobiologie gelöst werden kann: Das Gehirn extrahiert keine Informationen aus einer äußeren Umgebung, statt dessen erzeugt es eigene AM-Muster aufgrund der Bedeutung, die ein Stimulus für das Individuum hat.
Es ist Freemans ausdrückliches Bekenntnis zum Pragmatismus, das uns letztlich zur Beantwortung der eingangs gestellten Frage führt, nämlich: Warum wird der Input unserer Erfah-

rung von unserem Gehirn direkt als signifikant erfahren? Die Antwort lautet: Weil die Signifikanz ja erst intern im Gehirn erzeugt wird.
Merleau-Pontys Idee, unsere motorischen Fertigkeiten seien magisch auf einen optimalen Zugriff, auf die Beseitigung eines Ungleichgewichts des Selbst und seiner Welt ausgerichtet, findet bei Freeman seine Entsprechung im Konzept der Assimilation. Das Gehirn lerne über seine Umgebung, indem es sich in bestimmten ausgewählten Aspekten dieser angleicht. Dabei werde im Gehirn eine Kopie (ein AM-Muster) erzeugt. Diese Kopie sei jedoch kein Abbild einer außergeistigen Realität, sie sei dahingegen abgestimmt auf die Ziele und Lebensgeschichte des jeweiligen Organismus. Wir sehen die Welt nämlich nicht durch die Brille eines Betrachters, sondern als Beteiligte und das Be-greifen der Dinge erfolgt entlang einer Trajektorie, die uns in die Nähe eines Zustands der optimalen Assimilation des Selbst an ein Objekt führt. Freeman versteht einen solchen Zustand als ein Attraktorbecken.
Er spricht in diesem Zusammenhang von der Unidirektionalität der Wahrnehmung und meint damit das folgende: Wir handeln mit unseren motorischen Fertigkeiten in eine Welt hinein. Was wir zurückbekommen, sind jedoch nicht die Konturen einer in Folge unseres Handelns geänderten Welt, sondern die Konturen eines durch den Feedback unseres Handelns geänderten Selbst!
Wortwörtlich schreibt Freeman: „We act in the world and then change ourselves in accordance with the impact the world has on our bodies following our actions." (a.a.O., S. 87) Aus diesem Grunde sei auch die Fabrik, in der Bedeutungen erzeugt werden, ein geschlossenes System (vgl. a.a.O., S. 38).
Dieser *epistemologische Solipsismus* ist freilich nicht ganz neu. So sprach bereits Humberto Maturana in den achtziger Jahren des vergangenen Jahrhunderts von der „operationalen Geschlossenheit des Nervensystems" (1987, S. 179).
Dies führt uns zu folgenden kritischen Anmerkungen. Die neurobiologische Interpretation phänomenologischer Befunde, wie wir sie bei Merleau-Ponty und Heidegger vorfinden, darf uns über eines nicht hinwegtäuschen: In seinem Bemühen, drei zunächst gänzlich verschiedene Diskurse zusammenzuführen, nämlich Neurobiologie, Chaostheorie und schließlich Phänomenologie, muss an Freeman die Frage gestellt werden, inwiefern er den damit scheinbar gewonnenen Erkenntniszuwachs letzten Endes nicht nur einer metaphorischen Rhetorik verdankt.
Dazu drei Fragen: 1) Chaotische Attraktoren in der Chaostheorie sind genau definierte mathematische Konstrukte. Inwiefern lassen sich AM-Muster direkt – und eben nicht nur metaphorisch – in der Sprache der Chaostheorie rekonstruktiv beschreiben? 2) Was ist unter der Angleichung des Gehirns an das Objekt, worauf es intentional bezogen ist, näherhin zu verstehen? Freemans Antwort darauf ist eher vage und bedient sich der Rhetorik von Merleau-Ponty, also eben jener Rhetorik, die es ja neurobiologisch zu erklären gelte (vgl. Freeman, 2000, S. 165). 3) Inwiefern lässt sich überhaupt „intentionality", die Gerichtetheit eines denkenden Individuums auf ein Objekt oder Sachverhalt rein biologisch, als makroskopische Prozesse im Gehirn beschreiben? Inwiefern kann also Freeman das im Titel seines Buches enthaltene Versprechen, nämlich erklären zu können, „how brains make up their mind", auch tatsächlich einlösen?
Letztere Frage bedarf einer genaueren Behandlung. Wie ich bereits weiter oben dargelegt habe, unterscheidet Hilary Putnam in seinem Buch *Renewing Philosophy* zwischen einem „succesful belief" und einem „true belief". Nur wir als denkende Wesen seien dazu in der Lage,

unterscheiden zu können zwischem demjenigen, das nur so ausschaut wie und jenem, das tatsächlich so ist. Geben wir statt dessen einem Hund ein Stück Fleisch bestehend aus pflanzlichen Proteinen, so wird er keinen Unterschied zu echtem Fleisch feststellen, sofern auch ersteres die gleiche Funktion erfüllt wie letzteres. Es ist nämlich eine wesentliche Eigenschaft von uns als denkenden Wesen zwischen Schein und Sein, zwischen Täuschung und Wahrheit unterscheiden zu können. Träume, Halluzinationen, visuelle Illusionen sagen uns, dass die Welt, die wir erfahren, nicht die gleiche ist, wie die externe Welt, die durch die Welt der Erfahrung zuallererst re-präsentiert wird. Doch was bedeutet dies nun? Ist der Repräsentationalismus entgegen der Annahme von Dreyfus ein unverzichtbares Element bei der Beschreibung von uns als denkenden, reflektierenden Wesen?

Konfrontiert man das Problem der Fehlrepräsentationen mit Freeman's Attraktortheorie, so stellt sich heraus, dass die Neurobiologie hier zu kurz greift, um eine adäquate Erklärung anbieten zu können. Der Grund hierfür liegt an Freeman's Pragmatismus. Würde uns beispielsweise tatsächlich ein Stück pflanzlicher Proteine gereicht, die von natürlichem Fleisch ununterscheidbar sind, so würde dieser Stimulus in die exakt gleiche Attraktorschale fallen wie echtes Fleisch. Die Attraktortheorie gibt uns keinerlei Möglichkeiten in die Hand, hier einen Unterschied feststellen zu können. Interessant ist in diesem Zusammenhang eine Frage, die Stuart Dreyfus seinem Bruder bezüglich seines Artikels „Heideggerian AI" gestellt hat: Gibt es Attraktoren für Karotten, Sellerie usw. oder handelt es sich hierbei nur um verschiedene Essensanreize (vgl. Dreyfus, 2006, S. 21)? Um diese Frage aber beantworten zu können, müssten wir, so denke ich jedenfalls, unterscheiden können zwischen einem „true belief" und einem „successful belief". Dies bringt uns aber zu einem grundsätzlichen Problem der Heideggerian AI.

Ihre Idee, der Input unserer Erfahrung werde direkt als signifikant erfahren und nicht erst hinterher von einem kognitiven System mit Bedeutungen ausgestattet, ist von dem Bemühen geleitet, den Repräsentationalismus und die damit verbundene Subjekt-Objekt-Spaltung der westlichen Philosophie überwinden zu können. Dabei bedient sie sich aber eines bestimmten Tricks. Die Dinge, auf die wir intentional bezogen sind, werden nicht mehr als das Andere unseres Denkens und Wahrnehmens erfahren, sondern nurmehr als „Zeug". Dahinter steht eine pragmatisch verkürzte Heideggerdeutung. So hat bereits in den 50er Jahren des vorigen Jahrhunderts Alfred Delp Heidegger ironisch vorgeworfen, dieser verwandle den Menschen in den Besitzer eines riesigen Zeughausladens. Diese pragmatische Heideggerauslegung übersieht jedoch, dass der in Sein und Zeit beschriebene Zeugcharakter der Dinge nur ein erster Schritt ist, um die prinzipielle Offenheit der menschlichen Existenz aufzeigen zu können.

Tatsächlich versucht Heidegger die Probleme der klassischen Erkenntnistheorie, die noch von einem weltlosen Subjekt und einem ihr gegenüber unabhängigen Objekt ausgeht, durch Rekurs auf einen ursprünglichen Bezug des Menschen zu den Dingen aufzulösen. Nicht die Tatsache, dass bislang noch kein Beweis für die Existenz der Außenwelt erbracht wurde, sei, so meint er jedenfalls, der größte Skandal der Erkenntnistheorie, sondern vielmehr, dass überhaupt nach einem solchen gefragt wurde.

Dreyfus spricht unter Bezugnahme auf Heidegger von einer „originary transcendence", die jeder Subjekt-Objekt-Spaltung und jeder representationalistischen Interpretation des Denkens noch voraus ginge (vgl. z.B. a.a.O., S. 59). Doch was ist unter dieser „originary transcendence" zu verstehen? Jeder intentionale Bezug, sei es zu Dingen des alltäglichen Gebrauchs oder zu theoretischen Entitäten gründe in einem *Seinsverständnis*, das der Mensch aufgrund seiner

Seinsverfassung noch vor jeder Ausbildung einer konkreten Ontologie immer schon mitbringt. Heidegger bezeichnet deshalb diese Seinsverfassung des Menschen als *Dasein*. Das Seinsverständnis wiederum sei zurückzuführen auf eine vorgängige Vertrautheit mit den Dingen in ihrer Ganzheit, welche von Heidegger als In-der-Welt-sein des Menschen bezeichnet wird.

„In describing the phenomenon of world Heidegger seeks to get behind the kind of intentionality of subjects directed towards objects discussed and distorted by the tradition, and even behind the more basic intentionality of everyday coping, to the context or background, on the basis of which every kind of directedness takes place." (a.a.O., S. 88)

Ob allerdings diese Interpretation einer genaueren Analyse von Heideggers äußerst komplexem Weltbegriff standhält, sei vorläufig dahingestellt. Ich werde weiter unten auf diese Problematik näher eingehen. Dreyfus meint jedenfalls das folgende: Betreten wir beispielsweise ein Zimmer, so nehmen wir nicht einzelne Gegenstände nacheinander wahr, um sie dann nachträglich erst als Teile eines Zimmers zusammenzufügen, sondern das umkehrte ist der Fall: Die einzelnen Dinge werden nur im Kontext des ganzen Zimmers als solche wahrgenommen:

„The basic idea is that for a particular person to be directed toward a particular piece of equipment, whether using it, perceiving it, or whatever, there must be a correlation between that person's general skills for coping and the interconnected equipmental whole in which the thing has a place." (Drefyus, 1991, S. 102)

„Originary transcendence" meint also eine vorgängige – jedem Bezug zu einzelnen Dingen vorausgehende – Erschlossenheit des Seienden in seiner Gesamtheit. Indem der Mensch von Hause aus bereits eine Welt mitbringt, sei er zuallererst dazu in der Lage, in einen Bezug zu einzelnem Seiendem zu stehen.

Dreyfus unterscheidet terminologisch diese „originary transcendence" (Erschlossenheit von Welt) von einer „ontic transcendence" (Bezug zu einzelnem Seienden). Interessant ist nun aber, wie er diesen Unterschied interpretiert: Ontische Transzendenz, der Bezug zu einzelnem Seienden, sei ein „transparent coping with specific things" (Dreyfus, 1991, 107). Er meint damit das folgende: Verlassen wir die philosophische Gelehrtenstube und betrachten wir die Art und Weise, wie uns die Dinge bei der Betreibung unserer Alltagsgeschäfte zunächst und zumeist begegnen, so stellt sich heraus: Gegenstände des alltäglichen Gebrauchs sind gerade dann sie selber, wenn sie aus dem Zentrum unserer Aufmerksamkeit verschwinden, für unsere bewusste Wahrnehmung also unsichtbar (transparent) bleiben. Beispiele hierfür finden sich im ersten Abschnitt von Sein und Zeit, etwa der Hammer in der Werkstatt, der seine Funktion gerade dann erfüllt, wenn wir nicht eigens auf ihn acht geben. Gegenüber dieser ontischen Transzendenz sei nun die Erschlossenheit von Welt nichts anderes als die menschliche Fähigkeit, sich in einer Welt als Ganzes zurechtfinden zu können. Diese Interpretation markiert einen zentralen Punkt in Dreyfus' Einschätzung von Heideggers philosophischem Denken. Sie soll daher in voller Länge zitiert werden:

„One needs to be finding one's way about in the world in order to use equipment, but finding one's way about is just more coping. Any specific activity of coping takes place on the background of more general coping. Being-in-the-world is, indeed, ontologically prior (..) as the ontological condition of the possibilty of specific activities, yet being-in-the-world is just more skilled activity." (Dreyfus, 1991, S. 107)

Der Repräsentationalismus und die Subjekt-Objekt-Spaltung der abendländischen Philsophie sei dahingegen nur ein abgeleiteter Modus gegenüber diesem ursprünglichen Bezug des Menschen zu den Dingen. Reflexion, Selbstbewußtsein, Rationalität gehören sozusagen nicht zur ontologischen „Grundausstattung" des Menschen, sondern seien nur ein kultureller Unfall in der Menschheitsgeshcichte, der aber korrigiert werden kann. So meint Dreyfus: „A simplified culture in an earthly paradise is conceivable in which members' skills mesh with the world so well that one need never do anything deliberately or entertain explicit plans and goals." (Dreyfus, 1991, 85)

Trotz dieser Ablehnung des Rationalismus will Dreyfus seine Beschreibung von Heideggers Bezug zu den Dingen als „transparent coping with specific things" nicht als einen nur instinkthaften, gewissermaßen „bewußtlosen" Umgang mit den Dingen verstanden wissen. Menschliches Handeln sei „*purposive* without the actor having in mind a purpose"(93). In unserem Handeln verfolgen wir nicht die mentale Vorstellung eines Ziels. „Phenomonological examination confirms that in a wide variety of situations human beings relate to the world in an organized manner without the constant accompaniment of representational states that specify what the action is aimed at accomplishing." (93)

Obwohl keine bewußte Handlung, unterscheidet sich der menschliche Umgang mit den Dingen in mehreren Punkten vom mechanischen Verhalten eines Roboters oder Insekts (vgl. 68): Zwei dieser Punkte verdienen es besonders hervorgehoben zu werden: Wenn etwas schief geht, so sind wir überrascht. Dies sei nur möglich, da menschliches Verhalten auf die Zukunft ausgerichtet ist. Zweitens: Jeder menschliche Umgang mit den Dingen hat seine Sicht. Es ist eine Grunderfahrung, durch die sich uns zuallererst eine Welt und Dinge in dieser Welt erschließt. Diese Offenheit zu den Dingen ist Grundlage für unser Seinsverständnis. Und erst dieses Seinsverständnis ermöglicht intentionales Verhalten (12, 53, 88-89, 107).

Dreyfus hat diese Interpretation in einem viel beachteten Kommentar zu *Sein und Zeit* vorgestellt. Die eigentliche Leistung dieses Kommentars besteht nicht nur darin, Heideggers schwierigen Text erst einem englischsprachigen Publikum zu gänglich gemacht zu haben. Auch der deutschsprachige Leser findet hier eine einfürhende Interpretation vor, der es gelungen ist, Heideggers mitunter prätentiöse Ausdrucksweise in eine klare und kohärente Sprache übersetzt zu haben. Es ist aber zugleich die Verständlichkeit dieses Kommentars, die eine sachliche Kritik ermöglicht und damit neue Zugänge zu Heidegger erschließt, an die Dreyfus selber möglicherweise gar nicht gedacht hat. Es sind vor allem zwei Fragen, die sich bei der Lektüre von Dreyfus' Kommentar einstellen:

Die erste Frage lautet: Inwiefern ist Heideggers Transzendenz als jene Vorbedingung, die noch vor jeder Reflexion so etwas wie einen ursprünglichen Bezug zu den Dingen ausmacht, tatsächlich nichts anderes als ein „holistic background coping" (vgl. 104)? Anders gefragt: Ist Heideggers In-der-Welt-sein „as the ontological condition of the possibility of specific activities" nichts anderes als "just more skilled activity"? (vgl. 107)

Eindringlich warnt Heidegger vor einer derartigen pragmatischen Verkürzung seines Weltbegriffs: „Wenn man gar den ontischen Zusammenhang der Gebrauchsdinge mit der Welt identifiziert und das In-der-Welt-sein als Umgang mit den Gebrauchsdingen auslegt, dann ist ein Verständnis der Transzendenz als In-der-Welt-sein im Sinne einer Grundverfassung des Daseins aussichstlos." (VWG, 153f.)

Ist das Reflexionslose, vorbewußte In-der-Welt-sein wirklich gleichzusetzen mit einer Daseinsform, die am Ende noch gar nicht den Unterschied von Schein und Sein kennt? So be-

hauptet jedenfalls Dreyfus: „Coping practices do not represent and so cannot misrepresent“ (Dreyfus, 1991, 249). Dieser Mangel an möglicher Fehlrepräsentation erinnert entfernt an Fodors psychophysikalische Prädikate, die nicht den Komplexitätsgrad von Begriffen haben (man denke an den Unterschied von ‚seeing’ und ‚seeing as’). Was versteht also Heidegger tatsächlich unter diesem vorprädikativen Bezug zu den Dingen?

Darüber hinaus ist als zweites zu fragen, welchen Stellenwert Heidegger reflexivem Denken und theoretischem Erkennen nun tatsächlich bei der Beschreibung von Dasein beimisst. Kann das ontologisch „spätere“, aus dem ursprünglichen In-der-Welt-sein schließlich nur abgeleitete, tatsächlich bei der Beschreibung der ontologischen Grundverfassung von Dasein ausgeklammert werden? Ist nicht vielmehr auch das theoretische Erkennen eine „Seinsart“ von Dasein?

Erst wenn diese beiden Fragen geklärt worden sind, kann nämlich überhaupt angemessen geprüft werden, inwiefern so etwas wie ‚Heideggerian AI’ – eine naturalistische Übersetzung von Heidegger – möglich ist. Erst dann lässt sich auch klären, inwiefern die Sprache der Neurobiologie überhaupt für die von Heidegger gesehene ‚Sache’ die angemessene Begrifflichkeit zur Verfügung stellen kann. Dies führt uns aber zu einem ganz neuen und eigenständigen Thema, zu einer Auseinandersetzung mit Heideggers philosophischem Denken, wobei als Einstieg der Kommentar von Dreyfus herangezogen wird.